Local to Global Multi-Scale Multimedia Modeling of Chemical Fate and Population Exposure

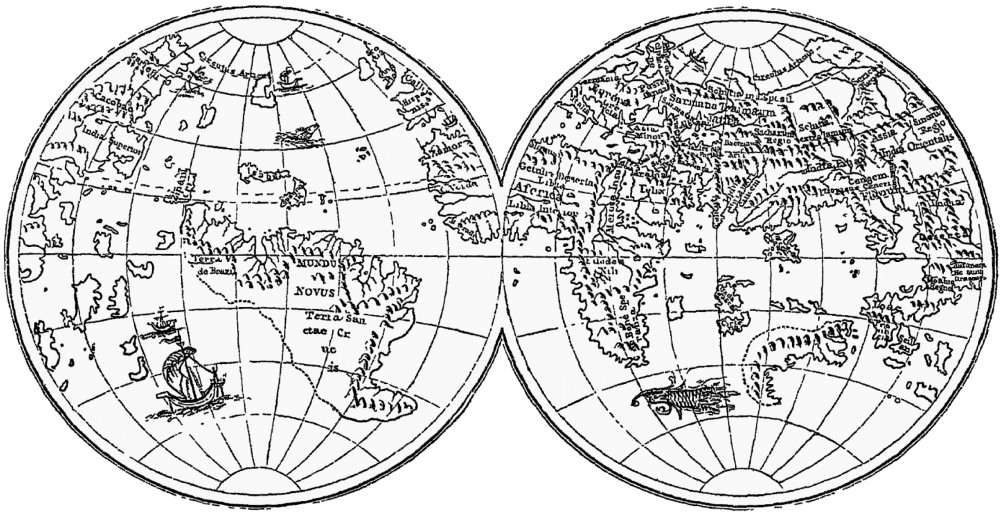
by

Cedric Wannaz

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Environmental Health Sciences)
in the University of Michigan
2017

Doctoral Committee:

Professor Olivier Jolliet, Chair
Professor Stuart A. Batterman
Associate Professor J. Tim Dvonch
Professor Carl P. Simon



HIC SUNT DRACONES

Cédric Wannaz

wannaz@umich.edu

ORCID iD: 0000-0002-1885-8713

Dedication

I am dedicating this attempt to produce *something of value* to my grandmother, Angèle Schlub (November 6th, 1919 – March 17th, 2017), who spent a selfless life behaving according to what she thought was fundamentally right, regardless of consequences for her. Having lived a lifetime of modesty, she belonged to an aristocracy of the heart that is universally the most endangered species.

Je dédie cette tentative de produire *quelque chose de valeur* à ma grand-maman, Angèle Schlub (6 novembre 1919 – 17 mars 2017), qui a passé sa vie à agir en accord avec ce qu'elle pensait être fondamentalement Juste, en faisant abstraction de sa propre personne. Ayant vécu une vie fondée sur la modestie, elle appartenait à une Aristocratie de Cœur qui est universellement l'espèce la plus menacée.

Cédric Wannaz, April 10th, 2017

Acknowledgements

While studying physics at the *Ecole Polytechnique Fédérale de Lausanne* (EPFL) in Switzerland, I became progressively aware of environment- and health-related global issues and their coupling. I had studied and worked in electronic engineering and computer science in the industry, worked in autonomous robotics, and was finalizing my master thesis in mathematical physics under Tudor Ratiu at the mathematics department of EPFL. I had a strong motivation to combine computer science, mathematics, and physics, and to apply them to the study of problems with a direct human dimension, which are typical to the fields of public and environmental health. I hence took an extra-curricular class on *sustainable developments* taught by Olivier Jolliet, who then became my advisor. He gave me the opportunity to tackle such interdisciplinary research, proposing a research topic about global impacts of pollutants following a spatial multi-scale approach, which led to the development of the *Pangea* model.

Most life trajectories are altered by unforeseen circumstances, however. Making no exception to this rule, mine happened in a rather severe fashion through health issues that struck first in 2007, and then much more severely from 2009 to 2016. These recurring autoimmune type of events would leave me partly unable to use my hands for typing. After recovering adequately in 2016, I could finally finish my thesis.

I therefore want to give my most profound and sincere thanks to my wife, Elizabeth Hedgeman, who went through all of this with me. She is a very unique person, if only because she can cope with me (illness set aside!). This appreciation must also be extended to our two families for their perpetual kindness.

I want to express my sincere gratitude to my advisor, Olivier Jolliet. He is the human definition of energy and positivity and has been loyally finding projects for helping me progress through this at a limited pace.

I also want to express my sincere gratitude to Kristin Keith, PT (another human definition of positivity) from the U of M Hospital, who gave me the tools for finding a path to recovery.

I would also like to thank my thesis committee members, Stuart Batterman, Carl Simon, and Tim Dvonch, who have been understanding and flexible given the circumstances.

The iMod group core (may visitors forgive me) has been a formidable aggregate of modeling nerds and friends. I want to thank (alphabetical order) Susan Csiszar, Alexi Ernstoff, Lei Huang, Dingsheng Li, Meghan Milbrath, Vy Nguyen, Katerina Stylianou, Karin Veltman, and Yvan Wenger for their friendship and the very pleasant interactions that we had.

I have to address special thanks to special friends (alphabetical order): Maïté Andres, Peter Fantke, Andrew Henderson, and Shanna Shaked, who are all unique in their own exceptional way. It is a privilege to know them.

Peter Fantke has been a *Pangea*-enthusiast since the early days of the framework, and a constant, reliable support. His ability to answer questions the *day after* at the latest – even when it involves reviewing a full article overnight – never ceases to impress me. I want to deeply thank him for this.

I also want to thank Antonio Franco and John Kilgallon, from the Unilever Safety and Environmental Assurance Centre (SEAC) team, who have been extremely supportive with the research presented in Chapter 3. They brought expertise, data, and a vision/scope for integrating *Pangea* into a longer-term tiered approach for analyzing impacts of chemicals on ecosystems.

I finally want to thank the administrative staff of EHS, and in particular (alphabetical order): Sue Crawford, Brenda Hadley, Kelsey Hargesheimer, Nancy Polderdyke, Beverly Slane, Patrice Somerville, and Cecilia Young, who probably saved my administrative life more times than I can count. Without them, nothing would have been possible!

Table of Content

List of Tables

List of Figures

List of Appendices

List of Acronyms

CTD	Characteristic travel distance
DALY	Disability Adjusted Life Years
EM	[S*pecific to Pangea*] Environmental Model Examples: atmospheric EM (wind model), hydrological EM (e.g. fresh water)
EPM	[S*pecific to Pangea*] Environmental/Exposure Process Model Examples: EPM characterizing chemical degradation in fresh water, EPM characterizing transfer of chemical between air and fresh water, EPM characterizing exposure through inhalation
FAO	Food and Agriculture Organization of the United Nations
GIS	Geographic Information System Examples: ESRI/ArcGIS, Quantum GIS
iF	Intake fraction
LCIA	Life Cycle Impact Assessment
LRT(P)	Long Range Transport (Potential)
MTD	Mean Travel Distance
PEC	Predicted Environmental Concentration
PNEC	Predicted No Effect Concentration
PiF	Product intake fraction
POP	Persistent Organic Pollutant
P_{ov}	Overall environmental persistence

RE	[S*pecific to Pangea*] Re-indexing Engine Component of Pangea from/to grid-specific indexing schemes to/from the virtual indexing schema
RP	[S*pecific to Pangea*] Refinement Potential Scalar field (or raster) that defines the need for grid refinement at each point of the globe
VOC	Volatile Organic Compound

Abstract

To assess environmental and human exposure to chemical emissions, two types of approaches are available: 1. spatially and temporally intermediate- to high-resolution, substance- and location-specific analyses, and 2. lower resolution and less specific analyses aiming for broad coverage. The first category is time and resource intensive, which limits its utility, while the second may be less accurate but allows for evaluation of large numbers of substances and situations. None is well suited for analyzing local to global population exposure associated with point sources (e.g. incinerators). There is a need for a multi-scale approach of intermediate complexity that bridges the advantages of both approaches: high resolution when relevant, the ability to evaluate large numbers of substances, and a level of accuracy that is "useful" (for decision-makers).

This thesis aims to 1. develop a multi-scale, multimedia fate and transport, and multi-pathway population exposure modeling framework, 2. evaluate the modeling of fate and transport using large-scale inventories of emissions and measured environmental concentrations, 3. evaluate local to global population exposure associated with large sets of point sources covering a wide variety of local contexts (e.g. up/down-wind/stream from large populations, important water bodies, or agricultural resources), and 4. simulate a large national inventory of emissions and perform multi-media source apportionment.

Coupling a geographic information system (GIS) and a computation engine, we develop the *Pangea* framework, which offers a unique ability to discretize the globe using three-dimensional (3D) multi-scale grids, to overlay Eulerian fate and transport multimedia models, and to compute multi-pathway population exposure. This flexible platform enables us to build study-specific global grids with high resolution where relevant, to spatialize existing reference non-spatial multi-media models (e.g. USEtox), and to parameterize them by gridding spatial global data sets and models that characterize e.g. the meteorology, the hydrology, the land cover, and the distributions of population and food production.

We first apply this framework to predict the fate and transport of home and personal care (HPC) chemicals in all of Asia, with a focus on freshwater concentrations and ecosystem exposure. This study provides both a large-scale high-resolution spatial inventory of emissions and a large data set of more than 1,600 monitoring values in fresh water and sediments. We compare predicted environmental concentrations (PECs) and measurements and find good agreement for the long-lived triclosan in fresh

water (Pearson r=0.82), moderate agreement for shorter-lived substances, and a large discrepancy specifically for parabens in sediments. We analyze a dilution effect in large rivers, particularly for short-lived substances, leading to higher concentrations in tributaries. This study also reveals the limitation of the present underlying gridded hydrological data set (WWDRII) when comparison with measurements at monitoring sites is required, which prompts the evaluation of a finer, catchment-based hydrological data set (HydroBASINS).

We then focus on human exposure and analyze the evolution of the population intake fraction – the fraction of the emission *taken in* by the population – as a function of the distance from the source. We simulate emissions from 126 point sources (stacks) corresponding to the location of solid waste treatment plants (SWTPs) in France, and compute the radial distribution of population intake fractions through inhalation and ingestion. We determine that a substantial fraction of emissions may be taken in by the population farther than 100 km away from point sources (78.5% of the inhaled benzene and 54.1% of the ingested 2,3,7,8-TCDD). We also demonstrate the feasibility of simulating large numbers of emission scenarios by extending the study to 10,000 emission point sources densely covering France.

We finally extend the previous emitter-oriented studies and framework to support receptor-oriented analyses, i.e. source apportionment. We simulate the fate and transport of 43 substances emitted from 4,101 point sources defined by the Australian National Pollutant Inventory (NPI) for the period 2014-2015. We compute subsequent population exposure, effects (cases of non-cancer and cancer), and severity (DALY). Formaldehyde, benzene, and styrene are found to be the three top contributors in terms of DALYs. We demonstrate the technical feasibility of multimedia, large-scale source apportionment. Exploiting NPI meta-information and source apportionment allows *Pangea* to analyze impacts per industrial sector, per industry, per region, or overall, and to trace back to major contributors (from emission inventories) associated with environmental levels and population exposure at any point on the globe.

This research opens new perspectives in spatial, local to large-scale fate and exposure modeling by enabling “real-time” adaption of the model geometry to project-specific requirements. The flexibility of *Pangea* enables the performance of sensitivity studies to the geometry and to the underlying data sets and models, which could not be achieved before. Finally, the global multi-scale approach facilitates a systematic evaluation of the evolution of population exposure as a function of the distance from point sources, within the same consistent framework. The major limitations of the framework are the underlying first-order fate and transport models and a limited availability in up-to-date global data sets.

Chapter 1 Introduction

1.1 Overarching project

The underlying scientific question that motivates this research is ***how to determine and compare, within the same consistent methodological framework, local to global impacts of substances (e.g. pollutants) that undergo a multimedia fate and lead to a population multi-pathway exposure?***

This work is part of a larger, overarching research effort undertaken by the *Impact and Risk Modeling Laboratory* (iMod) over the past 15 years, which aims to develop an evaluative modeling chain from molecule (produced, disposed, distributed in products) to impacts on human health and ecosystems. Research topics include environmental fate and far-field population exposure, food and goods production, near-field consumer exposure to chemicals in products, and internal exposure. The approach is top-down, often driven by needs from Life Cycle Impact Assessment (LCIA), striving for parsimony (in terms of data requirements, computation time, and complexity), targeting methods that capture main trends and allowing for evaluation of large numbers of substances and scenarios.

This introduction has four aims:

- Present a schematic of the research undertaken at iMod and situate this thesis in this context.
- Discuss modeling requirements and model availability to address the main scientific question.
- Define the general objectives of this research.
- Define the four specific aims of the thesis.

Figure 1.1 illustrates key contributions from former and current iMod collaborators, with a general classification per main research lines. These domains converge toward high-throughput evaluation of the total exposure, both at the individual and population levels, and to subsequent health impacts. In a sense, this work is well in line with new trends in “exposomics”, but from a top-down, LCIA point of view.

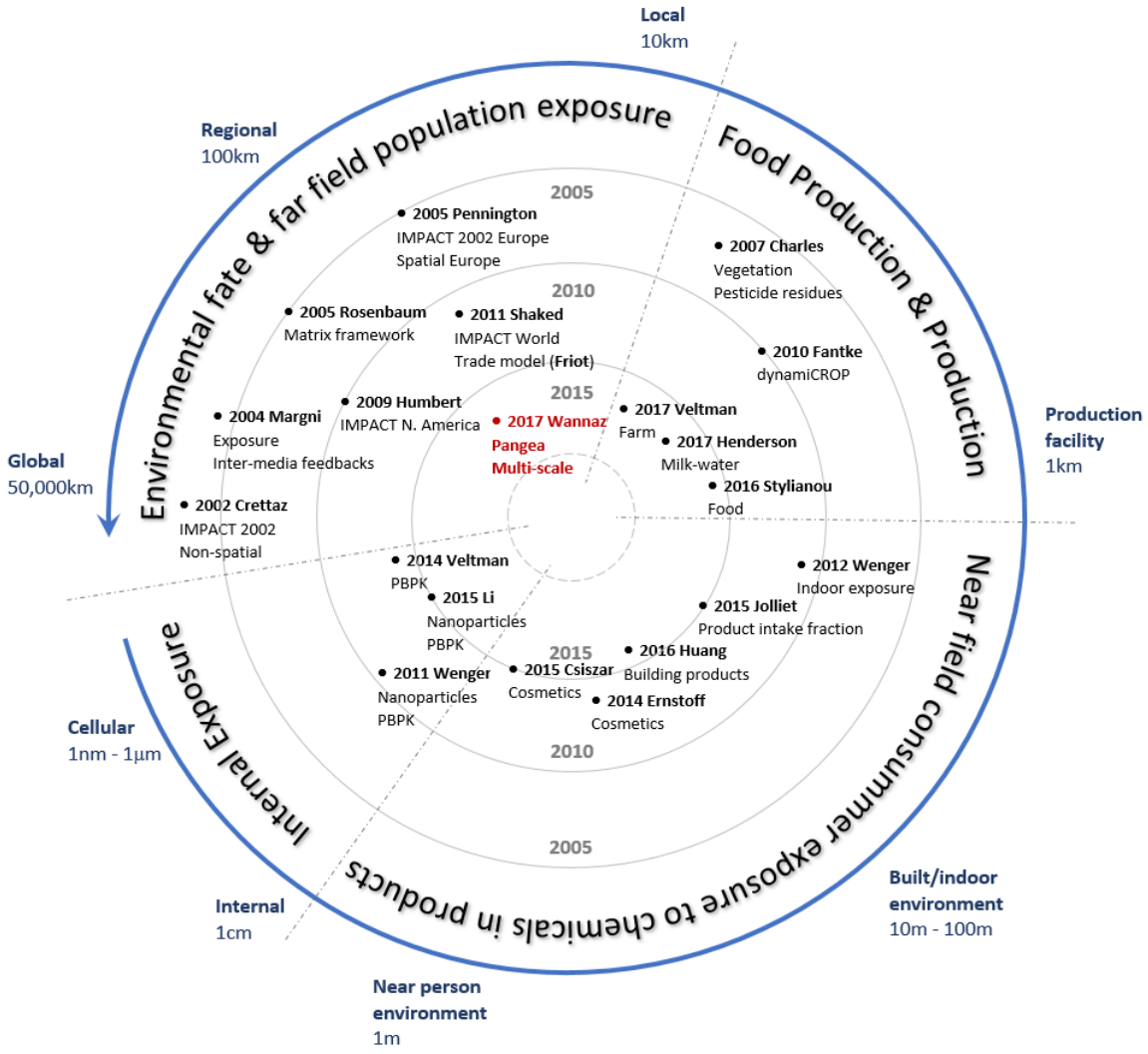


Figure 1.1 – General structure of the research undertaken at iMod, and the location of this research.

This thesis, and the development of the multi-scale model called *Pangea* lies in the line of spatial fate and exposure modeling (Figure 1.1), building on the work from Crettaz et al. (2002), Margni et al. (2004), Pennington et al. (2005), and more recently from Shaked (2011) and Fantke et al. (2010), who contributed to *Pangea* by helping define the environmental process models used in *Pangea*. This work constitutes the next logical step with the development of *Pangea*, a spatial multi-scale fate and exposure model that enables the evaluation of population exposure from local (~10 km) to global scales and for multiple emission scenarios.

1.2 Modeling requirements and model availability

Several approaches are available to model the fate and transport of substances, from local to global scales, and the subsequent exposure of populations to those substances. This section reviews mechanisms relevant to modeling fate and transport and the relevant models that could help to address the scientific questions studied in this thesis.

1.2.1 Role of multiple media in modeling source to intake

Presence of pollutants in multiple media – The literature targeting the fate of chemicals in the environment is clear in stating that pollutants can undergo multimedia transport. Scheringer and Wania (2003), for example, have described persistent organic pollutants (POPs) as "*multimedia chemicals with notable amounts in several environmental media, and reversible exchange processes between these media*", and have indicated that "*single-media approaches tend to be insufficient when aiming for a comprehensive understanding of the long-term fate of POPs*".

Therefore, several studies have reported and developed the presence of pollutants in multiple media, as well as the role of these media and the exchanges between them.

Several studies have evidenced that pollutant transport through *rivers* and *oceanic currents* is an important mechanism, and that many pollutants emitted into the air reach substantial steady-state concentrations in media other than air. Using the general circulation model ECHAM5 coupled with the aerosol model HAM, Sehili and Lammel (2007) have demonstrated that the polycyclic aromatic hydrocarbons (PAHs) benzo[a]pyrene (B[a]P) and fluoranthene (FLT) emitted in Europe and Russia are mostly distributed to soils (64-97% by mass) after five years, that only 1-5% are present in the air medium, and that 2-33% are present in oceans after 10 years. Using a global multimedia, zonally averaged model, Wania et al. (1999a; 1999b) have described the Arctic Ocean – the sea water medium – as the final reservoir for alpha-Hexachlorohexane (α-HCH).

Sediment and *soil* can also be significant: Gouin and Harner (2003) have revealed that polybrominated diphenyl ethers (PBDEs) will largely partition into organic carbon in these media, while Sehili and Lammel (2007) have reported that B[a]P and FLT are mostly distributed to soils five years after their emission.

Vegetation has been studied by Horstmann and McLachlan (1998), who have demonstrated the importance of forests through a series of measurements of semi-volatile organic compounds (SVOCs) in a coniferous and a deciduous forest canopy. Following this work, Wania and McLachlan (2001) have extended an existent non-steady-state multimedia fate model and shown that, during the growing season,

the inclusion of a canopy compartment in a model can decrease the average air concentrations by a factor of five. Similarly, Prevedouros et al. (2004b) have presented evidence of pollutant scavenging by forests and subsequent accumulation by organic-rich terrestrial surfaces. Multimedia models have hence been extended to include vegetation compartments, as described in Cousins and Mackay (2001), Bennett et al. (2002), and Margni (2003) for the IMPACT 2002 model.

Other media such as *snowpack* (Gouin et al., 2002; Meyer et al., 2009a, 2009b, Meyer and Wania, 2011, 2008) and *paddy fields* (Wei et al., 2008) have also been determined to play a significant role and therefore have been integrated into a few multimedia models.

More generally, the importance of accounting for multiple media (i.e. beyond atmospheric transport only) has been emphasized by Bennett et al. (2002), Webster et al. (2004, 1998), Pennington et al. (2005), EEA Report (2007, review from G. Lammel, p.114), Mackay and Webster (2006), and especially (MacLeod et al., 2004), who have demonstrated the crucial importance of emission media and the source(s) location(s), or in other words, the *mode of entry* of pollutants into the environment.

Fate and transport – Atmospheric transport is a key factor in the long-range transport of substances. Nevertheless, multiple other media can also play a significant role in pollutant transport. Wania and Mackay (1999) and Scheringer and Wania (2003) have reported that efficient *coupling of adjacent zones by large rivers* is lacking in current multimedia models. They have also found that *oceanic currents* are as vital as the atmosphere for the northbound transport of α-HCH. Lammel et al. (2007) have supported this, concluding that current models should be improved by (better) accounting for oceanic currents.

Another critical aspect of pollutant fate and transport is the potential for *multi-hop* or "grasshopper" transport, caused by the temperature-dependent volatility of certain pollutants (e.g. POPs), which can lead to several consecutive depositions and re-emissions (Fernández and Grimalt, 2003; Hansen et al., 2004; Wania and Mackay, 1996). Gouin and Harner (2003) have established that PBDEs are sensitive to diurnally and seasonally fluctuating temperatures, leading to situations where 50% of the total mass of PBDE-47 that is deposited to vegetation is re-emitted to the atmosphere.

The metrics that characterizes fate and transport reflect this dependence between media: parameters such as *overall persistence in the environment* (P_{ov}), or the *characteristic transport distance* (CTD) may vary significantly when evaluated with multimedia and single-medium approaches. The definition of P_{ov} and CTD for multimedia environments has been extensively studied, with a few examples mentioned here. Müller-Herold (1996) has defined and related effective multi-compartment decay rates to single-compartment rates, Bennett et al. (1998) have proposed a methodology for determining the CTD in a

multimedia environment, and Webster et al. (1998) and Mackay and Webster (2006) have evidenced how multimedia persistence is related to the mode of entry and partitioning.

Population exposure – Six or seven pathways are currently considered important for studying an environmentally mediated multi-pathway exposure (Rosenbaum et al., 2008): inhalation and ingestion, with the latter through drinking water, fish, beef, eggs, above-ground vegetation (e.g. cereals, fruits, and vegetables) and below-ground vegetation (e.g. carrots and potatoes). For example, an individual can be exposed directly through air inhalation, but also through beef ingestion, the result of cow inhalation (and potential bio-concentration of the substance). This example indicates that the inclusion of multi-pathway exposure requires a multimedia evaluation of the fate and transport.

Emissions can vary depending on the technology and scenario used. It is therefore essential for a model design to be independent of emissions intensities so that it is flexible enough to be applied to many emission substances and scenarios.[1] Rather than calculating a total intake, Evans et al. (2002, under the name *exposure efficiency*) and (Bennett et al., 2002a, 2002b), in building on discussions among experts from indoor air pollution, risk assessment, and life-cycle assessment, have proposed to calculate the *intake fraction* (iF), which refers to the fraction of emitted pollutant that is eventually taken in by a population.[2] The iF is a product of the fate and the exposure through the different pathways (MacLeod and McKone, 2004). Since its definition, the iF has been widely used to characterize source to intake at a population level (e.g. Wambaugh et al., 2013).

General trends in multimedia modeling – Over the last decades, the need for multimedia-based approaches has led to the development and improvement of several multimedia models of varying complexity: BETR (MacLeod et al., 2001; Woodfine et al., 2001), CalTox (McKone and Enoch, 2002), ChemRange (Scheringer, 1996; Scheringer et al., 2002), ClimoChem (Scheringer et al., 2000), CoZMo-POP 2 (Wania et al., 2006), EQC (Centre for Environmental Modelling and Chemistry, CEMC), EUSES (Attias et al., 2005; Vermeire et al., 1997), G-CIEMS (Suzuki et al., 2004), IMPACT 2002 (Margni et al., 2004; Pennington et al., 2005), SimpleBox (Hollander et al., 2016), TaPL3 (CEMC), TRIM.FaTE (US-EPA), and USEtox (Hauschild et al., 2008; Rosenbaum et al., 2008). This growth in modeling has also prompted the extension of single-medium models (such as atmospheric transport models) to include other media, to some extent. These extended models are sometimes referred to as "atmospheric

[1] This is valid as long as the application of the models is limited to cases in which linear behaviors dominate. The practical interest of the intake fraction vanishes for non-linear cases, where e.g. saturation is important (i.e. the fraction of the emissions that is taken in depends on the intensity of the emissions).

[2] Intake fractions associated with inhalation and ingestion are usually computed and discussed separately. They summarize the fate and exposure processes, and they are what we are ultimately going to compare at local and global scales.

dispersion models adapted for POPs", and have been described in Koziol and Pudykiewicz (2001), Shatalov et al. (2016, EMEP/MSC-E), and Roeckner et al. (2003, ECHAM5). However, they generally suffer several limitations; few are global, they mostly model other media as reservoirs (two-dimensional as described in Semeena et al. (2006), or with no reverse flows), and they do not model the exchanges between these reservoirs (such as between soil and plants).

Needs – There is a clear need to utilize a multimedia approach to characterize the source-to-dose fate and multi-pathway exposure. If only the fate was of concern, it could be possible to use and manually cascade some single-medium models critical to fate and transport. However, the need to include multi-pathway exposure (with ingestion pathways associated with most aforementioned media) imposes the use of a multimedia model. The alternative would be to employ up to seven single-medium models reflecting the different exposure pathways, which would make inter-media exchange and feedbacks difficult to model. It is therefore better to adapt and extend current multimedia modeling tools and techniques, improving them if necessary, to achieve our goal of estimating local to global exposure to pollutants. The treatment of both local and global scales implies that it is of interest to further disaggregate the overall iF – as a well-suited and appropriate exposure metric – by region of exposure.

1.2.2 Spatial resolution

Multimedia models can be split into two broad categories: generic (or non-spatial) multimedia models and spatially resolved (or spatial) multimedia models. In order to address both local and global impacts, the model must be able to operate at different scales, and a review of existing multimedia models is necessary.

Non-spatial models – Generic multimedia models such as CalTox, EUSES, CoZMo-POP, SimpleBox, and USEtox have proven useful for screening and comparing various chemical emissions (Beyer et al., 2000; EEA, 2007; Pennington et al., 2005). They are therefore beneficial for performing chemical risk assessment, as well as a decision-support tool for policy makers. The main advantages of non-spatial models are that they have low complexity and low data requirements, but maintain a multimedia approach.

Spatial models – Many of the factors that are crucial to the study of pollutant fate may vary spatially, such as the aforementioned presence of forests, lakes, and rivers, as well as the temperature or the hydroxyl radical concentration (Anderson and Hites, 1996), for example. Pennington et al. (2005) have emphasized this, highlighting the need for less spatially resolved models to at least differentiate emissions to water that go directly into the seas from those emissions to water that flow into lakes with long residence times. On the exposure side, population density (for inhalation-dominant substances) and

food production intensity (for pollutants that bio-concentrate in the food chain) have been identified as the most relevant parameters (MacLeod and McKone, 2004; Pennington et al., 2005; Rochat et al., 2006) that vary extensively with location. These parameters are all available as spatial, geo-referenced data sets.

Multiple spatial multimedia models are currently available, with various regions covered and resolutions. Examples are BETR North America (MacLeod et al., 2001; Woodfine et al., 2001), BETR Global (MacLeod et al., 2005), BETR Europe (Prevedouros et al., 2004a, 2004b), IMPACT 2002 (Margni et al., 2004; Pennington et al., 2005), Impact World (Shaked, 2011, with continent-sized regions), Impact North America (Humbert et al., 2009, with North America nested in a one-box world), Globo-POP (Wania and Mackay, 2000 and references therein, global latitudinal), ClimoChem (Scheringer et al., 2000, global latitudinal), ChemRange (Scheringer, 1996, global 1D-circular), G-CIEMS (Suzuki et al., 2004, Japan), and NIAES-MMM-Global (Wei et al., 2008, and references therein, global fate of pollutant emitted from Japan and all of Asia).

A category of geographic information system (GIS)-based models has recently emerged (Pistocchi, 2014) that can achieve particularly high spatial resolution. Because they are based on raster algebra, however, they can become data and computation intensive, especially in contexts where both high resolution at specific locations and global-scale coverage are required. While multimedia coupling (with feedback) is possible, it remains rather complicated to achieve, especially if one needs to tailor parameterization on a per-study basis.

Spatial and non-spatial models have been extensively compared by Klepper and den Hollander (1999), Fenner et al. (2005), Pennington et al. (2005), Rochat et al. (2006), Armitage et al. (2007), and Hollander et al. (2007). Fenner et al. (2005) have compared nine spatial and non-spatial multimedia models using a set of 3,175 hypothetical chemicals covering a broad range of chemical characteristics. They have demonstrated that overall persistence and *long-range transport potential* (LRTP), which are metrics for fate and transport, have almost identical ranking orders in spatial and non-spatial models for organic chemicals. However, problems with non-spatial models arise when seeking to obtain absolute values for P_{ov}, LRTP, or environmental concentrations, as Beyer et al. (2000), Pennington et al. (2005), Armitage et al. (2007), and Hollander et al. (2007) have stated. As a consequence, generic model outputs are not suited for comparison with measurements (Rochat et al., 2006; Webster et al., 2004).

Potential and limitations of spatial models – Most of the models mentioned above have medium to low resolution. The latitudinal models split the whole globe into several bands (accounting for the changes of the latitudinal gradient of Henry's law constants), BETR North America comprises 24

regions, IMPACT 2002 Europe and North America use a 2°×2.5° grid, and Impact World uses continent-sized regions. Only the regional and more recent model G-CIEMS reaches higher resolutions (5 km × 5 km fixed grid). On the one hand, authors have warned that a high spatial resolution is worthless if data sets are low resolution (EEA, 2007), but on the other hand, authors have acknowledged the need for a fine spatial resolution. Webster et al. (2004) have concluded that region-specific landscape parameters have a significant impact on the quality of the model output: "*While region-to-region differences are small compared to the chemical-to-chemical differences that may span many orders of magnitude for physical-chemical or degradation properties, chemical fate is not the same for regions of differing landscape parameters. … It is therefore concluded that the quality of results obtained from regional environmental fate models can be improved by the use of region-specific landscape parameters.*". Similarly, Scheringer and Wania (2003) and Wania and Mackay (1999) have cautioned us against artifacts due to averaging (mostly for longitudinal averaging that is assumed in latitudinal models).

Spatial artifacts constitute main incentives for using models with a higher resolution than is currently available in the realm of multimedia models. Figure 1.2 illustrates an artifact resulting from the instantaneous full-mixing assumption in compartmental models when large cells are involved: a cell as defined by the red box (close to a latitudinal cell) leads to a situation in which emissions in Australia are immediately diluted and transferred to both Africa and South America. This creates a coupling that may lower environmental concentrations near sources of emission, but also induces artificially high environmental concentrations in remote areas at steady state (long-time solution) for long-lived substances.

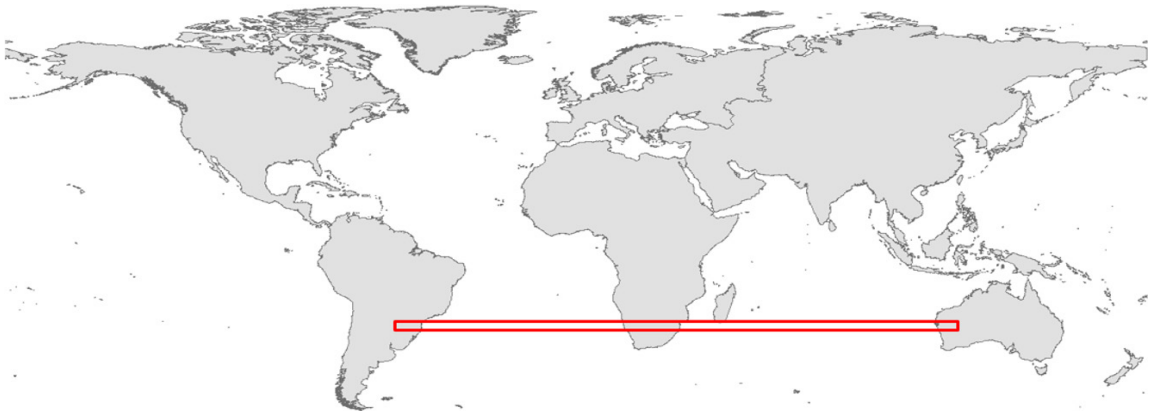


Figure 1.2 – Instantaneous homogeneity (full mixing): large grid cells, such as the red cell represented here (typical for latitudinal models), lead to a situation (artifact) in which emissions in e.g. Australia are immediately diluted and transferred to Africa and South America.

Figure 1.3 depicts a situation in which the world is discretized using a regular grid[3] that appears rather dense at first glance. However, zooming in on a 200 km × 200 km grid cell near Chicago (the background raster is the population density, with dark brown pixels representing highly populated regions of Chicago), we see that a model using this grid would not differentiate any of the four emission locations S1 to S4, since these sources happen to lie within a single grid cell. For such model, the highly urban S1 would then be equivalent to the rural S2 to S4, and vice versa. For some substances and exposure pathways, this difference could be critical. This example illustrates an artifact of non-localization/discrimination/separation of sources and receptors.

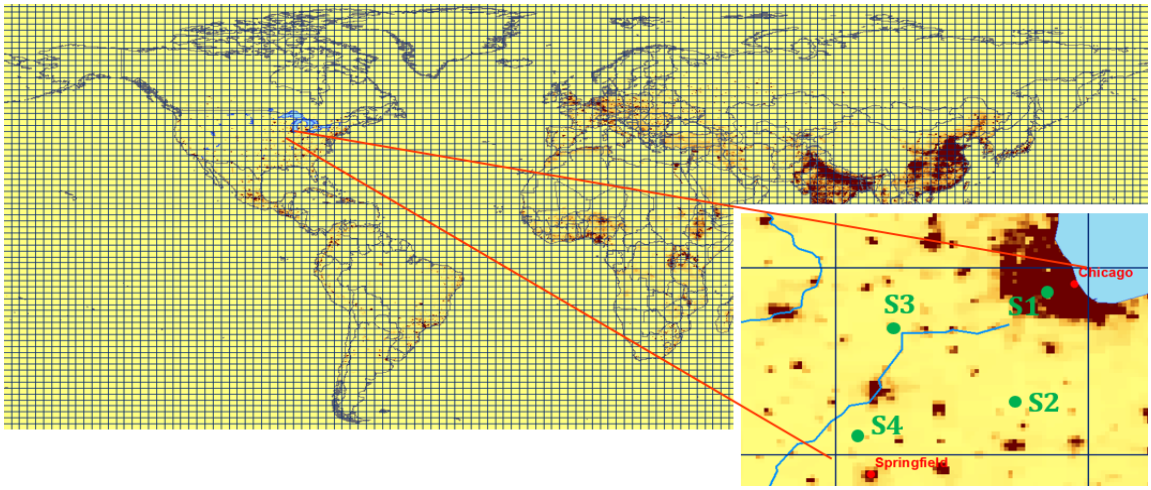


Figure 1.3 – Fixed grid, large number of cells for covering the world, raster: population density, cells near Chicago: ~200km × 250km, sources S1(urban), S2-S4(rural) are equivalent for a model at this resolution, artifact: non-localization/discrimination of sources and receptors + artificial, instantaneous dilution.

Finally, Figure 1.4 presents a situation in which the Rhône River in Switzerland flows down the Alps to Lake Léman, and then from Lake Léman to the Mediterranean Sea. The grid cell in green is the grid cell from the IMPACT 2002 Europe model; grey/red cells are finer cells from a hydrological model. We see that the IMPACT model is unable to differentiate between a pollutant emission upstream (cell 26368 of the hydrological model) and a pollutant emission downstream from a lake, despite a difference of around 3,000 days in water-residence time.

[3] In a space of longitudes and latitudes.

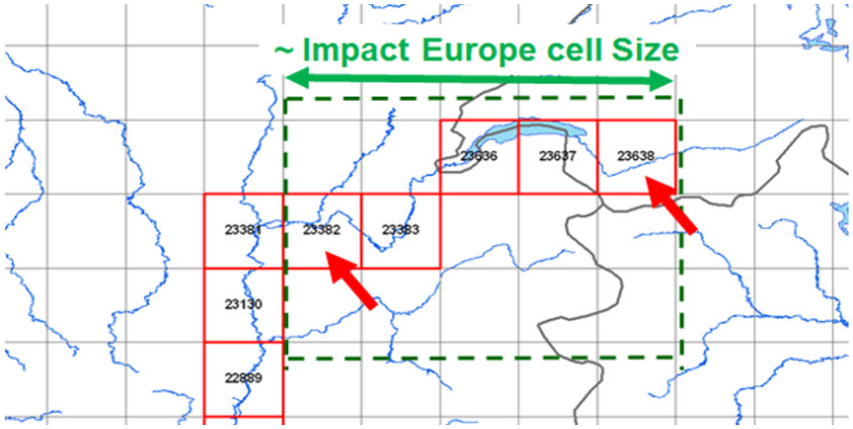


Figure 1.4 – IMPACT 2002 spatial Europe grid (green) versus WWDRII grid (grey/red) in Lake Leman (Switzerland) region. IMPACT 2002 does not differentiate between emissions in cells 23382 and 23638, while the lake substantially differentiates the fate between these cells, with around 3000 days of residence time for water.

In their comparison of multimedia box models and finer single-medium models (sometimes extended), Scheringer and Wania (2003) have mentioned an emerging category of models that lies in the middle, characterized by a multimedia structure with a "high" resolution. In a recent paper, Humbert et al. (2009a) have developed and used a static "multi-scale" version of Impact North America that can divide North America into indoor, urban, and regional zones with a medium resolution, and then nest it within a one-compartment world box. While taking a step in the direction of the multi-scale, this model and these introductory studies present drawbacks in that a significant part of the work (several person-months) that was undertaken is grid and data dependent, and that neither the grid nor the data sets are intended to evolve. Relatively minor changes, such as adjusting the region of study, adding or changing media or compartments, or even only updating some data sets and changing the grid while keeping the same region, would require much additional work. This restricts the performance of sensitivity studies to grid variations. In other words, it is impossible to determine the impact of the initial choice of grid(s), or the impact of a variation in the grid(s). Global scale (world) models usually deal with continent-size regions,[4] but it is not yet known whether this is sufficiently detailed.

Needs – None of the currently published and operational multimedia models is able to properly evaluate local to global impacts. There is a clear need for increasing their spatial resolution in order to assess exposure at local scales while avoiding the above-described artifacts in regions where they occur. Simultaneously, the evaluation of POPs that may travel worldwide requires that the global scale be taken into account.

[4] It goes up to 288 regions for BETR Global (MacLeod et al., 2005), but this model does not include exposure.

The authors are not aware of any model that has a fine enough local resolution, is multimedia for fate and transport, has a global scale, and includes multi-pathway exposure. Literature has not yet identified the resolution needed to accurately – or at least adequately (by minimizing artifacts) – predict the iF for a variety of substances. There is therefore a need to better understand and assess the influence of spatial resolution on multimedia modeling outcomes, and to identify critical spatial parameters that affect the evaluation of the fate and exposure. This requires a model with the necessary flexibility for varying grid geometries and for processing available spatial data efficiently.

1.2.3 Challenges and needs

In this review of modeling requirements and models, we have identified the need to address the following challenges:

- Estimate and compare population multi-pathway iFs (source to dose) of substances that undergo multimedia fate and transport, at small (local) to large (global) distances from emission sources. In other words, we want to determine the relative importance of local versus global impacts of emission sources within the same modeling framework (same systematic error).
- Minimize spatial artifacts in multimedia models. This work seeks to achieve high resolution where relevant, while keeping the approach parsimonious.

We have posited that in order to address these challenges, the following are needed:

1. A model with multimedia fate and transport, and multi-pathway population exposure.
2. A flexible model, capable of building project-specific geometries quickly and evaluating a variety of environmental media.
3. A model capable of accounting for spatial variability (using geo-referenced data sets) and context-specific geometries (e.g. the geometry of streams, if the objective is to compare predicted concentrations with measurements in fresh water).

This review has demonstrated that there is currently no tool that has these features. Thus, we propose to develop a multi-scale framework named *Pangea* that is built on the body of knowledge associated with current multimedia models (typically IMPACT 2002 and USEtox), but implemented in a structure that incorporates a computation engine and a GIS engine. This latter engine is instrumental to provide enough flexibility for building project-specific multi-scale grids and geometries.

Figure 1.5 displays the evolution of both the spatial and the non-spatial branches of IMPACT 2002-based models and the major connections between these branches. In the non-spatial branch, USEtox emerged as a combination and an update of SimpleBox fate and transport models and IMPACT 2002

exposure models. It started as a generic model and evolved towards a regional model that is not spatially explicit, but which allows for selecting a region parameterized by IMPACT World and simulating it within a global "box." The decision to break with the former Excel/VBA IMPACT 2002 basis for building *Pangea* is motivated by the rigidity of these models, the fact that ESRI started to support Python as a scripting language for ArcGIS,[5] and the need for a more specialized numerical computing package for handling systems with up to a million compartments. The decision to break with IMPACT 2002 fate, transport, and exposure models is prompted by USEtox having become "*a scientific consensus model endorsed by the UNEP/SETAC Life Cycle Initiative for characterizing human and ecotoxicological impacts of chemicals*", while IMPACT 2002 derivatives are proprietary.

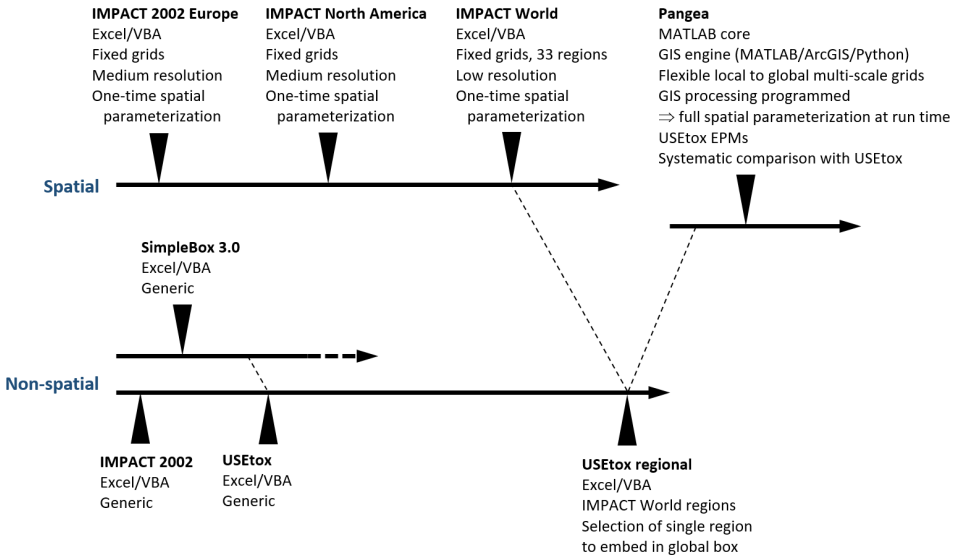


Figure 1.5 – Position of the *Pangea* framework in the evolution of the spatial and non-spatial multimedia fate and multi-pathways exposure models based on IMPACT 2002 and SimpleBox.

[5] More precisely, the availability of the ArcGIS toolbox through a Python "geoprocessor".

1.3 Objectives and specific aims

The main objective of this thesis is to ***determine and compare, within the same consistent methodological framework, local to global impacts of substances (chemicals, pollutants, etc.) that undergo a multimedia fate and lead to a population multi-pathway exposure.***

More specifically, in response to the aforementioned challenges and needs, we define the four following specific aims.

SPECIFIC AIM 1: Develop a flexible multi-scale modeling framework that models multimedia fate of substances (e.g. pollutants), the subsequent multi-pathway exposure of populations, and provides steady-state and dynamic solutions at local to global scales. More specifically, this aim entails: (1) Constructing a computational and GIS framework for modeling the fate and transport of chemicals, as well as human exposure. (2) Creating a system of grids for defining flexible global discrete geometries. (3) Implementing environmental models (EMs) and environmental process models (EPMs). (4) Developing an approach for building a mathematical compartmental system, and for solving equations of evolution.

SPECIFIC AIM 2: Evaluate the model and apply it to releases of chemicals used in HPC products for all of Asia. More specifically, this aim comprises: (1) Developing a spatial exposure modeling framework for HPC products, from market to environmental concentrations. (2) Applying the framework to compute spatial multimedia predicted environmental concentrations (PECs) of selected HPC chemicals across Asia. (3) Evaluating the framework by comparing PECs with USEtox results (model-model comparison with consensus model), and with monitored concentrations in fresh water and in sediments collected from literature. (4) Analyzing spatial variations of PECs across Asia and identifying key parameters/factors that are associated with high environmental concentrations.

SPECIFIC AIM 3: Study the radial distribution of iFs (inhalation, ingestion) of a selection of substances that are emitted from solid waste treatment plants (SWTPs) located in France. More specifically, this specific aim involves: (1) Presenting the structure and elements of the multi-scale *Pangea* model applied to multiple SWTPs. (2) Assessing the local to global fate and exposure associated with emissions of selected pollutants, analyzing the magnitude of the intake as a function of chemical properties and distance from a set of 126 point sources, for relevant exposure pathways. (3) Performing ~30,000 studies to build maps of global iFs associated with emission points that densely cover France.

SPECIFIC AIM 4: Study variations in populations exposure in Australia based on emissions defined by the NPI. Study populations exposure from an emitter perspective, per sector, and study the receptor perspective (source apportionment). More specifically, this aim entails: (1) Determining the population exposure resulting from the combined emissions of 4,101 point sources spread across Australia, identifying the main contributing sectors and the magnitude of their impact on human health. (2) Performing a source apportionment, identifying the main sources contributing to exposure at a given location, from a receptor perspective.

1.4 Framework development and previous contributions

Previous models have defined the needs and the rationale for developing a new framework. As discussed in Section 1.2.3 and illustrated in Figure 1.5, *Pangea* therefore departs from these models. The first aim of this thesis is to develop *Pangea*, from scratch, as a full substance fate and transport and human exposure modeling framework that benefits both from a recent (at the time we started) extension of GIS packages with programming languages (more specifically the availability of the full ArcGIS toolbox as a library for the Python language) and from the *Task Force on Toxic Impacts (TF LCIA 3) under the auspices of UNEP/SETAC Life Cycle Initiative* joint effort to produce the USEtox model. The most substantial external contribution to *Pangea* came from Peter Fantke, who collected and documented a selection of fate, transport, and human exposure models from both IMPACT 2002 and USEtox in order to facilitate their implementation in *Pangea*.[6]

1.5 Selection of case studies

The case studies presented in Chapters 3, 4, and 5 were selected to cover fate and transport of substances, populations exposure, from an emitter and receptor (source apportionment) perspectives.

Chapter 3 studies the fate and transport of HPC chemicals in Asia. This is an important case study, performed in collaboration with UNILEVER, because it allows for an evaluation of the fate and transport modeling aspect of *Pangea* against measurements based on spatial inventories of emissions and monitoring values. Such large-scale inventories are rare, especially because they were spatialized at the level of the *county*, which is high resolution in this context (2,852 counties for the sole China). Finally, this case study is interesting because it integrates *Pangea* into a large study oriented towards ecosystems, wherein UNILEVER is designing and funding current and future measurement campaigns in collaboration with Asian universities.

[6] The reasons for combining fate/transport/exposure models from both IMPACT 2002 and USEtox are that USEtox went through a long stabilization period (involving SimpleBox, another reference model), and that the full model specification/definition was only released in early 2017.

Chapter 4 extends the modeling of substances fate and transport with the modeling of subsequent population exposure. It radially summarizes spatial distributions of intakes (and iFs) associated with point sources located in France, with a focus on the evolution of the cumulative intake as a function of the distance from source, relative to the global intake. This study is specific to point sources and limited to a modeling result (including a Pangea/USEtox model-model comparison of global intakes) because there is no large-scale inventory of emissions or monitoring values.[7] The objective is to discuss the evolution of the intake as a function of the distance from source, as computed by a model of the category of USEtox, but spatialized, within the same framework (under the same systematic error). It is interesting because it helps to address a result by Lohman and Seigneur (2001), who have demonstrated that *up to 90% of polychlorinated dibenzo-p-dioxins and polychlorinated dibenzofurans (PCDD/F) emitted from stacks might be deposited farther than 100 km away from sources*. Finally, the second part of this study addresses the feasibility of studying large numbers of emission scenarios (10,000) in parallel, and compares distributions of global iFs through inhalation and ingestion computed with *Pangea*, with minimum and maximum values associated with urban and rural emissions computed with USEtox.

Chapter 5 extends previous studies to a large national inventory of emissions in Australia, under the usual emitter perspective and also under a receptor perspective (source apportionment). Australia is a notable because it is characterized by starkly contrasted regions in terms of population and environment, and also because of the availability of the NPI. This study "closes the loop" by demonstrating the feasibility of multimedia source apportionment despite the size of the underlying mathematical system.

1.6 Complementarity with other approaches

The objective of *Pangea* is to make a bridge between two types of approaches: generic, non-spatial evaluative models like USEtox, and models that are high resolution, local/regional, and principally atmospheric, such as AERMOD. We limit the modeling to first-order fate and transport processes[8] because the alternative would be global multi-physics, which is out of reach for a thesis (and more generally). This limits the maximum spatial resolution to scales and phenomena dominated by first-order processes.

[7] The first part of this study simulates emissions from real locations of SWTPs. The rationale for this is that we have an inventory of SWTPs (ISWA, 2012) that includes the tons of wastes treated per year (but not the emissions), and emissions per ton of waste for typical SWTPs have been published. While simulating estimates of true emissions would not enable us to compare simulation results with measurements (because SWTP are neither the only nor the major contributors to current environmental levels), it would enable us to perform an LCA type of comparative study, with a cost-benefit analysis that accounts for co-generation of heat and electricity.

[8] This is a modeling choice that is an absolute limitation. We currently also focus on linear processes, as this is the nature of fate and transport models in IMPACT 2002 and USEtox. However, this is not an absolute limitation.

The objective is hence to make *Pangea* capable of covering scales from the global down to the external spatial boundary of more specialized models (e.g. AERMOD) in order to allow for comparison (and possible calibration of *Pangea*) at the interface.

The case study on HPC chemicals in Asia is a fitting illustration of the applicability of *Pangea*, and it is complementarity with both non-spatial and more specialized models. Two options are currently available for studying impacts of HPC chemicals over ecosystems in Asia: a screening study based on non-spatial models like USEtox or SimpleBox, which informs about average behaviors and environmental levels, and a location-specific finer modeling of e.g. a segment of a river downstream from a treatment plant. The latter type of study cannot be performed everywhere at the scale of Asia (14 countries are involved in the study) to identify regions of interest and potential hotspots. It is also not able to cover multimedia transfers for chemicals such as D5, which are primarily emitted in media other than fresh water. In this context, *Pangea* offers an intermediary approach that is spatial and high enough resolution to account for local/regional variability. It offers insights regarding potential regions of interest and permits a selection for further, finer analysis. *Pangea* can therefore be instrumental for prioritizing finer studies.

Finally, non-spatial models estimate population exposure by exposing the whole population of large-scale or generic boxes to average environmental concentrations. By spatializing the approach, *Pangea* accounts for the spatial distribution of environmental concentrations, the spatial distribution of populations or more generally receptors, and antagonistic or synergistic effects between them, that former models cannot captured.

1.7 Outline of the dissertation

This dissertation is structured as follows: **Chapter 2** addresses Specific Aim 1, the model development. Introductory sections define the approach and provide insights into the developmental history. Further sections explain the general structure, as well as each major feature/mechanism separately (GIS engine, grids, grid refinement, the re-indexing engine, environmental and process models, and the computation engine). The chapter ends with a discussion of model limitations and a conclusion. **Chapter 3** addresses Specific Aim 2, the model applicability, which is evaluated in a study about HPC chemicals released in Asia. **Chapter 4** addresses Specific Aim 3, the local to global (radial) analysis of the iF, as it is associated with a set of 126 locations of SWTPs in France. **Chapter 5** addresses Specific Aim 4, the receptor perspective, by studying the variations in population exposure to multiple sector-specific sources from the NPI. **Chapter 6** concludes the paper and provides perspectives for the future.

Chapter 2 Modeling Framework Development

2.1 Introduction

This chapter and Specific Aim 1 develops a flexible multi-scale modeling framework called *Pangea*, which models multimedia fate of substances (e.g. pollutants) and the subsequent multi-pathway exposure of populations, and provides steady-state and dynamic solutions at local to global scales. It provides a sufficient level of detail for understanding the functioning of each *Pangea* component and sets the basis for the spatial analyses performed in subsequent chapters, while avoiding a lengthier, comprehensive, handbook-style explanation.

Pangea allows users to spatialize any set of linear environmental process models (EPMs) over the globe, spatially discretized by a set of multi-scale grids that cover relevant media, as illustrated in Figure 2.1. It is not a model, but a modeling framework. When associated with a set of EPMs and relevant data sets and parameters, it becomes a model specific to these inputs. *Pangea* is distributed with a default set of EPMs and with a selection of data sets and parameters, which define the "*Pangea* model" used in this thesis.

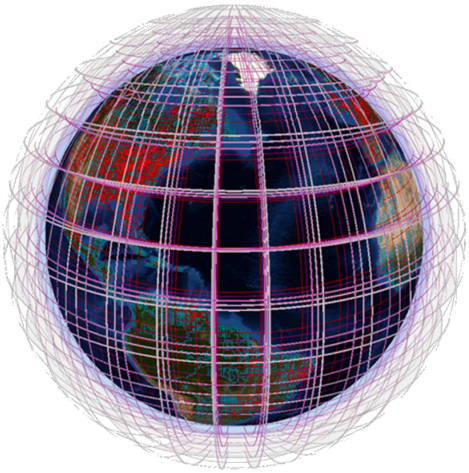


Figure 2.1 – Example of global multi-scale grid built by *Pangea*. Selection of a few layers from the 3D atmospheric grid: high-altitude fixed resolution layers (grey-pink), and highly refined ground-level layer (red).

More specifically, this chapter achieves the following: (1) defining the structure of the framework; (2) describing the GIS engine and the main grids present in *Pangea*, and explaining how grid refinement is implemented; (3) explaining the mechanism for passing from a geometric and geographic gridded world to a mathematical system that is better suited for expressing differential equations of fate and transport; (4) defining EMs and EPMs, and their respective roles; (5) describing the underlying mathematics and the computation engine; and (6) discussing key concepts, main limitations, and identifying conclusions.

Appendix B.1 provides the history of the developments and main challenges that have been addressed.

2.2 General structure

This section presents the structure of the *Pangea* framework from the point of view of a flow of operations between systems using engines. It concludes with an overview of most concepts and mechanisms developed in this chapter, illustrated with a focus on the atmospheric aspect of *Pangea*. This is essential because it visualizes a slice through the framework that explicitly contextualizes all constituting blocks of *Pangea,* which are explained separately in the following sections.

2.2.1 Systems, Engines, and Flow of Operations

Figure 2.2 represents the flow of operations in *Pangea*. Geo-referenced data that numerically describe the **natural system** are processed using the **GIS engine** (1). The output of this engine is a set of grids, gridded data sets, and geometric and topological parameters that form the **geometric system**. The geometric system represents the world, discretized by grids covering all relevant regions and media.

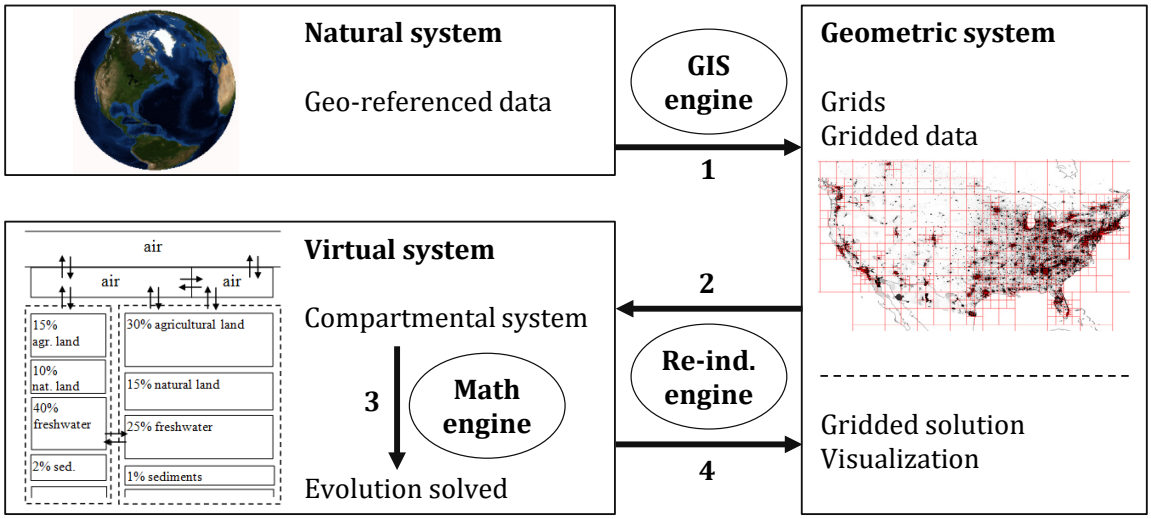


Figure 2.2 – Systems, main "engines", and flow of operations.

Grid cells can have inhomogeneous content; therefore, they are not suitable as compartments from a mathematical compartmental system perspective (that requires homogeneous content). In addition, depending on EPMs to build a set of (at least 70,000) coupled differential equations that iterates through grids/layers/cells/media and their connections to possibly many other grids/layers/cells/media and through contact surfaces areas with various mismatching geometries would make EPMs extremely complex. In view of this, a **re-indexing engine** (2) projects the geometric system into an abstract system called the **virtual system**.[9] The virtual system is a mathematical compartmental system; it is hence well

[9] It is abstract in the sense that the re-indexing engine transforms a system based on cells that have a geometric/geographic reality into a system of abstract homogeneous compartments (that represent fractions of cells defined by the proportion of each medium in each cell) connected through abstract contact surfaces.

suited for writing a set of differential equations that describe the evolution of environmental masses of substance, wherein masses are the dynamic variables associated with compartments. These equations are solved (3), and the solution is re-indexed back (4) to the geometrical system (onto the original grids) for visualization.

2.2.2 Framework components

Figure 2.3 illustrates the general structure of the framework. It depicts a *Pangea* **Core** component (Section 2.2.3) associated with a set of external groups of resources: EMs, EPMs, data sets, input/output processors, and tools.

Multiple **EMs** can be defined for each medium. The freshwater hydrological model, for example, can be based on the 0.5°×0.5° gridded World Water Development Report II (WWDRII, Section 2.8.3), or on a finer resolution, catchment-based HydroBASINS (Section 2.8.3). Multiple **EPMs** sets can be defined; for a given project, the default USEtox/SimpleBox-based EPMs set could be updated with an alternate sedimentation EPM, for example. The *Pangea*-based model that is ultimately used in a project is therefore project-specific and specific to all inputs (e.g. emissions and physical/chemical data), but also specific to the selection of EMs and EPMs. **Data pre-processors** are a set of tools for performing relevant operations on e.g. spatial time series (aggregation, statistics) or physical and chemical parameters, so they are well suited for *Pangea*. **Input/Output tools** essentially contain tools for importing data from chemical (e.g. parser for NPI or parser for EPISuite batch output) and tools for saving/exporting/displaying results and parameters, such as functions for building graphs, maps, and gridded maps, functions for saving data to relevant file formats, and functions for building simulation reports. Finally, the set of **tools** contains essentially the GIS engine and wrappers for external models (e.g. USEtox wrapper/connector).

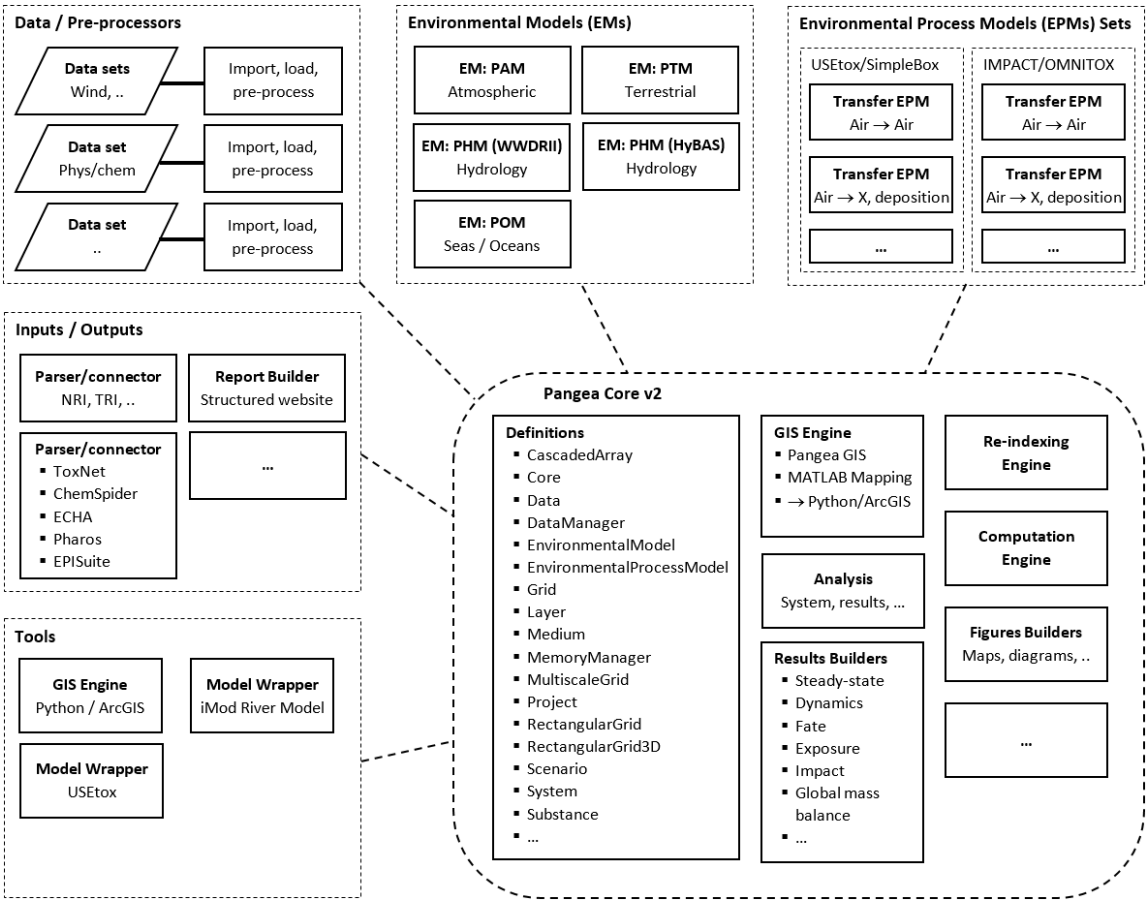


Figure 2.3 – General structure of the *Pangea* framework.

2.2.3 *Pangea* Core

Pangea Core is the set of resources that form the core of the framework. As Figure 2.3 indicates, it does not contain EMs, EPMs, data processors/connectors, or other tools, such as external models wrappers. *Pangea* is written in object-oriented (OOP) MATLAB. Its core is a MATLAB package (`Pangea`) that contains more than a hundred classes and six sub-packages. This currently represents more than 60,000 lines of code, which should be reduced during the next refactoring phase (to keep it "simple").

Core classes include for example *grids*, *layers*, and *media* (`Pangea.Grid`, `Pangea.RectangularGrid`, `Pangea.RectangularGrid3D`, `Pangea.MultiScaleGrid`, `Pangea.Layer`, `Pangea.Medium`), which can be seen as objects definitions, associated with tools for working with these objects (i.e. `Pangea.Grid` defines a generic type of grid and implements a set of relevant functions for working with generic grids).

The core provides mechanisms for managing models, projects, scenarios, and substances in a tree structure. A cascaded memory management system (Section 2.2.4) allows for the inheritance of objects, data, and parameters from upper levels down the tree (e.g. a grid or a data set defined at the project [top] level is inherited by all scenarios and all substances, which enables it to store it in memory only once).

2.2.4 Projects structure

Figure 2.4.A depicts the structure of projects: a tree with a top-level Model (object) that contains a series of projects. Each project contains a series of scenarios,[10] and each scenario contains a series of substances.

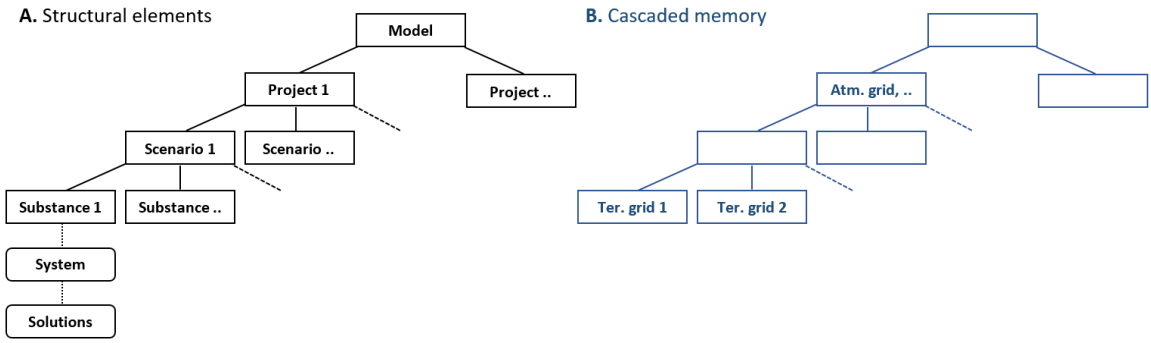


Figure 2.4 – Structure of projects (A), and corresponding cascaded memory (B).

In this structure, substances are called *end points*, and *Pangea* can build a compartmental system (compute a **K** matrix and define a source vector/array) and compute solutions for each end point.

Resources (grids, data, etc.) can be associated with each element of the tree, at each level. In most projects, however, some resources will be common (e.g. same atmospheric grid for all scenarios and all substances), while others will be specific (e.g. simple terrestrial grid for benzene, and finer version for PCBs). While common resources could be repeated/copied for each end point, it would be extremely resource intensive. This prompted the development of a cascaded memory manager that is parallel to the structural elements (projects, scenarios, substances), as shown in Figure 2.4.B (illustrating the above example with benzene and PCBs). The memory manager supports aggregating resources by branch (e.g. retrieving all grids defined above the current element into the tree), requesting first-defined elements and re-injecting elements in the structure at relevant places/levels.[11] This structure and mechanisms

[10] Scenarios and substances can be permuted in the structure. The mechanism covers two types of studies, e.g. the study of two technological options (scenarios first) with their specific lists of substances, and a Monte-Carlo analysis (as a series of scenarios) for each substance (substance first).

[11] The mechanism is too complex to fully develop in this thesis report; in short, a virtual system is built based on a defined set of grids, as the corresponding re-indexing engine (section 2.5). When grids are gathered along the branch of the first substance, *Pangea* can build and re-inject the re-indexing engine at the same level as where the

aimed at minimizing memory consumption are mandatory because *Pangea* is processing extensive data sets, performing operations with substantial GIS resources, and computing with large matrices.

2.2.5 Overview of components and mechanisms

In this section, we study a slice through the *Pangea* framework and its modeling chain, focusing on the atmospheric aspect, which illustrates a series of key concepts and mechanisms. This contributes to an understanding of the general structure of the framework and positions the material presented in the following sections.

Figure 2.5 presents the main components associated with atmospheric modeling: wind/precipitations data sets, a 3D multi-scale atmospheric grid, an atmospheric environmental model (EM) that describes the dynamics of the medium (air), and environmental process models (EPMs) that describe the dynamics of substances and output blocks of the **K** matrix, the matrix of transfer and elimination rate coefficients.

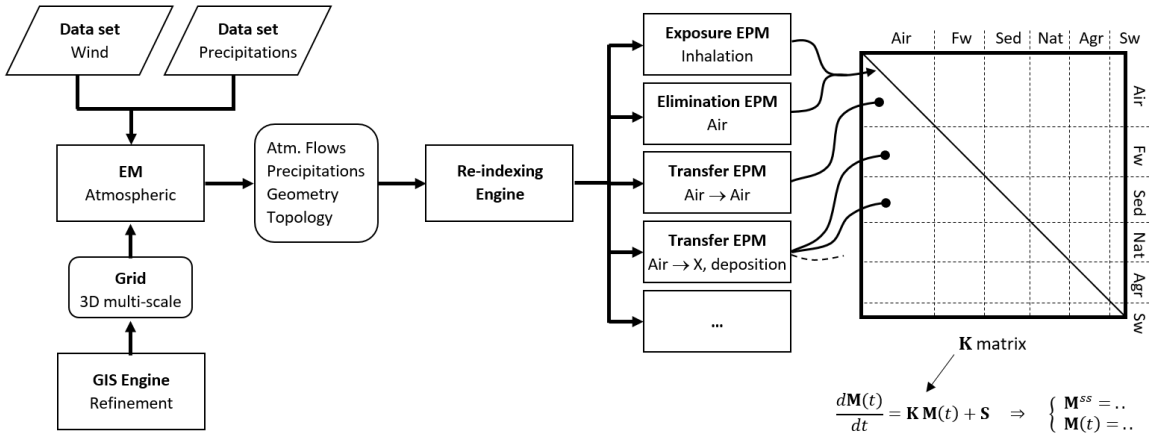


Figure 2.5 – Main blocks of *Pangea* associated with atmospheric modeling.

Starting from the left, the **atmospheric environmental model**[12] **(EM)** takes wind and precipitation data sets and a GIS-generated 3D multi-scale grid as inputs. The atmospheric EM computes 3D interpolation of wind velocities (generally planar, in u and v directions, by layer), horizontal flows through 3D grid cells walls, and vertical flows by continuity. It also supports pre-processing time series of wind velocities and computing directional monthly/yearly averages,[13] pre-processing time series of rain events, and building the 3D spherical topology.

bottom-most grid was found. If all grids are defined at the project level, this mechanism allows *Pangea* to build a single re-indexing engine (one-shot operation, and single instance in memory) for all scenarios and all substances.

[12] Referred to as PAM, for *Pangea* Atmospheric Model.

[13] This is performed keeping u^+, u^-, v^+, v^- separate, to account for bidirectional transfers, instead of cancelling them through averaging positive and negative values.

The **wind data set** provides time series of gridded u/v velocities for multiple global atmospheric layers. The default wind data set is a time series of six-hourly data as defined by GEOS-4 on 55 global layers with a 2°×2.5° grid resolution[14]. The atmospheric EM can work with any resolution, and a discussion of the validity of the data set and its resolution must be conducted on a per-project basis. *Pangea* uses two default precipitation data sets: a default, trivial set of constants that characterize rain events as defined in USEtox, and a gridded data set based on the NASA Tropical Rainfall Measuring Mission (TRMM).

The atmospheric EM principally outputs atmospheric flows between compartments of the atmospheric grid and geometric and topological parameters. These parameters are expressed in a grid-specific indexing (≅ numbering) schema and must be transferred to a more global/abstract/unified indexing schema called *virtual system*, which is better suited for computation (defining and solving a set of coupled differential equations). This is performed by a **re-indexing engine**,[15] which sorts re-indexed vectors and matrices by medium, giving them a medium-based block structure.

The output of the re-indexing engine is appropriate for parameterizing **EPMs**, which describe the dynamics of substances and their transfers between compartments.[16] This is done by block, which means that the transfer EPM air→fresh water (deposition), for example, receives a vector of air volume that is the "air" block of the vector of all volumes, a matrix of contact surfaces areas that is the "air/fresh water" block of the matrix of all contact surfaces areas, etc. Environmental process models associated with atmospheric modeling account for advection, diffusion, degradation, wet and dry deposition, population exposure as an elimination process, etc. Environmental process models output blocks of the **K** matrix, the matrix of transfer and fate rate coefficients, which parameterizes the set of linear ordinary differential equations that defines the evolution of the compartmental system. The **K** matrix has a medium-based block structure: the EPM air→fresh water (deposition), for example, computes the block in row 2, column 1 of the block **K** matrix, which describes the transfer between the atmospheric and freshwater compartments.

[14] Pangea can use GEOS-Chem 0.25° × 0.3125° (or lower resolution) three-hourly meteorological data since they are available as NetCDF-4 files. We keep the 2° × 2.5° data set as default because it was used as a basis for parameterizing other models (e.g. IMPACT World and USEtox), which simplifies models comparison.

[15] The (main) re-indexing engine is a complex machinery that accounts for fractions of contact surfaces area between grids with different, mismatching geometries, for fractions of media in heterogeneous cell contents, and for the nature of parameters (intensive or extensive). There are simpler/smaller re-indexing engines within multi-layer grids, which manage transitions between layer- and grid-indexing schemas.

[16] Environmental models describe the dynamics (or any other parameter) of media, while EPMs describe dynamics of substances (based on the dynamics of media).

This short introduction has illustrated key concepts and components, as well as the general flow of operations:

1. Multi-scale grids are created using the GIS engine.
2. Spatial data sets, which describe media and/or their dynamics, are processed by EMs.
3. Grid-specific outputs of EMs are re-indexed to medium-based block versions in the virtual system.
4. Parameters in this system (and other, non-spatial parameters, e.g. physico-chemical parameters) are used by EPMs for computing blocks of the **K** matrix.
5. The **K** matrix is used to define a matrix linear ordinary differential equation, which describes the evolution of the compartmental system.

2.3 GIS Engine

The GIS engine is a set of GIS resources integrated into a common framework that allows *Pangea* to process geo-referenced data in real time, to build project-specific grids and geometries, and to project spatial data sets onto these grids. It is technically one of the most challenging parts of *Pangea* due to the complexity of GIS processing. In practice, GIS processing is an art, especially when dealing with large and complex geo-referenced data sets. Generally, one must iteratively understand why a series of standard operations is failing, and subsequently test alternative approaches. Failures can be crashes (easy to detect and manage), but also inappropriate outputs, and there is often no algorithmic way to detect the latter; it can only be detected by the human eye, experience and intelligence, and the perception that "something does not look right" (which cannot be programmed easily, as illustrated in Appendix E.1).[17] This occurs in particular when GIS processing involves large data sets using complex geometries to globally combine data sets that are not fully global, or not global at all.[18] Problems arise especially when dealing with multi-scale geometries that combine large and small polygons (with orders of magnitude of difference in scales), where small, valid, high-resolution polygons are comparable in size, or even smaller than parasitic polygons that are generated by gaps, overlays, and operations with tolerance, and which arise when large polygons with (even slightly) mismatching geometries are involved in the processing.

[17] This comes from the wide variety of spatial references and projections (with their limitations and singularities), from the incompleteness of geo-referenced data sets, from the presence of tolerances and approximations at all levels, from the fact that faulty or singular geometries (easily generated) may not work, but may not crash either, from limitations of GIS tools, etc.

[18] If anything can fail anywhere because of singularities, it **will** fail in a global model.

Geographic information system tools and operations are generally slow as a result of the implementation of versatile methods in GIS libraries. Calling GIS tools from the Python/ArcGIS library is not a light operation, if only because it must establish a connection to a license server and validate licenses for ArcGIS and its toolboxes. On the MATLAB side, a mapping toolbox is available that provides tools for manipulating GIS objects and building maps. However, this toolbox is not a GIS, and it lacks most of the common GIS tools.[19] Other MATLAB libraries and other GIS packages (e.g. Quantum GIS) exist, and these offer various tools with their limitations.

This led to the **GIS engin**e being implemented in *Pangea* as a **cascaded** set of **wrappers for GIS libraries**, including a *Pangea* GIS library that I developed in pure MATLAB, which provides GIS tools specific to *Pangea* (e.g. grid refinement, zonal statistics optimized for rectangular grids). The GIS engine provides top-level functions that cascade to relevant libraries based on the nature of their inputs. Figure 2.6 illustrates part of the cascade associated with a top-level `refineGrid` method/function from the *Pangea* GIS library. The function checks whether the grid is rectangular or "generic". If it is rectangular, it delegates refinement to specialized functions of *Pangea* GIS which are implemented in MATLAB and are optimized for rectangular grids. If the grid is not rectangular, or if all refinement methods for rectangular grids fail, *Pangea* GIS delegates refinement to the slower but more versatile and robust Python/ArcGIS library, which also implements a series of refinement functions.

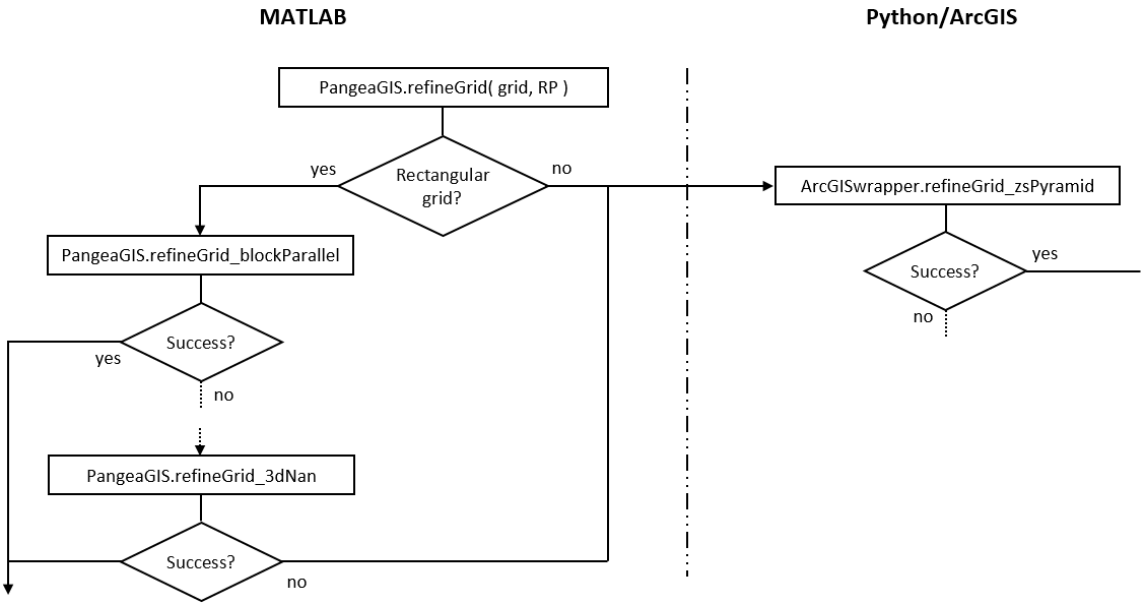


Figure 2.6 – GIS engine cascade of libraries for grid refinement.

[19] Performing a basic "intersect" between two polygon feature classes, for example, is not trivial.

This cascaded approach facilitates adding new libraries in a fashion that is transparent for most users (including EMs) who consistently use top-level functions.

2.4 Grids structure and the geometric system

2.4.1 Introduction

There are several types of **grids** in *Pangea* that are mainly used for spatially discretizing all relevant media, as well as for display or analytical purposes. All grid types are built on and extend the same fundamental type/class of generic grid with the following properties[20]:

- Name, identifier (ID)
- List of layers
- List of media
- Structure array, as an $n^{cells} \times n^{media}$ matrix of proportion of each medium in each cell

Layers can be seen as GIS polygon feature classes, i.e. as sets of polygons defined by the coordinates of their nodes. These polygons are called **cells** and have a geometric and a geographic reality. Each **medium** has a unique ID that can be associated with grids. When multiple media are associated with a grid, a **structure** array must define the proportion of each medium inside each grid cell. This array is automatically a column vector of 1s for single-medium grids, and computed based on a land-cover data set for the terrestrial grid for example.

Based on these simple definitions, the atmospheric grid contains a list of 17 layers (for 17 atmospheric layers by default), a single air medium, and a trivial $n^{cells} \times 1$ structure array of 1s. In contrast, the terrestrial grid contains a single layer, a list of three media (fresh water, natural land, and agricultural land), and a $n^{cells} \times 3$ structure array of proportions defined by the land cover data set.

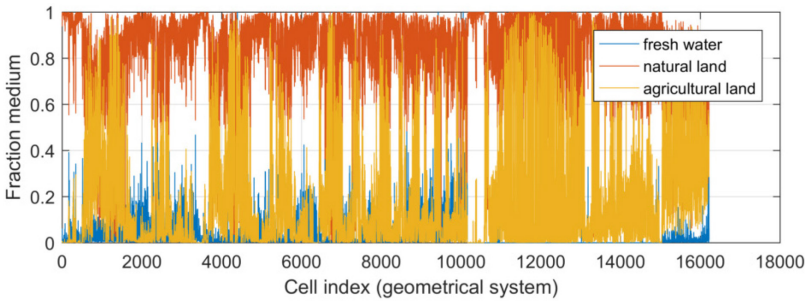


Figure 2.7 – Elements of the structure array of a terrestrial grid.

[20] Using an OOP terminology, all grids are subclasses of a `Pangea.Grid` superclass, which provides the most common properties and methods specific to grids.

Figure 2.7 presents the structure array of a terrestrial grid, plotting the proportion of each of the three media as a function of the cell ID (x axis) for a grid with ~16,200 terrestrial grid cells.

Other, more complex grids contain an internal re-indexing engine that enables passage from a "whole grid" indexing schema to layer-specific indexing schema and vice versa, as well as other, more advanced features.

This section describes the most common grids in *Pangea*. Multi-scale grids are built by grid refinement, starting from a low-resolution background grid. They are used as a basis for building the *results* grid – the grid on which most results are ultimately projected/re-gridded for comparison – and the atmospheric grid. The terrestrial grid involves another type of refinement that is constrained by the hydrological model. The (freshwater) sediments grid is built as an image (piggy-back) of the terrestrial grids, limited to cells that contain fresh water.

Finally, we define the **geometric system** as the system of all grids and their connections. It has a geometric/geographic reality: grid cells delineate regions of the globe that have a physical reality. However, as mentioned, it is too complex to serve as a basis for defining a mathematical compartmental system, as its components (grid cells) can delineate regions with heterogeneous content that are connected by mismatching, complex geometries.

2.4.2 Special grid types/classes

Generic grids can be refined into multi-scale grids using Python/ArcGIS and a GIS-based refinement algorithm from *Pangea* v1 (Section 2.5), which iterates zonal statistics and grid creation through polygons splitting. This approach is highly versatile but very slow. *Pangea* v2 implements a series of special grids that greatly accelerate GIS processing. The simplest example is a class of rectangular grids for which optimized, MATLAB-only algorithms benefitting from simpler and more specific geometry can be used for grid refinement. This class contains 2D rectangular grids that are made of rectangular, non-disjointed polygons (that form a partition of the geographic coordinates), as well as 3D rectangular grids that stack 2D rectangular grids (as layers) and incorporate an internal re-indexing engine for facilitating the passage from layers to the 3D grid indexing. A final type of grid is a class of multi-scale grids, which store a main 2D multi-scale "final" rectangular grid as well as all refinement steps from a low-resolution background grid.

To illustrate, the 3D atmospheric grid is a 3D rectangular grid built based on a multi-scale grid and all its refinement steps, which are stacked vertically with a decreasing resolution.

2.4.3 The Background Grid

The background grid is a low-resolution global grid that defines the starting point/resolution for the refinement algorithm. Any grid can be used, but *Pangea* creates by default a regular grid (as shown in Figure 2.8), resolution of which can be defined by the user (e.g. by defining the number of rows and columns global).

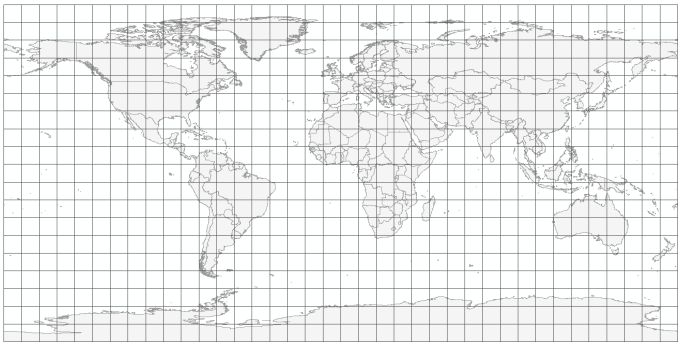


Figure 2.8 – Example of background grid.

2.4.4 The Results Grid

The *results* grid is a grid onto which all results are ultimately projected/regridded, essentially for comparison. Any grid can be used, but the most common choice is to define a multi-scale grid that is used both as a *results* grid and as the first layer of the atmospheric grid. Figure 2.9 provides an example over Europe, which is the *results* grid used in Chapter 4.

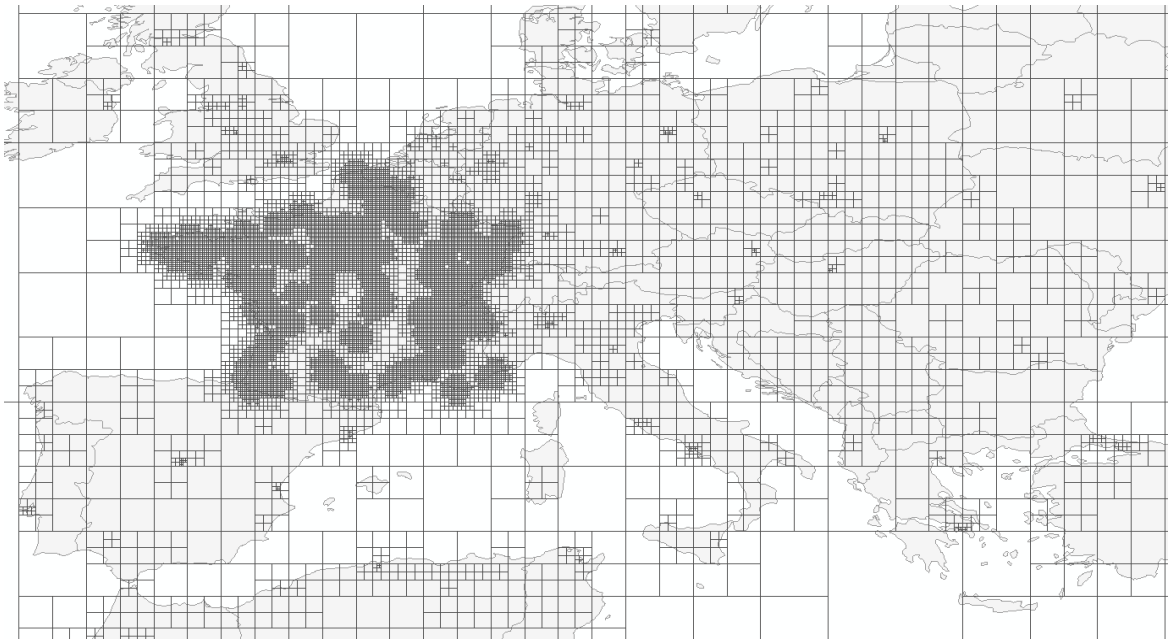


Figure 2.9 – Example of *results* grid.

2.4.5 The Atmospheric Grid

The atmospheric grid is built based on a multi-scale grid with all intermediary steps of the grid refinement procedure (including the initial, background grid). The objective is to create a 3D multi-scale grid, with maximum refinement at ground level (usually where emitters and receptors are located) and a "smoothly" decreasing resolution with altitude.

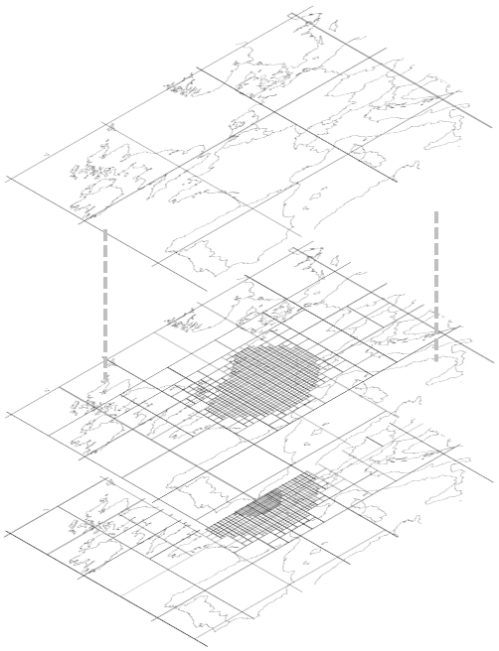


Figure 2.10 – Example of atmospheric grid.

The default atmospheric grid is based on the default 55 GEOS-Chem layers (or 72 when using GEOS-FP) or any subset of them, usually chosen to cover altitudes ranging from about 100m to 16km from the ground. However, any number of layers and altitudes can be chosen independently of GEOS-Chem, as the wind data set is ultimately interpolated in 3D for computing air flows associated with any rectangular geometry.

By default, *Pangea* builds the vertical geometry by stacking a multi-scale grid and all its refinement steps vertically with a decreasing resolution, and by repeating the background grid as needed for higher altitudes.

2.4.6 The Terrestrial Grid

The terrestrial grid is a global "grid" of polygons limited to land. It is defined by the hydrological model, as polygons have to respect the watershed structure of the hydrological network. It is called a grid because the first hydrological model implemented (WWDRII) is a global gridded network, but it can be composed of polygons with any geometries (e.g. HydroBASINS catchments). The terrestrial grid is multi-scale, but to a lesser extent than the atmospheric grid. Grid refinement (or aggregation) algorithms can up-scale or down-scale the resolution, but they are constrained by the maximum resolution and geometry of the underlying hydrological model.

Current terrestrial grids are built using a geometry defined by WWDRII, but we are transitioning towards HydroBASINS and its finer geometry, for the reasons developed in Chapter 3.

Terrestrial grid based on WWDRII – The current, default terrestrial grid is based on the Word Water Development Report II (WWDRII; Vörösmarty, 2002), a global gridded $0.5° \times 0.5°$ water network.

The native grid, shown in Figure 2.11, defines ~65,000 cells, and is thus too large for most projects (generally regional or continental)[21]. *Pangea* aggregates this native grid into clusters, respecting the watershed structure of the hydrological network. This procedure is weakly constrained by the refinement potential (RP) (section 2.5)

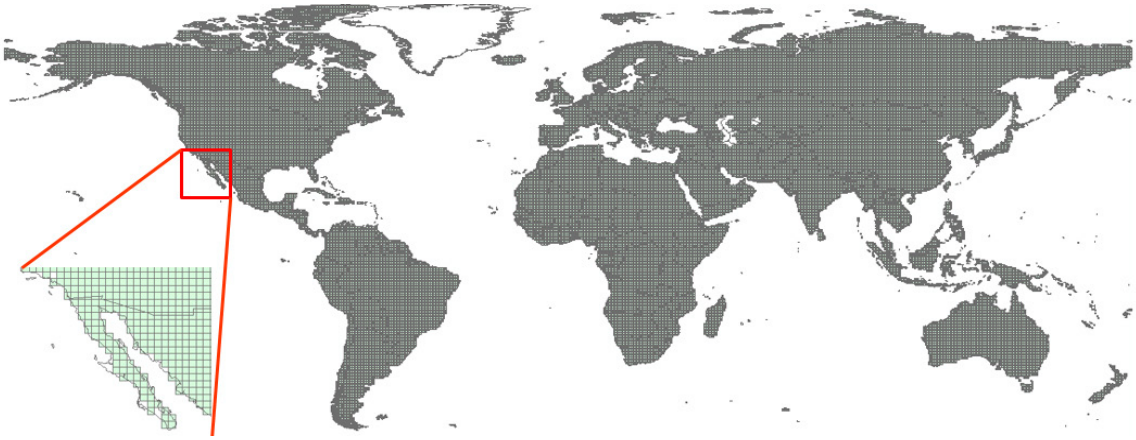


Figure 2.11 – Native WWDRII global 0.5° × 0.5° grid.

Figure 2.12 presents the outcome of the clustering procedure. Even constrained by the RP, the cluster resolution is generally not high enough over specific regions of interest. As a final step, *Pangea* therefore un-clusters the grid over regions of interest.

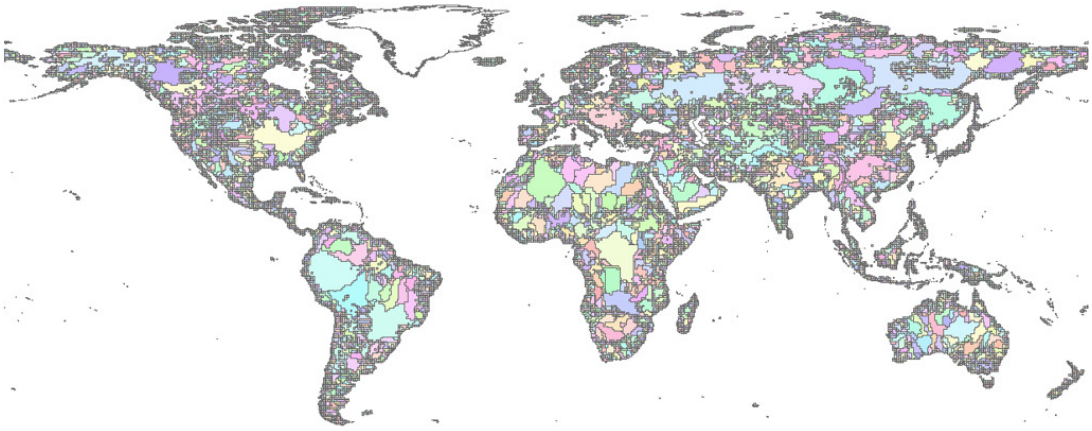


Figure 2.12 – Example of clustered terrestrial grid.

Figure 2.13 illustrates this, showing the terrestrial grid used in Chapter 3, where the region of interest encompassed Asia (extended) and Australia, and the terrestrial grid was un-clustered accordingly.

[21] *Pangea* builds between 3 (default) and 22 compartments per cell depending the aggregation of land cover categories. This means ~200,000 compartments just for the terrestrial grid in the case of maximal aggregation of land cover categories (best case).

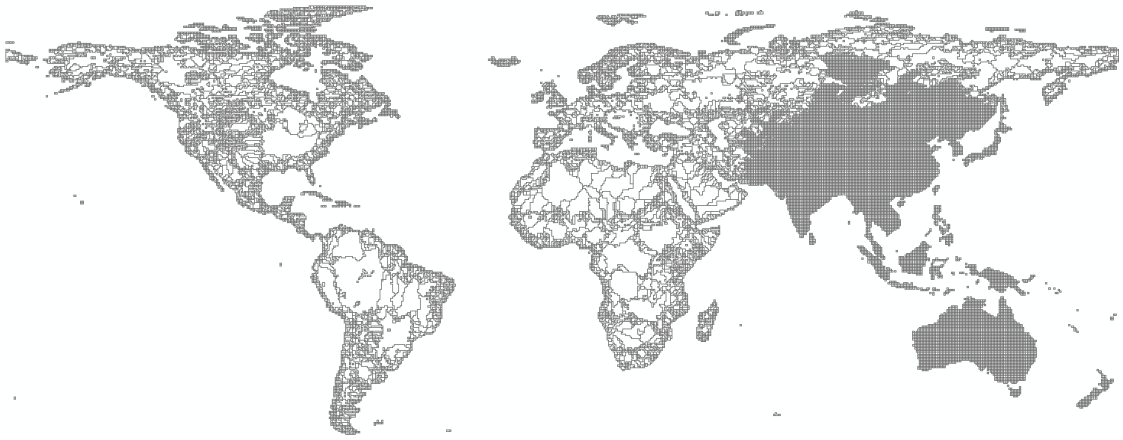


Figure 2.13 – Example of clustered terrestrial grid, un-clustered over a region of interest (here Asia and Australia).

Terrestrial grid based on HydroBASINS – A new hydrological model based on HydroBASINS is evaluated and used as a comparison element in Chapter 3. HydroBASINS is global and has 12 levels of resolution that are defined by polygons, the boundaries of which match natural catchments. Figure 2.14 illustrates levels 4, 6, 8, and 10, and displays the WWDRII native grid for comparison. Combining polygons from multiple levels allows us to define multi-scale grids. This data set has the advantage of coming with other resources, such as HydroATLAS (data associated with catchments) and HydroLAKES (lakes as polygons, and pour points).

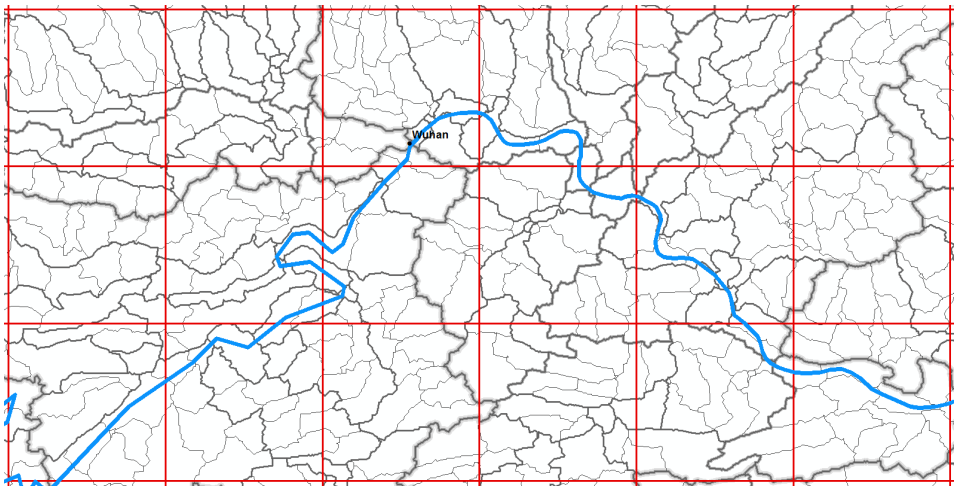


Figure 2.14 – Levels 4, 6, 8, and 10 of HydroBASINS (polygons with grey boundaries) along the Yangtze river (blue polyline). The WWDRII 0.5° × 0.5° grid (red) is overlaid for comparison.

2.4.7 The Sediments Grid

The sediments grid is built to match the terrestrial grid, but limited to terrestrial grid cells that contain fresh water. The reason for this is that the presence of sediment grid cells with no sediment (because there is no fresh water) would generate singularities in the mathematical system.

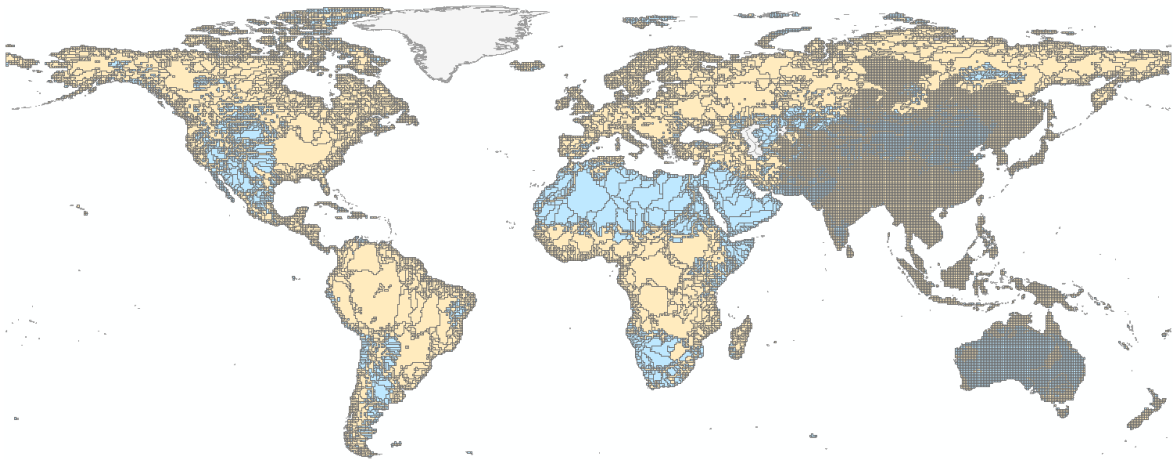


Figure 2.15 – Example of sediments grid (beige) overlaying the terrestrial grid (blue). Holes in the sediments grid indicate regions where there is no water defined by WWDRII.

Figure 2.15 highlights the sediment grid in beige, overlaying the terrestrial grid in blue. Blue regions are regions where there are not fresh water sediments defined because WWDRII does not define any fresh water. One reason for updating the hydrological model to the version based on HydroBASINS is that it would reduce the extent of regions with no fresh water and no sediments. Even if there is a relatively high uncertainty on HydroBASINS hydrological parameters in these regions, this would at least allow *Pangea* to build compartments that correspond to these regions.

2.4.8 The Seas/oceans Grid

For the studies presented in this thesis, the oceans grid was a single polygon, defined as the world complement of lands. The rational for this is given in Section 2.8.5.

2.5 Grid refinement methods and Multi-scale grids

2.5.1 Introduction

Multi-scale grids are built using a refinement algorithm. One of the first technical questions that had to be addressed is: how to identify an iterative refinement procedure where high resolution is needed? The retained solution has been to define a global scalar field, called the *refinement potential* (RP), that defines the "local need for high resolution" at each point on the globe. This was a first step: integrating this field over polygons would provide an integral (or "sum") per polygon, and we could have each polygon whose integral is above a given threshold refined (e.g. split in quads). This iterative process would then be repeated.

This unconstrained process (iterative, independently of the geometry) works for the atmospheric grid, but not for grids or media with geometrical constraints.

This section provides the rationale behind needs for refinement, defines the RP, and discusses refinement methods associated with major grids in *Pangea* (air, WWDRII hydrology, HydroBASINS hydrology, oceans).

2.5.2 Needs for refinement

There are four main reasons for grid refinement, with some overlap:

1. To reduce artifacts presented in illustrated in Figure 1.2 to Figure 1.4, in particular artificial instantaneous dilution (full mixing) and non-discrimination of sources and receptors.
2. To account for spatial variability around emitters and/or receptors, and to focus on regions of interest.
3. To meet more advanced criteria based on a first output, e.g. to limit the relative importance of removal in each cell with respect to the total output, or to target regions with the highest gradients in concentrations.
4. To account for specific geometries when the media associated with transport or measurements are highly localized.

The first example of the third point refers to a study by Warren et al. (2009), who have defined a criterion for identifying cases in which the well-mixed box assumption can deviate significantly. This is briefly discussed at the end of Section 2.5.4.

A typical situation for illustrating the fourth point is the hydrology, when two important streams that belong to completely distinct basins get close to each other. We have an example in Switzerland, where the Rhine River and the Rhone River are both sourced from neighboring regions of the Alps. The Rhine

River progresses north through highly populated regions of Germany and the Netherlands, and eventually it flows into the North Sea. Meanwhile, the Rhone River runs south through France and eventually flows into the Mediterranean Sea. Using a rough grid geometry over this region of the Alps could distribute an emission of pollutants that effectively enters the Rhone River partly, or even fully, through the Rhine River. Similarly, measurements made in a specific river can be associated with another river, as defined by a rough hydrological model, due to a routing algorithm that displaces rivers according to a prioritization schema that may not be compatible with local needs.

2.5.3 Refinement potential

The RP mechanism addresses the question of how to define a 2D or 3D multi-scale grid based on study-specific refinement criteria. The RP is defined as a "global 2D scalar field whose value at each location of the globe defines the modeler's interest for having a high resolution at this point." This definition is general, but in practice, the RP is often defined by a raster (a discretized version of the scalar field), i.e. a geo-referenced "image" whose pixels' values define the intensity of the potential at each pixel's center point. Such "discrete RP" is often imposed by the presence of data-based components that are directly available as rasters (e.g. raster of population density), and not as mathematical functions.

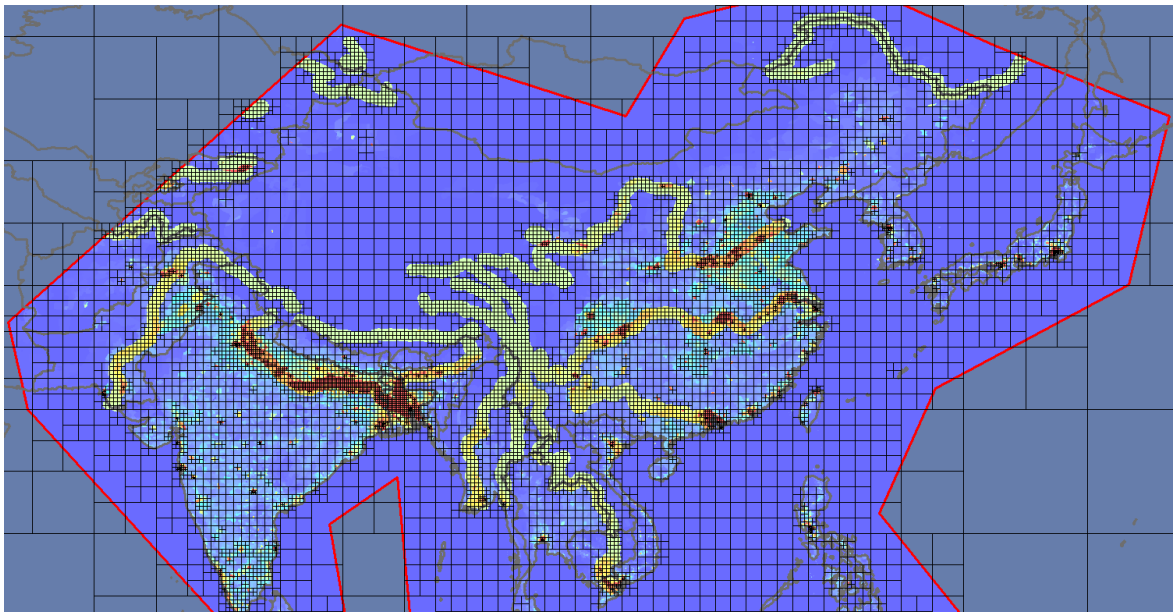


Figure 2.16 – Example of refinement potential (RP) and multi-scale grid.

The RP can also be defined by a mathematical expression that involves of a series of components, such as the population density (refinement where high density of receptors) and the distance from source(s) (refinement decreases with the distance from source). A simple, effective illustration of the mathematical expression is a basic weighted sum of components, with user-defined weights that signify the importance of each component.[22] Figure 2.16 illustrates an RP over Asia as a raster with dark blue (low value) to dark red (high value) pixels. A polygon with red boundaries indicates a "region of interest" component, combining a population component as well as a component that targets the proximity to large rivers. The final grid is overlaid in black, and we see that the higher the value of the potential, the higher the resolution.

Table 2.1 – List of most usual components of refinement potentials.

Component	Type	Description
Region of interest	Vector	Generally, a set of polygons (e.g. shapefile) that delineate one or more regions of interest.
Population counts	Raster	E.g. CIESIN global raster of population counts estimates for 2010.
Flag "in land"	Vector	Two multi-parts polygons: seas (value 0) and land (value 1).
Emissions	Vector/raster	Point feature class for point sources, or raster of emission intensities for diffuse emissions, or any combination.
Distance to feature	Vector/raster	Often "multiple rings buffer" based on any geometrical feature, giving values to rings (e.g. decreasing with ring index).
$\sin(\emptyset + \frac{\pi}{2})$ with $\emptyset$ the latitude	Mathematical	Compensates for the fact that the background grid (usually a regular grid in a space of latitudes and longitudes) defines cells whose area on ground decreases with latitudes moving to the poles. The RP is multiplied by a function whose value varies between 1 at the equator and 0 at the poles.
$\alpha e^{-\beta\lvert\vec{x}-\vec{x}_0\rvert}$	Mathematical	Exponential decreasing with the distance from point in $\vec{x}_0$. Often use in conjunction with a "point" feature class of point source(s)
Functions of output of previous run, e.g. based on components of $\mathbf{K}$ or $\overrightarrow{\mathrm{m}}_{ss}$	Vector	Outcome of a former simulation, expressed on relevant grid(s), constraining the refinement for the current simulation (iterative approach).

Three types of components are supported: raster components, vector components, and mathematical functions. A raster of population counts or density is a good example of a raster component. The region

[22] Users of GIS can also see the process as raster algebra between raster components, and vector and mathematical components converted to raster.

of interest in Figure 2.16 illustrates a vector component (defined as a polygon). Finally, an exponential decrease with the distance from a point source is a simple example of mathematical component. Any geo-referenced data set or mathematical function can hence be used as RP components, but the most common components are listed in Table 2.1.

Pangea provides mechanisms for converting vector and mathematical components to raster, for extending regional rasters to the globe, and for normalizing components. Expressions like that in Eq. 2.1 are basic and common:

$$\mathbf{RP} = \delta_{land} \times \delta_{RoI} \times \left(\mathrm{d}_{rivers} + \sum_{i=1}^{n_s} \alpha_i e^{\beta_i |\vec{\mathrm{x}} - \vec{\mathrm{x}}_{0,i}|} \right) \tag{2.1}$$

where all variables/components are global rasters, δ_{land} is a flag "in land" whose pixels are in $\{0,1\}$, δ_{RoI} is a flag "in region of interest" whose pixels are in $\{0,1\}$, d_{rivers} is a discretized function that decreases with the distance from rivers,[23] and each of the $\alpha_i e^{\beta_i |\vec{\mathrm{x}} - \vec{\mathrm{x}}_{0,i}|}$ is a function that decreases with the distance from one of n_s point sources. In this example, we see that the two flags act as abrupt antagonistic components that cancel the potential outside of lands and outside of the region of interest, whereas other components act synergistically.

This mechanism controls where high resolution occurs, and it therefore allows for retaining the numerical size of the restricted system. While it is not the focus of this thesis, controlling numerical size becomes crucial when simulating multiple substances simultaneously and when solving the dynamics.

Finally, given the limited time required for running projects and rebuilding geometries in *Pangea* v2 (1 to 10 minutes), it is possible to follow an iterative approach for updating the RP as a function of initial runs. A relevant example is associated with the criterion defined in Warren et al. (2009): "*if over 25% of the chemical entering a box is removed, applying* [the] *well-mixed assumption can lead to substantial error*". They advise diminishing the size of boxes until this criterion is met. Figure 2.17 visualizes the percent output removed by elimination, for atmospheric compartments. The computation involves elements of the **K** matrix associated with a first system.[24] The criterion is fulfilled, but if it were not, cells corresponding to compartments above 25% could be flagged for refinement until the criterion is met.

[23] In practice, this is often achieved using a "multiple-rings buffer" (from the Python/ArcGIS based GIS-library), associating values to rings, and converting to raster.
[24] Resulting from a first grid generation through refinement based on a first RP, data projection, etc.

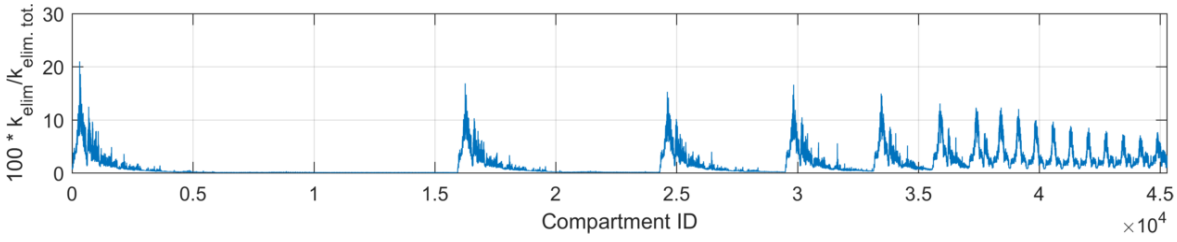


Figure 2.17 – Percent of output removed by elimination, for atmospheric compartments.

2.5.4 Base refinement procedure

The refinement algorithm consists in integrating the RP over a low-resolution background grid, refining all cells with an integrated RP above a user-defined threshold, and iterating using the new grid for integration. Figure 2.18 illustrates this process. When the RP is defined as a raster, the integration becomes a zonal statistic. The refinement potential, the background grid, the threshold, factors for scale transitions (i.e. refine by splitting in 2×2 sub-cells), and the refinement maximum depth are user-defined and project-specific.

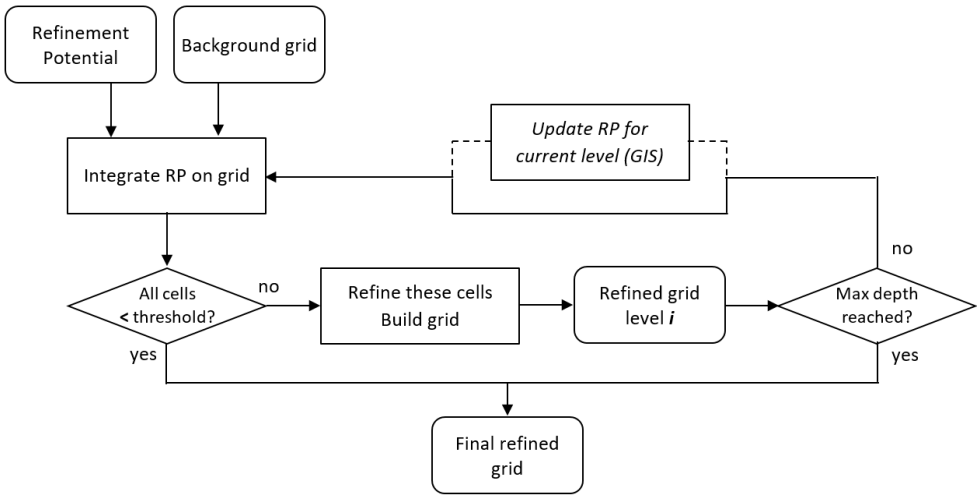


Figure 2.18 – Base grid refinement algorithm.

Two refinement implementations are available. The *Pangea v1* implementation, based on Python/ArcGIS, is versatile but very slow; it can refine most types of grids and update the RP as a function of the refinement depth, thus accounting for certain geo-referenced features only at given scales (e.g. account for roads when grid cells are finer than 20 km × 20 km). This approach is not maintained because we never used its full power and because it is extremely time consuming. Being GIS-based and using slow operations (generic zonal statistics with polygon feature class as zone definition, and high

resolution RP raster), it can take a few hours per grid. The *Pangea v2* implementation is specific to rectangular grids and optimized for this; it is MATLAB-based, uses a fast zonal statistics algorithm, and takes generally less than 10 seconds per grid.

Figure 2.19 illustrates this procedure, which is applied to an RP that is defined by the raster of population counts only. The background grid is shown at depth 0, and we see the iterations though the loop of Figure 2.18, with increasing depths and maximum resolution. Cells stop being refined progressively because the integral of the RP over their extent is below threshold, but some cells are refined until maximum depth is reached. Setting the maximum depth allows for setting the maximum resolution (in degree latitude and longitude).

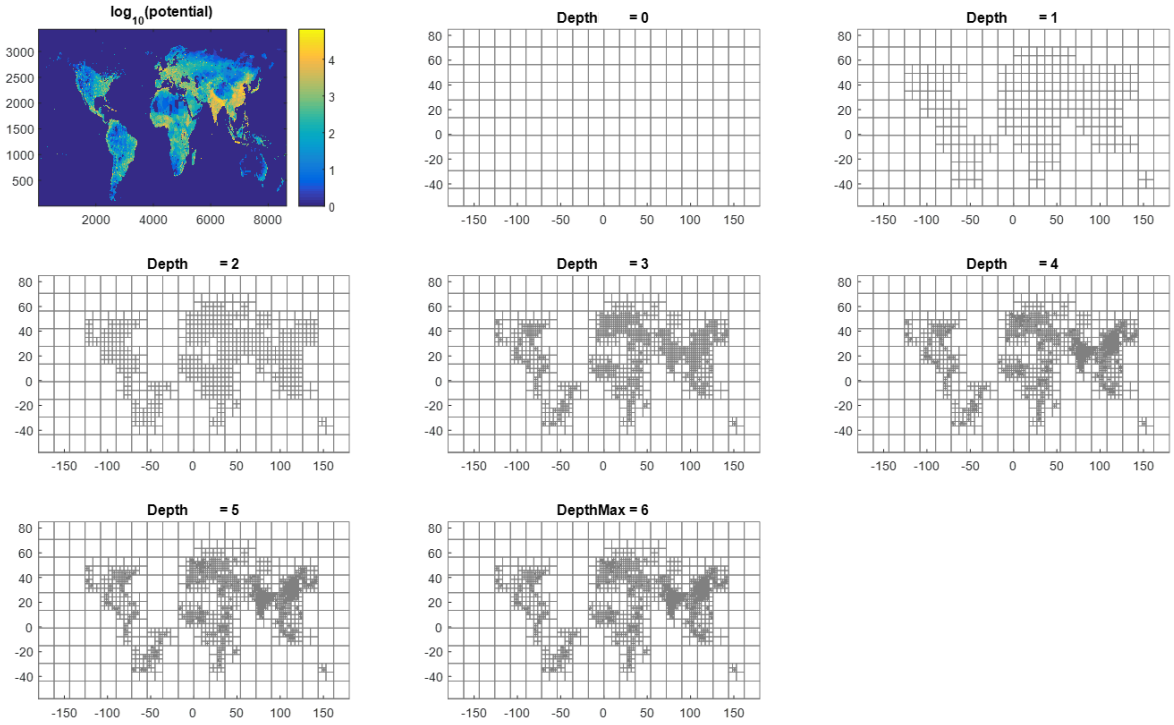


Figure 2.19 – Refinement procedure applied a raster of population counts as RP.

2.5.5 Refinement of the WWDRII native grid

The WWDRII native grid, shown in Figure 2.11, is provided at the maximal 0.5°×0.5° resolution of this model/data set. It is not possible to refine it in a fashion that is consistent with the underlying hydrology, but it is also not practical to keep this grid at its maximal resolution globally. Therefore, *Pangea* performs a clustering of this grid based on its topology (respecting its watershed structure), and travel time. The RP weakly constrains this clustering operation, leading to clusters comparable to those in Figure 2.12. *Pangea* can then "explode" clusters that touch the region of interest, as illustrated in

Figure 2.13. A full description of the *Pangea* hydrological model based on WWDRII and the tools that it provides (e.g. clustering) is outside of the scope of my thesis report, especially since *Pangea* is transitioning towards the HydroBASINS hydrological model.

2.5.6 Refinement of HydroBASINS catchments

HydroBASINS is distributed as a set of catchments defined at 10 scales. *Pangea* implements a (test) algorithm for selecting polygons of scales 4, 6, 8, or 10 according to the RP, and building a partition (non-disjoint) of the land out of it. Figure 2.20 illustrates the outcome of this operation, showing a background RP focusing (blue pixels = low potential, to red pixels = high potential) on a segment of the Yangtze River and highly populated areas. The grid built by *Pangea* based on HydroBASINS levels 6 and 8 displays level 8 (finer) polygons/catchments (black boundaries) that cover regions with high potential, and level 6 (coarser) polygons/catchments (brown boundaries) elsewhere.

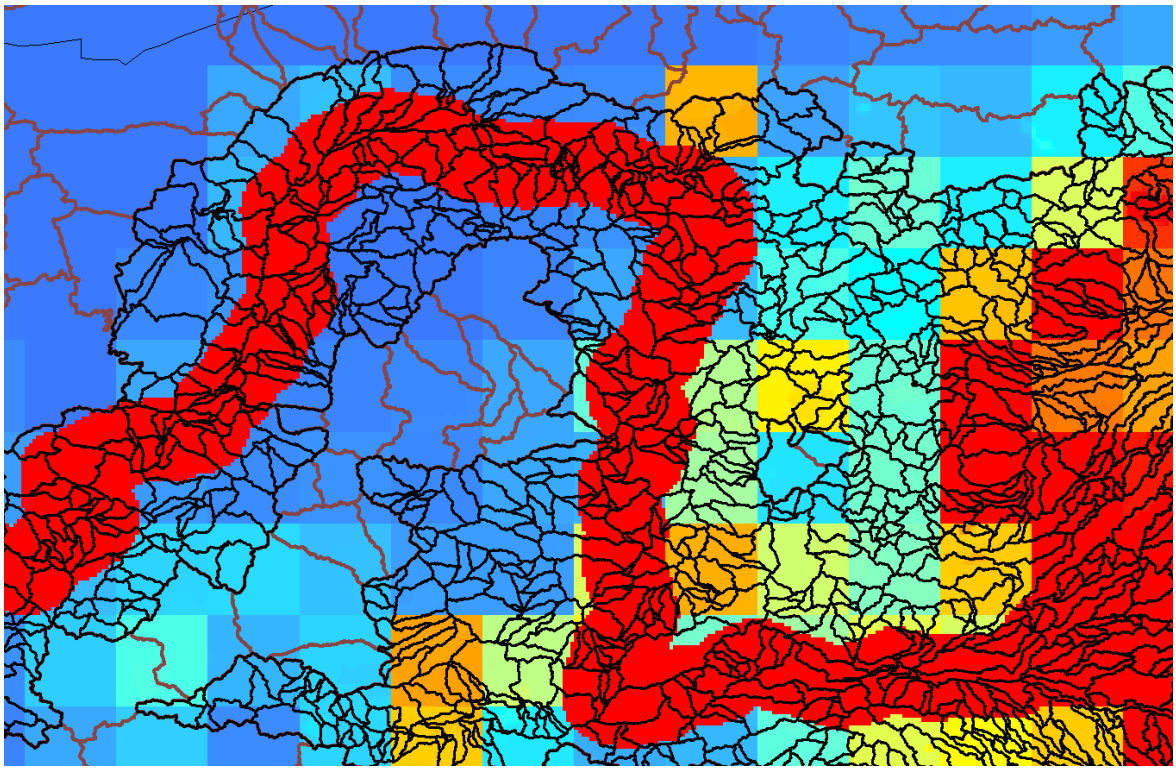


Figure 2.20 – Refinement procedure based on HydroBASINS, levels 6 and 8, and a refinement potential.

The HydroBASINS-based *Pangea* hydrological model is in an evaluation phase, and it serves as a comparison element in Chapter 3 to examine whether it can potentially address some of the shortcomings of WWDRII.

2.6 Re-indexing Engine and Virtual System

2.6.1 Description

For modeling the fate and transport of chemical substances, *Pangea* must build a mass-balance equation, as presented below in Equation 2.4 (section 2.7.2). This equation defines the evolution of a mathematical compartmental system, a system of compartments characterized by a single dynamical variable per compartment. In our context, it translates into the necessity for each compartment to represent the volume of a single, homogeneous medium. This describes an abstract situation that is often represented by a diagram (Figure 2.21).

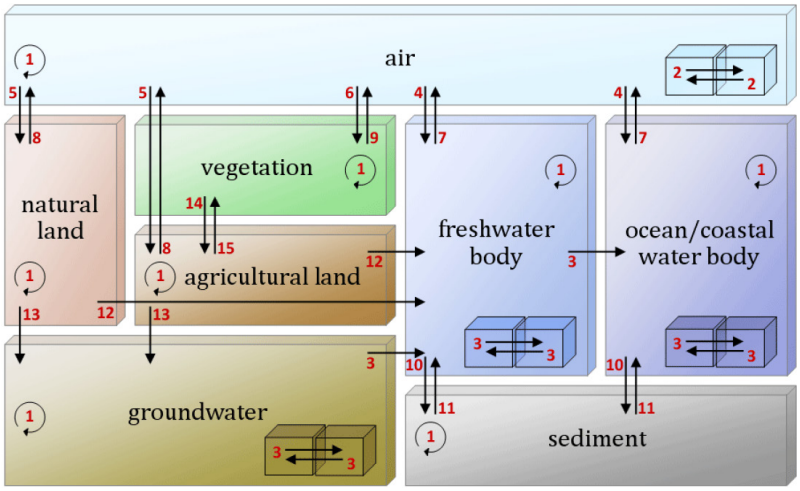


Figure 2.21 – Non-spatial, abstract compartmental system. Compartments represent homogeneous media.

However, the geometric system, with mismatching grid geometries covering complex heterogeneous terrestrial regions, is not as simple as this abstract system. Grid cells can delineate regions with inhomogeneous content, as demonstrated in Figure 2.22, which represents the interface between the first layer of an atmospheric grid (rectangular with black edges), a terrestrial grid based on HydroBASINS (irregular polygons with white boundaries), and the underlying raster of land cover, the pixels of which code for 22 types of land cover (Section 2.8.4).

This situation highlights complex and irregular connections (contact surfaces) between atmospheric cells with homogeneous content (air) and terrestrial cells with inhomogeneous content (22 types of land cover, which Pangea aggregates into fresh water, natural land, and agricultural land by default).

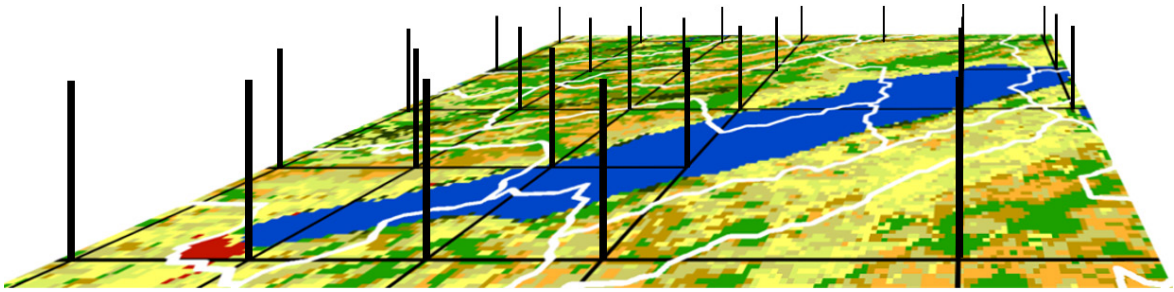


Figure 2.22 – Non-spatial, abstract compartmental system. Compartments represent homogeneous media.

The geometric system and its grid cells are thus not suitable as a set of compartments from a mathematical compartmental system point of view. A **re-indexing engine** therefore projects elements of the geometric system (grids, geometric and topological parameters, spatial data) into an abstract system called the **virtual system**.[25] The virtual system is a mathematical compartmental system, and is hence well suited for writing a set of differential equations that describe the evolution of environmental masses of substance, wherein masses are the dynamic variables associated with compartments.

This warrants an important terminological definition specific to Pangea:

- **Cells** are grid cells, or elements of the geometric system. They have a geometric and geographical reality, and they can have a heterogeneous content in terms of media.
- **Compartments** are elements of the virtual system. They are abstract; a compartment may represent the 20% agricultural component of a cell, which is spread arbitrarily across the cell. Compartments are connected through abstract contact surfaces areas, and they have a homogeneous content.

The re-indexing engine is the tool that allows for passage from one to the other in both directions.

2.6.2 The re-indexing engine

Each spatial parameter that describes a grid cell property (e.g. volume) or a projected data set (e.g. population per grid cell) is expressed as a vector, the size of which equals the number of cells of the grid. Similarly, each parameter that describes interactions/connections between two grids (e.g. advection between atmospheric grid cells or deposition from atmosphere to freshwater grid cells) is expressed as a matrix, with a size defined by the size of the grid(s) involved (e.g. $n_{atmospheric\ cells} \times n_{terretrial\ cells}$). These grid-specific vectors and matrices characterize (or are expressed in) the geometric system. To

[25] It is abstract in the sense that the re-indexing engine transforms a system based on cells that have a geometric/geographic reality into a system of abstract homogeneous compartments (that represent fractions of cells defined by the proportion of each medium in each cell) connected through abstract contact surfaces.

efficiently parameterize a global mass balance equation based on generic EPMs,[26] we perform a re-indexing operation that splits vectors and matrices from the geometric system according to the proportion of each medium in each grid cell and distributes them into the virtual system. The virtual system has a block structure (for vectors/matrices) that is organized by medium type, as seen in Figure 2.23. Let δ be any a priori grid-specific parameter, such as the cells' base surface areas; they are initially defined in the geometric system (and/or computed by the GIS) as vectors $\vec{\delta}^{atm,L1}$ to $\vec{\delta}^{atm,L17}$ associated with the 17 layers of the 3D atmospheric grid, and $\vec{\delta}^{ter}$ associated with the terrestrial grid (composed of fresh water, natural land, and agricultural land). Re-indexing operations (blue arrows) distribute components of these vectors into relevant blocks of $\vec{\delta}^{v}$, the vector of all base surfaces areas in the virtual system, according to the proportion of the media that compose each grid cell (i.e. 1:1 to air for the atmospheric grid, and according the land cover composition for the terrestrial grid).

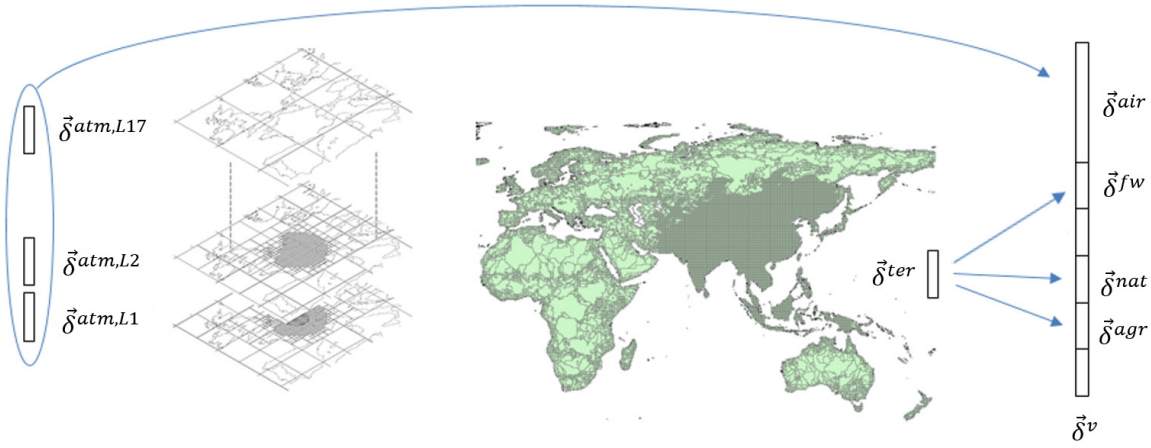


Figure 2.23 – Re-indexing: geometric system to virtual system.

Each element of vectors and matrices of the virtual system is associated with a single medium (vectors) or with a single pair of media (matrices). The size of the virtual system is noted n^v, which corresponds to the total number of compartments in the system ($n^v = n^v_{air} + n^v_{fresh\ water} + n^v_{sediments} + ..$).

Mathematically, a simplistic but useful explanation of re-indexing is to present it as a matrix multiplication:

[26] Environmental process models cannot implement the complex task of determining connections (topology and geometry) relevant to their computations. That would make their adaptation from e.g. USEtox to *Pangea* a daunting task. *Pangea* moved this complexity outside of the EPMs, keeping EPMs "generic": instead of computing one transfer rate coefficient from fresh water to air based on one contact surface area (as implemented in non-spatial models), they need to be able to output a matrix of coefficients matching the size of their input, when passed a matrix of contact surfaces areas. This is generic in the sense that they work with abstract areas, as they do not need to know the location or true geometry of these areas.

$$\vec{\mathrm{x}}^{v} = \boldsymbol{\mathcal{R}}^{G1\to v}\mathrm{x}^{G1} \text{ for vectors, and } \mathbf{M}^{v} = \boldsymbol{\mathcal{R}}^{G2\to v}\,\mathbf{M}^{G1,G2}\;\;{}^{t}\boldsymbol{\mathcal{R}}^{G1\to v} \text{ for matrices} \tag{2.1}$$

For the vector case, $\vec{\mathrm{x}}^{G1} \in \mathbb{R}^{n^{G1}\times 1}$ is a vector[27] associated with grid #1 (numbered for the purpose of the matrix example), $\boldsymbol{\mathcal{R}}^{G1\to v} \in \mathbb{R}^{n^{v}\times n^{G1}}$ is a re-indexing matrix[28] from grid #1 to the virtual system, and $\vec{\mathrm{x}}^{v} \in \mathbb{R}^{n^{v}\times 1}$ is the re-indexed, virtual version of x^{G1}, split by medium and distributed appropriately. For the matrix case, $\mathbf{M}^{G1,G2} \in \mathbb{R}^{n^{G2}\times n^{G1}}$ is a matrix associated with grids #1 and #2, $\boldsymbol{\mathcal{R}}^{G1\to v} \in \mathbb{R}^{n^{v}\times n^{G1}}$ is a re-indexing matrix from grid #1 to the virtual system and its transpose ${}^{t}\boldsymbol{\mathcal{R}}^{G1\to v}$, $\boldsymbol{\mathcal{R}}^{G2\to v} \in \mathbb{R}^{n^{v}\times n^{G2}}$ is a re-indexing matrix from grid #2 to the virtual system, and $\mathbf{M}^{v} \in \mathbb{R}^{n^{v}\times 1}$ is the re-indexed, virtual version of $\mathbf{M}^{G1,G2}$, split by medium and distributed appropriately.

If $\mathbf{M}^{G1,G2}$ is the matrix of contact surfaces areas between grid #1 and grid #2, for example, $\mathbf{M}^{v}$ is the matrix of abstract contact areas between all compartments associated with all cells and media of grid #1, and all compartments associated with all cells and media of grid #2.

This last example illustrates well that a large part of *Pangea*'s complexity lies in the construction of re-indexing matrices, which contain and translate mismatching geometries and heterogeneous cell contents into their homogeneous components based on grids structure (array of proportions of media for each cell).

In practice, re-indexing is more complex because extensive and intensive quantities are not re-indexed the same way. In addition, one often needs to associate a quantity with a specific medium. For example, the depth of fresh water is a vector associated with the terrestrial grid, but its components should not be distributed into all terrestrial media components; rather, they should be "dealt" to fresh water only, using a medium-restricted re-indexing matrix. Re-indexing back is also more complex than it looks; given $\vec{\mathrm{m}}_{ss}^{v}$, the vector of environmental masses of substance at steady-state expressed in the virtual system, a direct back-re-indexing onto the terrestrial grid would compute a weighted sum of masses associated with all terrestrial media, which would be meaningless. Finally, re-indexing operations can be associated with grid layers instead of full grids. The population count of people breathing at ground level is associated with layer #1 of the atmospheric grid, and must be re-indexed from the layer (part of the grid) to the virtual system. This means that instead of having a limited set of re-indexing matrices $\{\boldsymbol{\mathcal{R}}^{G1\to v}, \boldsymbol{\mathcal{R}}^{G2\to v}, \dots\}$, re-indexing requires a much larger set $\left\{\boldsymbol{\mathcal{R}}_{intensive}^{G1\to v}, \boldsymbol{\mathcal{R}}_{extensive}^{G1\to v}, \boldsymbol{\mathcal{R}}_{deal}^{G1\to v}, \boldsymbol{\mathcal{R}}_{intensive}^{G2\to v}, \dots\right\}$, as well as mechanisms for building back-re-indexing

[27] This works for $\mathrm{x}^{G1} \in \mathbb{R}^{n^{G1}\times n}$ for an arbitrary n.
[28] Numerically, these matrices are very large and they are stored as sparse matrices.

matrices and mechanisms for restricting these matrices by layer and/or by medium, e.g. $\boldsymbol{\mathcal{R}}_{intensive}^{G2,L1\to v}$ or $\boldsymbol{\mathcal{R}}_{intensive}^{G1,L3,fw\to v}$.

In summary, the re-indexing engine is the module of *Pangea* in charge of building re-indexing matrices and managing all declinations of re-indexing operations: back and forth, with restriction by medium and/or by layer, for vectors and matrices, and for extensive and intensive parameters.

Finally, the re-indexing engine manages re-gridding operations, which, like re-indexing, are a function of the nature of the parameters to be regridded: intensive or extensive.

2.6.3 Structure of the Virtual System

The virtual system has a block structure by medium. Elements of vectors of this system are associated with specific media, and elements of square matrices with pairs of specific media. Elements are associated with compartments, and compartments represent homogeneous components in terms of the media that compose grid cells. There is therefore a relationship between compartment IDs, grid IDs, layer IDs, cell IDs, and media IDs. Figure 2.25 reports part of the output of the re-indexing engine self-analysis tool associated with a typical project characterized by a system made of 99,149 compartments with blocks sizes given in Figure 2.24. Sub-plots of Figure 2.25 display relevant IDs (e.g. grid ID, cell ID) as a function of the compartment ID for all compartments present in the system. The boundaries of the blocks by medium are overlaid in red. The "Sys. Grid ID" plot demonstrates that the whole "air" block is associated with grid #1 (terrestrial), the "freshwater" block with grid #2 (terrestrial), the "sediments" block with grid #3 (sediments), and then the two "land" media with grid #2 again, as both are defined on the terrestrial grid. The "Sys. Sediments ID" plot trivially shows a monotonous increase in the medium ID. The "Grid layer ID" plot indicates an increase in layers from 1 to 17 in the atmospheric grid, and that all other grids are single layer. The "Grid cell ID" and "Layer cell ID" plots exhibit the evolution of the ID of each cell in the grid "total" structure and index schema, and within each layer indexing schema respectively.[29] Sudden increases in these IDs in all three blocks associated with terrestrial media (most visible in the "freshwater" block) highlights the absence of each medium in blocks of cells with contiguous IDs.

General
Virtual sys. size : 99149

Virtual system structure

Medium	Size	Index Range
air	45274	1 - 45274
fresh water	11329	45275 - 56603
sediments	11329	56604 - 67932
natural land	16212	67933 - 84144
agricultural land	15004	84145 - 99148
sea water	1	99149 - 99149

Figure 2.24 – Virtual system blocks sizes

[29] This illustrates what multi-layers "small" internal re-indexing mechanisms deal with.

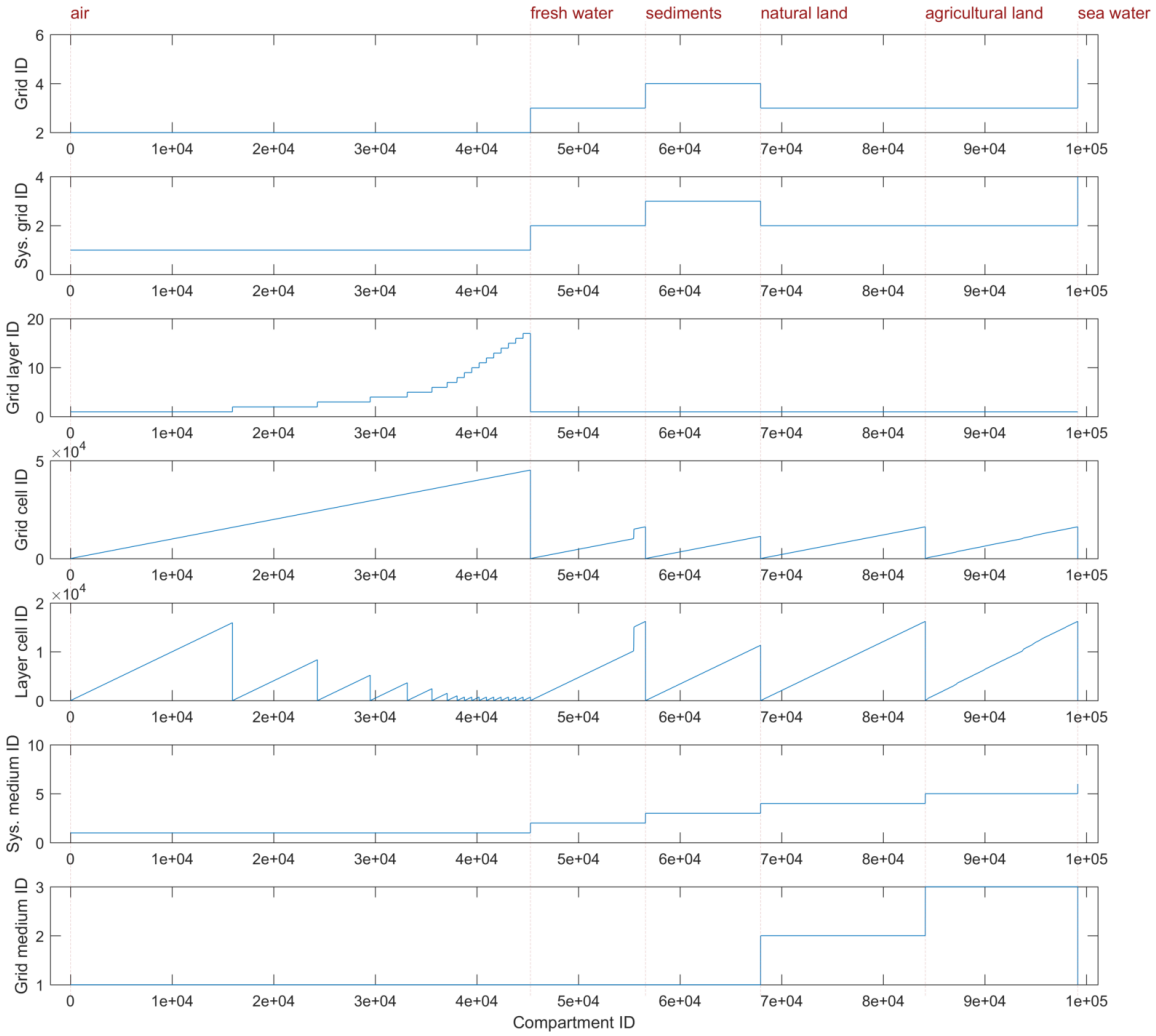


Figure 2.25 – Structure of the virtual system and relationship with media and geometric objects.

Analyzing this output is crucial for developing a better understanding of the relationships between a variety of indexing schemas and objects (grids, layers, cells, and media). It is also the basis for analyzing and debugging singular cases by back-tracking values from a gridded output back to the virtual system, and then back again to original gridded data sets when necessary.

2.7 Compartmental System, fate, and exposure

2.7.1 Introduction

This section builds on the theory of linear compartmental systems presented in Appendix C and extends it in a way that is better suited for large systems. Appendix C shows that a system of coupled linear differential equations can be expressed in a matrix form (Eq. C.11) as:

$$\frac{d\vec{\mathrm{m}}(t)}{dt} = \mathbf{K}\,\vec{\mathrm{m}}(t) + \vec{\mathrm{s}} \tag{2.2}$$

with

$$\vec{\mathrm{s}} = \begin{pmatrix} \mathrm{s}_1 \\ \vdots \\ \mathrm{s}_\mathrm{n} \end{pmatrix}, \quad \text{a constant source vector}$$

$$\vec{\mathrm{m}}(t) = \begin{pmatrix} \mathrm{m}_1(t) \\ \vdots \\ \mathrm{m}_\mathrm{n}(t) \end{pmatrix}, \quad \text{a time-varying mass vector}$$

and where **K** is a matrix of rate coefficients with diagonal elements defined by:

$$\mathrm{k}_{ii} = -\mathrm{k}_{i,deg} - \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{k}_{ji} \tag{2.3}$$

Non-diagonal elements of **K** are positive or null, and represent inputs into compartments. Element k_{23}, for example, is the transfer coefficient from compartment 3 to compartment 2; it appears in Equation C.11 multiplied by $\mathrm{m}_3(t)$ as input to compartment 2. Diagonal elements are negative[30] and represent total elimination as a sum of degradation and outputs to all other compartments. Element k_{44}, for example, is minus the sum of the degradation in compartment 4 and transfer coefficients from compartment 4 to all other compartments. It appears in Equation 2.2 multiplied by $\mathrm{m}_4(t)$, which defines the total output/elimination from compartment 4.

2.7.2 Fate and transport

Pangea builds on Equation 2.2 with constant coefficients **K** and $\vec{\mathrm{s}}$ when the focus is on the steady-state solution. We leave systems with time-dependent coefficients (Eq. C.16) aside in what follows, but most of the material holds for that case.[31] In Equation 2.2, elimination (degradation, exposure) and

[30] They must even be strictly negative, as a zero on the diagonal means no elimination and would lead to a divergent solution.

[31] The difference is that we do not compute and analyze the steady-state solution (that does not exist) when coefficients are not constant, but rather the dynamic solution computed with a numerical solver.

transfer processes (transport to connected compartments) are defined through their rate coefficients, which are stored in the **K** matrix.

Emission scenarios – *Pangea* extends Equation 2.2 by supporting matrices of emissions and masses, defined as arrays of vectors written in column, that represent emission scenarios (for the same system defined by **K**) and corresponding masses. Explicitly, we note n^{es} the number of emission scenarios, and we define $\mathbf{S}^v = \left(\vec{s}_1^v, .., \vec{s}_{n^{es}}^v\right)$ and $\mathbf{M}^v(t) = \left(\vec{m}_1^v(t), .., \vec{m}_{n^{es}}^v(t)\right)$, where $\vec{s}_i^v$ is the emission vector (distribution) for emission scenario i, and $\vec{m}_i^v(t)$ is the corresponding vector (distribution) of mass at time t.

Mass balance equation – Given the properties of the matrix multiplication, we can directly rewrite Equation 2.2 as follows:

$$\frac{\mathrm{d}\mathbf{M}^v(t)}{\mathrm{d}t} = \mathbf{K}\,\mathbf{M}^v(t) + \mathbf{S}^v \tag{2.4}$$

with $\mathbf{S}^v \in \mathbb{R}^{n^v \times n^{es}}$ a matrix (array of vectors) of constant emission scenarios [kg s^{-1}] written in column, $\mathbf{M}^v(t) \in \mathbb{R}^{n^v \times n^{es}}$ the corresponding matrix of masses [kg] at time t [s] written in column, and $\mathbf{K} \in \mathbb{R}^{n^v \times n^v}$ a matrix of transfer and elimination rate coefficients [s^{-1}]. **K** is a sparse matrix, with dimensions typically in the range of 70,000 × 70,000 to 500,000 × 500,000; it can be factorized, and the solving of Equation 2.4 can be parallelized in cases where n^{es} is large (e.g. 10,000, as in the study presented in Chapter 4).

Construction of the K matrix – While being mathematically simple and easy to solve numerically, Equation. 2.4 is complex to build, because components of **K** depend on most spatial parameters (e.g. contact surface areas between mismatching cells geometries, volumes, projected spatial data sets), EMs, and EPMs that are specific to each configuration of grids. The consequent work (several person-months when done manually) is usually performed only once in traditional models with fixed grids. The GIS and computation engines integrated in *Pangea* instead provide a solution for addressing this challenging task dynamically, at runtime, with a re-projection of all spatial data (e.g. land cover, demographic data, atmospheric flows, flows through the freshwater network) each time a new grid is generated. Based on spatial parameters and projected data, environmental non-spatial parameters, and physicochemical parameters, a set of EPMs is used to compute the components of the **K** matrix. The EPMs are specific to each medium for elimination processes (degradation), to each pair of media for transfer processes (advection and diffusion), and to each exposure pathway (inhalation and all ingestion routes). They are currently based on IMPACT 2002 (Margni et al., 2004; Pennington et al., 2005) and USEtox (Rosenbaum et al., 2008), and are adapted to take spatial data into account.

Solution of the fate and transport – The solution of Equation 2.4 is the distribution of pollutant mass at steady-state $\mathbf{M}_{ss}^{v}$ (Section 2.10.3), or as a function of time $\mathbf{M}^{v}(t)$ when we solve the dynamics (section 2.10.4). These solutions are back-re-indexed to relevant grids to obtain e.g. $\mathbf{M}_{ss}^{atm,L1}$, the spatial distribution of masses in the first layer of the atmospheric grid, or $\mathbf{M}_{ss}^{fw}$, the spatial distribution of masses in the fresh water network.

2.7.3 Population Exposure

Exposure pathways considered by default are inhalation and ingestion of freshwater and food (fish, meat, milk, unexposed produce/root crops, and exposed produce). The population intake through both inhalation and ingestion is computed and defined as:

$$\mathbf{IN}_{ss}^{v} = f\left(\vec{\mathrm{p}}^{v}, \mathbf{IR}^{v}, \mathbf{BAF}^{v}, \mathbf{C}_{ss}^{v}\right) \tag{2.5}$$

where $\mathbf{C}_{ss}^{v}$ is the array of environmental concentrations corresponding to $\mathbf{M}_{ss}^{v}$, $\mathbf{BAF}^{v}$ is an array of generalized bioaccumulation factors (BAFs) that can be interpreted as an application from environmental concentrations to concentrations in air, water, and food items, $\mathbf{IR}^{v}$ is an array of generalized individual intake rates, that can be interpreted as an application from concentrations in air, water, and food items, to masses taken in, $\vec{\mathrm{p}}^{v}$ is a vector of population counts, f is an appropriate product between its arguments, and $\mathbf{IN}_{ss}^{v} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ is a 3D array of population intakes [kg s^{-1}] with n^{ep} as the number of exposure pathways (inhalation, drinking, ingestion per food item category). Finally, the population iF is defined as:

$$\left(\mathbf{iF}_{ss}^{v}\right)_{ijk} = \left(\sum_{r=1}^{n^{v}} \mathbf{S}_{rj}^{v}\right)^{-1} \left(\mathbf{IN}_{ss}^{v}\right)_{ijk} \tag{2.6}$$

Fixing the emission scenario (index j, 2nd dimension), this formulation aggregates population iFs over compartments (1st dimension) and/or exposure pathways (3rd dimension). *Pangea* considers only exposure through inhalation to be direct; all other exposure routes, including water consumption, are considered production based. We define the total cumulative *production based population intake and iF* as the sum of iFs inducted by production (of fresh water and food items). Direct and production-based iFs cannot be summed; the former happens locally, whereas the latter happens wherever production items are ultimately consumed. The missing link is a trade model such as the one implemented in IMPACT World (Shaked, 2011), which is not implemented in *Pangea* because of the difficulty to make it multi-scale.

2.8 Environmental Models

2.8.1 Introduction

Environmental models (EMs) are models that define the dynamics (or any other parameter) of media. The atmospheric EM, for example, defines air flows through walls of the cells of a 3D global multi-scale atmospheric grid (i.e. the dynamics of air as a transport medium, and not the dynamics of substances). The terrestrial model, on the other hand, describes the composition (land cover) of cells of the terrestrial grid. Four EMs are present in *Pangea*:

- Pangea Atmospheric Model (PAM): dynamics of air as a transport medium
- Pangea Hydrological Model (PHM): global hydrology of fresh water
- Pangea Terrestrial Model (PTM): land cover (+aggregation and correction)
- Pangea Oceans Model (POM): seas and oceans circulation

2.8.2 Pangea Atmospheric Model

PAM is the EM responsible for the following:

- Processing an input 3D (multi-scale) grid, such as that described in Section 2.4.5, and checking for compatibility with atmospheric grids requirements (e.g. contiguity between layers)
- Computing internal connections geometry and topology, and especially the 3D spherical topology.
- Pre-processing time series of wind velocities and computing directional monthly/yearly averages[32]
- Managing nested meteorological fields from GEOS-Chem
- Pre-processing time series of rain events
- Processing u10 (wind speed at 10m) data set
- Computing 3D interpolation of wind velocities (generally planar, in u and v directions), oversampling on cells walls for further "integration"
- Computing horizontal flows through 3D grid cell walls, and vertical flows by continuity

The techniques involved are fairly standard, but are implemented to take advantage of the 3D rectangular geometry.

By default, PAM uses a 55-layers, 2°×2.5° gridded global wind data set built on GEOS-4 meteorological six-hourly data for the year 2005. This data set was used for parameterizing other models such as IMPACT World and USEtox (for regional transfers), because 2005 was a stable year from a meteorological perspective. However, PAM supports more recent meteorological data sets, including

[32] This is done keeping u^+, u^-, v^+, v^- separate in order to account for bidirectional transfers instead of cancelling them.

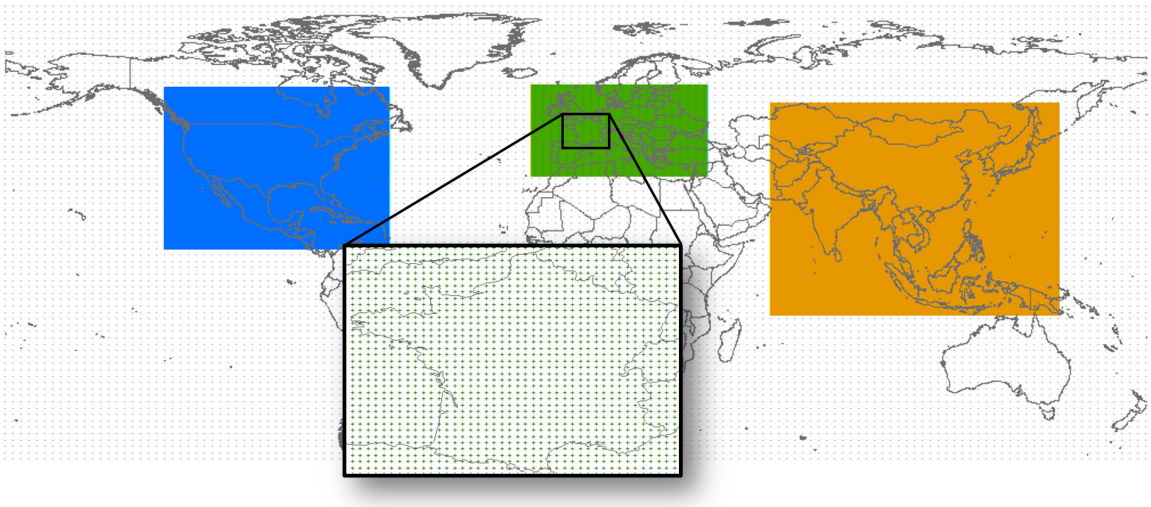


Figure 2.26 – GEOS-FP global 2° × 2.5° grid (light grey), and nested 0.25° × 0.3125° grids for North America (blue), Europe (green, with zoom on France), and China and Southeast Asia (orange).

higher resolution nested fields, principally based on GEOS-FP. GEOS-FP is the current operational meteorological data product from NASA/GMAO; it is a 72-layers, natively 0.25°×0.3125° gridded global wind data set built with three-hourly (hourly for some parameters) data. Due to the size of the data,[33] *Pangea* currently uses a nested approach based on a global 72-layers 2°×2.5° grid, in which it can nest 0.25°×0.3125° regions, as shown in Figure 2.26. PAM manages processing meteorological data at various scales (e.g. computing directional weighted averages over arbitrary periods of time), and nesting them.

2.8.3 Pangea Hydrological Model

PHM is the EM responsible for the following:

- Defining the terrestrial grid (because the terrestrial grid must be compatible with the hydrology defined by PHM)
- Defining main hydrological parameters: network topology, discharges, depth of freshwater bodies
- Defining other hydrological parameters: freshwater body surface area, lake parameters, etc
- Managing corrections based of the land cover data set.[34]

The current and default version of PHM is based on the global 0.5°×0.5° gridded hydrological model WWDRII (World Water Development Report). WWDRII was implemented in Pangea v1 because it is global and comes with a set of consistent gridded data sets that covers most of our needs. However, we

[33] 117GB for the 2°×2.5° grid per year, and 93GB per year for the 0.25°×0.3125° grid covering Europe for example.
[34] This is PHM type/version-specific; with the WWDRII-based PHM, it essentially means managing singular cases (where the hydrology is singular) and managing cells that have a seawater component.

encountered limitations inherent to its routing on a 0.5°×0.5° grid. This is discussed in Section 3.6.1, which provides an explicit example; in short, rivers may be routed farther than 0.5° (~70 km on the equator) away from their true location, which becomes a major issue when comparing modeled and measured concentrations at point sampling sites (leading to associations between sample sites and wrong rivers). For this reason, we have evaluated a HydroBASINS-based version of PHM, which will replace WWDRII in mid-2017.

WWDRII-based PHM – Typical grids associated with the WWDRII-based model are shown in Figure 2.11 through Figure 2.13. The current implementation of this model is based on the iMod River Model (iRM), a MATLAB class developed for *Pangea* that encapsulates WWDRII and extends it with small and large river lengths estimates (Helmes et al., 2012), and a clustering algorithm described in Section 2.5.5.

Based on HydroBASINS – Typical grids associated with the HydroBASINS-based model are presented in Figure 2.14, which displays the WWDRII grid for comparison. The current implementation, in test, is a MATLAB class that encapsulates HydroBASINS (shapes + basic properties), HydroATLAS (data sets), and HydroLAKES, and provides tools specific to *Pangea* for building multi-scale grids (section 2.5.6), building a hydrological **K** matrix, and integrating lakes.

2.8.4 Pangea Terrestrial Model

PTM is the EM responsible for the following:

- Defining the proportions of land-cover categories in terrestrial cells
- Defining the proportions of land-cover categories in countries, which is necessary for spatializing national inventories of food production
- Computing distribution key for distributing data based on land cover, between countries and terrestrial cells

The current underlying land cover data set is GlobCover 2.3 (ESA GlobCover 2009 Project), distributed as a high-resolution raster (129,600 × 55,800) with pixels that code for 22 (23 with “no data”) categories of land cover, as shown in Figure 2.27.

While the tasks that PTM must perform are straightforward in theory, the large size of the raster (~6.7GB uncompressed) makes it technically challenging. For this reason, PTM employs a pre-aggregated GlobCover-based data set of proportions summarized (by zonal statistics) on polygons resulting from the intersection of countries and the WWDRII native grid, represented in Figure 2.27 by polygons with black edges and white edges, respectively. This lengthy one-shot operation (neither countries nor the

native grid are project-specific) provides proportions of land cover for polygons that can be quickly aggregated into the WWDRII native grid (and further into project-specific terrestrial clusters), or into countries.

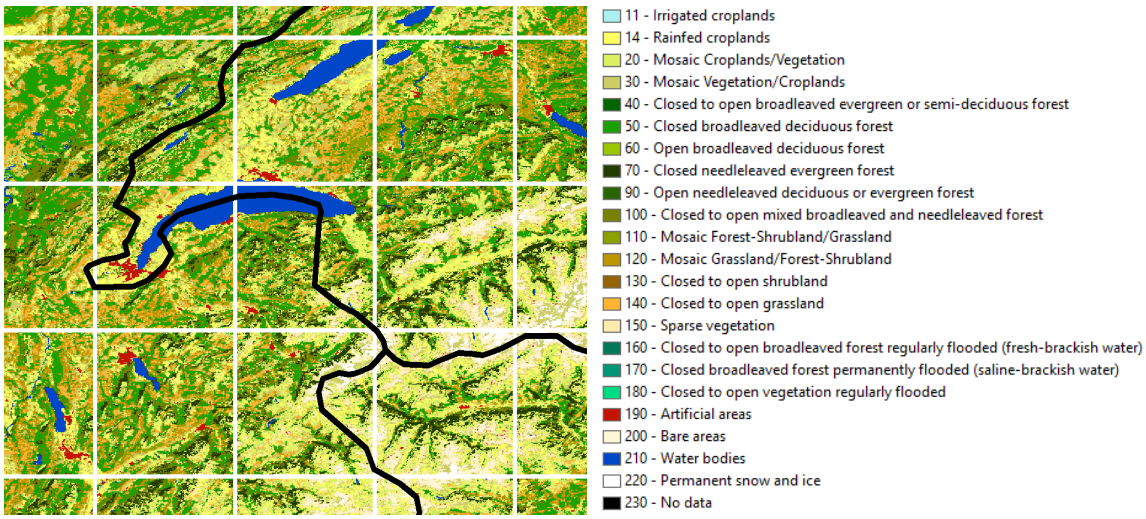


Figure 2.27 – Land cover (GlobCover 2.3), native WWDRII grid (white), and countries (black).

Finally, PTM provides mechanisms for spatially distributing data based on proportions of land cover. This is used in particular for spatializing food production data per country, using land cover as a proxy. This is not a trivial operation, and an easy way to represent what it entails is to understand that *Pangea* distributes food items across countries according to proportions of land cover within each country, and then gathers these food items on the relevant grid (typically the grid of terrestrial clusters).

2.8.5 Pangea Oceans Model

Pangea ocean model can grid oceans using an approach comparable to the atmospheric grid, but limited in depth by the bathymetry. This was not used in this thesis, because none of the studies was focusing on oceans, and because the main modeling difficulty with oceans is the connection with the freshwater hydrological network through a coastal zone for which we have no global data set of marine currents.

2.9 Environmental Process Models

Definition – Environmental process models (EPMs) define environmental transfer and elimination rate constant for substances, between and within media respectively. They describe first order processes (in *Pangea*) that define the evolution of environmental masses of substances. We add population exposure to these processes (because it is an elimination process in the sense that it is an output from environmental compartments), as well as transformation processes (substance to substance). While the computation engine supports this latter category of EPMs (Section 2.10.2), they are out of the scope of this thesis and were not implemented.

Main processes involved in EPMs – The *Pangea* EPM's handbook describes the main processes and parameters that are involved.[35] Parameters are split into substance-specific, spatial, and generic (global, substance-independent values) parameters. The GIS engine technically allows for spatializing any parameter by using vector or raster definitions (in a fashion comparable to the definition of the RP), but the limitations generally derive from a lack of data availability (especially at the global scale).

Implementation – Environmental process models work with abstract scalar/vector/matrix parameters, which is the beauty of this system. The computation engine (Section 2.10) "feeds" them with a block of data/parameters relevant to their functioning, which enables the implementation of computations in a direct, simple fashion. To illustrate, the air → agricultural soil transfer/deposition EPM "does not care" about the complex geometry of the contact surface areas between air cells and the distribution of agricultural soil within non-matching terrestrial cells; it receives a matrix of corresponding abstract contact surfaces areas (that account already for all proportions and mismatches), and it can implement direct, simple computations with them.

EPMs sets – We call sets of EPMs that cover all the needs of a simulation, *EPMs sets*. *Pangea* provides a default EPMs set (soon to be USEtox EPMs), but advanced users can add/remove/update EPMs from sets, or even create full sets. Having EPMs separate from the computation engines brings modularity and allows it to run and compare simulations based on various sets. This in turn enables, for example, the re-analysis of former projects using the USEtox EPMs set, and a comparison with former results.

Pangea currently implements EPMs based on IMPACT2002, OMNIITOX, and USEtox, which are documented in the *Pangea* EPM's handbook. We are, however, currently transitioning towards a set of EPMs that are strictly based on USEtox ("*scientific consensus model endorsed by the UNEP/SETAC Life Cycle Initiative for characterizing human and ecotoxicological impacts of chemicals*".

[35] The EPM's handbook is available here: http://pangea-model.org/documents/Pangea EPMs v10.pdf

2.10 Computation Engine

2.10.1 Introduction

The computation engine is the set of resources of the *Pangea* core that manages the creation of the mathematical system (parameters of the matrix differential equation), the computation of the steady-state and dynamical solutions of the fate and transport, and subsequent computations (population exposure, impacts on human health, etc.). The main operation for creating the mathematical system is the definition of the **K** matrix. This is done using (and specific to) an EPM set as defined in the previous section.

2.10.2 Matrix differential equation and virtual system structure

The matrix ordinary differential equation that *Pangea* must build and solve is Equation 2.4:

$$\frac{\mathrm{d}\mathbf{M}^{v}(t)}{\mathrm{d}t} = \mathbf{K}\,\mathbf{M}^{v}(t) + \mathbf{S}^{v}$$

where **K**, the matrix of transfer and elimination rate coefficients, is the sum of a matrix of transfer coefficients, a matrix of elimination coefficients, and a matrix of exposure coefficients, the last of which is an elimination as well:

$$\begin{aligned}\mathbf{K} &= \mathbf{K}^{\text{transfer}} + \mathbf{K}^{\text{elimnation}} + \mathbf{K}^{\text{exposure}} \\ &= \begin{bmatrix} \mathbf{K}^{\text{transfer}}_{\text{air}\to\text{air}} & \mathbf{K}^{\text{transfer}}_{\text{f.water}\to\text{air}} & \cdots & \mathbf{K}^{\text{transfer}}_{\text{agr.soils}\to\text{air}} \\ \mathbf{K}^{\text{transfer}}_{\text{air}\to\text{f.water}} & \mathbf{K}^{\text{transfer}}_{\text{f.water}\to\text{f.water}} & \cdots & \mathbf{K}^{\text{transfer}}_{\text{agr.soils}\to\text{f.water}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}^{\text{transfer}}_{\text{air}\to\text{agr.soils}} & \mathbf{K}^{\text{transfer}}_{\text{f.water}\to\text{agr.soils}} & \cdots & \mathbf{K}^{\text{transfer}}_{\text{agr.soils}\to\text{agr.soils}} \end{bmatrix} + \\ &\quad \begin{bmatrix} \mathbf{K}^{\text{elimination}}_{\text{air}} & 0 & \cdots & 0 \\ 0 & \mathbf{K}^{\text{elimination}}_{\text{f.water}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{K}^{\text{elimination}}_{\text{agr.soils}} \end{bmatrix} + \\ &\quad \begin{bmatrix} \mathbf{K}^{\text{exposure}}_{\text{air}} & 0 & \cdots & 0 \\ 0 & \mathbf{K}^{\text{exposure}}_{\text{f.water}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{K}^{\text{exposure}}_{\text{agr.soils}} \end{bmatrix}\end{aligned} \tag{2.7}$$

where each block is associated with a single medium (diagonal) or two media (off-diagonal). Figure 2.28 demonstrates the filling and the block structure of a real **K** matrix. Each colored point in the matrix represents a single, scalar rate coefficient. For off-diagonal (transfer) elements, given the structure of Equation 2.4, columns are associated with source compartments, and rows with destination compartments. The circled block is hence the block of all transfer coefficients from natural land

compartments to fresh water compartments. It represents the freshwater runoff. It is not a straight diagonal, because not all terrestrial cells have both natural land and freshwater components.

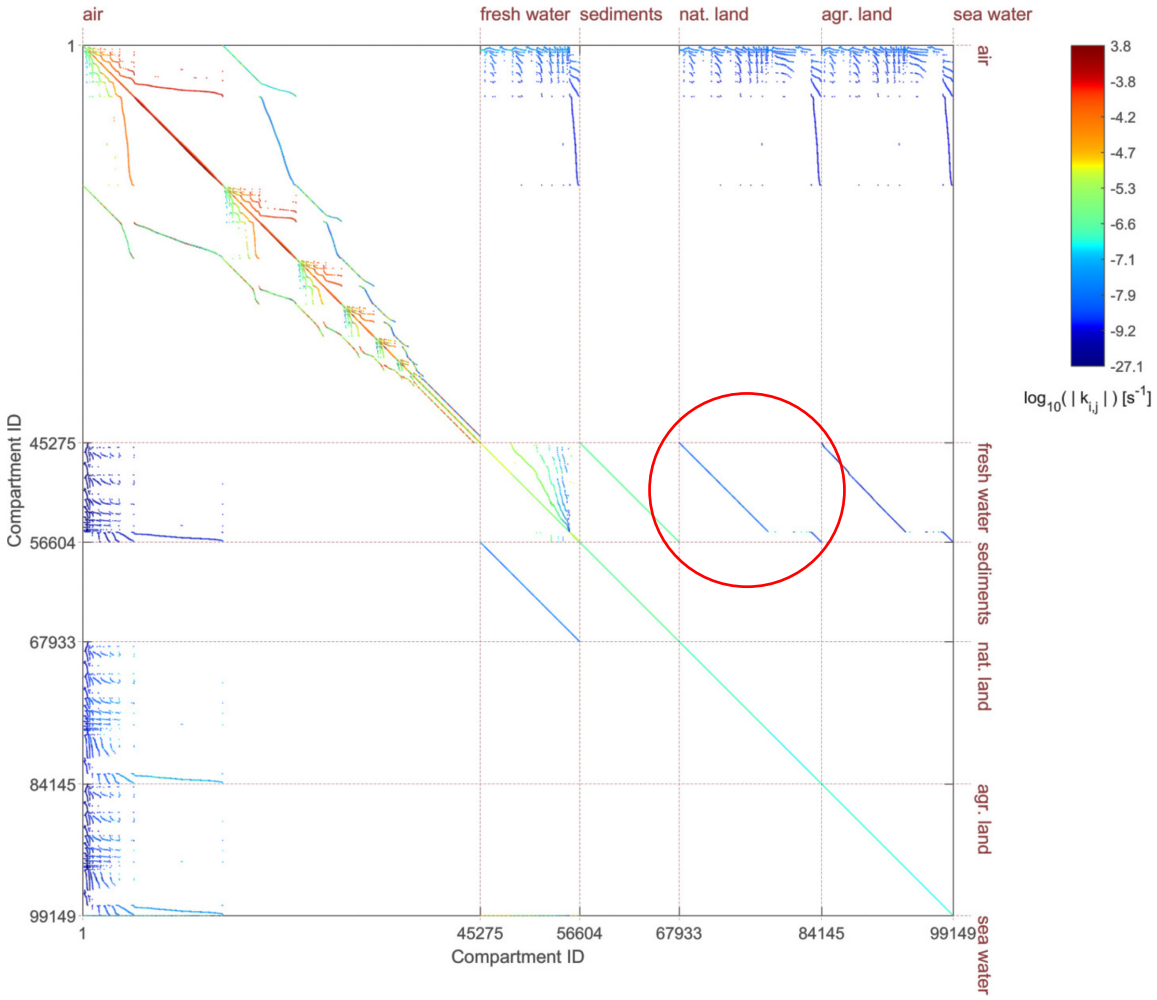


Figure 2.28 – Filling of a **K** matrix. The circled block (row #2: fresh water, column #4: natural land) is the block of transfer rate coefficients from natural land compartments to fresh water compartments.

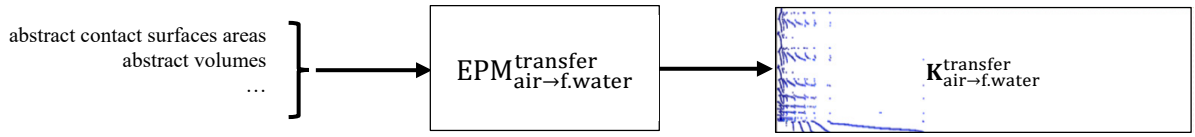
Blocks of the **K** matrix are computed by and for a given EPMs set, which can be represented as set of matrices of EPMs, that match the structure of the **K** matrix:

$$\mathrm{EPMs}^{\mathrm{transfer}} = \begin{bmatrix} \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{air}\rightarrow\mathrm{air}} & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{f.water}\rightarrow\mathrm{air}} & \cdots & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{s.water}\rightarrow\mathrm{air}} \\ \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{air}\rightarrow\mathrm{f.water}} & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{f.water}\rightarrow\mathrm{f.water}} & \cdots & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{s.water}\rightarrow\mathrm{f.water}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{air}\rightarrow\mathrm{s.water}} & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{f.water}\rightarrow\mathrm{s.water}} & \cdots & \mathrm{EPM}^{\mathrm{transfer}}_{\mathrm{s.water}\rightarrow\mathrm{s.water}} \end{bmatrix}$$

$$\mathrm{EPMs}^{\mathrm{elimination}} = \begin{bmatrix} \mathrm{EPM}_{\mathrm{air}}^{\mathrm{elimination}} & - & \cdots & - \\ - & \mathrm{EPM}_{\mathrm{f.water}}^{\mathrm{elimination}} & \cdots & - \\ \vdots & \vdots & \ddots & \vdots \\ - & - & \cdots & \mathrm{EPM}_{\mathrm{s.water}}^{\mathrm{elimination}} \end{bmatrix}$$

$$\mathrm{EPMs}^{\mathrm{exposure}} = \begin{bmatrix} \mathrm{EPM}_{\mathrm{air}}^{\mathrm{exposure}} & - & \cdots & - \\ - & \mathrm{EPM}_{\mathrm{f.water}}^{\mathrm{exposure}} & \cdots & - \\ \vdots & \vdots & \ddots & \vdots \\ - & - & \cdots & \mathrm{EPM}_{\mathrm{s.water}}^{\mathrm{exposure}} \end{bmatrix}$$

The circled block in Figure 2.28 is computed by the corresponding (same row/column) EPM in the matrix of transfer EPMs, which is $\mathrm{EPM}_{\mathrm{nat.land}\rightarrow\mathrm{f.water}}^{\mathrm{transfer}}$. The interest of this system is that EPMs are generic/abstract: $\mathrm{EPM}_{\mathrm{air}\rightarrow\mathrm{f.water}}^{\mathrm{transfer}}$, for example, can be a "simple" deposition model (e.g. from USEtox or SimpleBox), slightly extended to accept vector/matrix parameters instead of scalars (i.e. matrix of abstract contact surfaces areas, vector of volumes) and usual parameters for deposition models, and it outputs the corresponding matrix of transfer coefficients.



Some of these models require the output of more complex models. For example, atmospheric flows are parameterized by a standard run of GEOS-Chem interpolated in 3D over *Pangea* project-specific geometries for the atmospheric grid. The hydrological EPM is based on the full WWDRII hydrological model, and soon on HydroBASINS.

Finally, we do not use currently use transformation EPMs, but *Pangea* can build a composite system parameterized by a composite **K** matrix, as indicated in Figure 2.29, with substance-specific **K** matrices as diagonal block elements and transformation off-diagonal blocks. Part of *Pangea* Core supports this, but implementing transformation EPMs was out of the scope of this thesis. The size of a composite system is the size of the virtual system for a single substance, multiplied by the number of simultaneous substances. The very large size of such systems limits the maximum number of emission scenarios that can be treated in parallel.

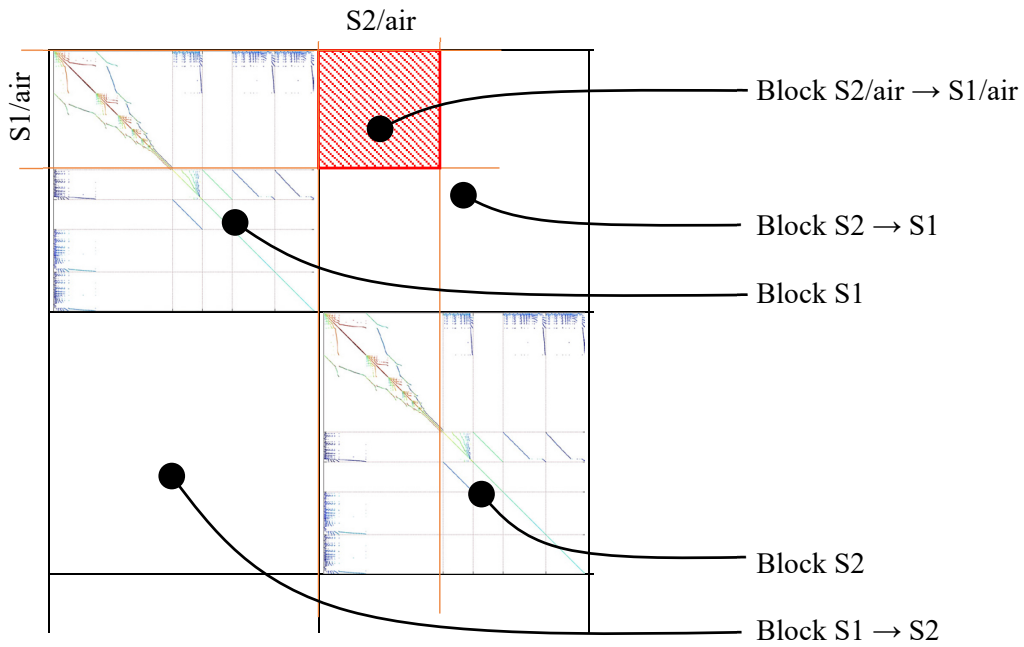


Figure 2.29 – Filling of a composite **K** matrix for multiple substances that can transform into each other.

2.10.3 Steady-state solution of fate and transport

Steady-state – *Pangea* computes the steady-state of systems with constant coefficients **K** and $\mathbf{S}^v$ by imposing a null derivative in Equation 2.4:

$$\frac{d\mathbf{M}_{ss}^v}{dt} = \mathbf{0}^v = \mathbf{K}\,\mathbf{M}_{ss}^v + \mathbf{S}^v \tag{2.8}$$

where $\mathbf{M}_{ss}^v$ is the matrix of vectors of masses at steady-state for all emission scenarios, written in column. Solving for $\mathbf{M}_{ss}^v$, we get:

$$\mathbf{K}\,\mathbf{M}_{ss}^v + \mathbf{S}^v = \mathbf{0}^v \quad \Rightarrow \quad \mathbf{K}\,\mathbf{M}_{ss}^v = -\mathbf{S}^v \quad \Rightarrow \quad \mathbf{K}^{-1}\mathbf{K}\,\mathbf{M}_{ss}^v = -\mathbf{K}^{-1}\,\mathbf{S}^v$$

and finally:

$$\mathbf{M}_{ss}^v = -\mathbf{K}^{-1}\,\mathbf{S}^v \quad \text{sometimes noted} \quad \mathbf{M}_{ss}^v = \mathbf{FF}\,\mathbf{S}^v \quad \text{with} \quad \mathbf{FF} = -\mathbf{K}^{-1}$$

As explained in 0, for larger systems like in *Pangea*, it is not possible to invert **K**, and therefore not possible to obtain **FF**. Yet, the linear system:

$$\mathbf{K}\,\mathbf{M}_{ss}^v = -\mathbf{S}^v \tag{2.9}$$

can be solved numerically for $\mathbf{M}_{ss}^v$.

Numerical approach – This operation can be time and memory consuming ($\mathbf{M}_{ss}^{v}$ is dense even when $\mathbf{S}^{v}$ is sparse) when the number of emission scenarios (= number of column of $\mathbf{M}_{ss}^{v}$ and $\mathbf{S}^{v}$) is large. For this reason, *Pangea* solves Equation 2.9 by block using a parallel approach (if enabled). Two options were available a priori: CPU and GPU. Given the size in memory of typical sparse **K** matrices (<20MB for typical system sizes), we can assume that transferring **K** to GPU arrays would be efficient. However, MATLAB "backslash operator" on GPU does not support sparse matrices (as of 03/2017). The approach currently implemented uses MATLAB backslash operator (direct, multi-frontal method, Davis et al., 2016) and an LU factorization described and implemented by Tim Davis in his "Factorize" MATLAB package. Numerically, we solve in parallel (PARFOR loop):

$$\mathbf{M}_{ss}^{v}|_{\text{block } i} = -\mathbf{Q} * (\mathbf{U} \setminus (\mathbf{L} \setminus (\mathbf{P} * (\mathbf{R} \setminus \mathbf{S}^{v}|_{\text{block } i})))) \tag{2.10}$$

where **L** and **U** are respectively unit lower triangular and upper triangular matrices, **P** and **Q** are permutation matrices, and **R** is a diagonal scaling matrix, obtained through LU factorization of the **K** matrix (optimized for sparse matrices). This approach is very advantageous because when solving:

$$\mathbf{M}_{ss}^{v}|_{\text{block } i} = -\mathbf{K} \setminus \mathbf{S}^{v}|_{\text{block } i} \tag{2.11}$$

the most time-consuming internal operation is the factorization of **K**. If we were performing this computation within the PARFOR loop, the factorization of **K** would be performed at each iteration (for each block). Yet, the **K** matrix is not block-specific, so factorizing it can and should be done only once, before the PARFOR loop. Given the particular filling of the matrices resulting from the LU factorization, solving Equation 2.10 in the loop is much faster than solving Equation 2.11.

Fate-factors – Finally, if inverting **K** was possible, storing and manipulating **FF** would not be possible because it is dense (80GB for storing a single **FF** when $n^{v} = 100{,}000$). Yet *Pangea* can compute rows, columns, or blocks of **FF** for e.g. computing a source apportionment.

2.10.4 Dynamic solution of fate and transport

In addition to the steady-state solution on which this thesis focuses, the dynamic solution can also be computed by part over a partition of the overall time span. The size of blocks is a function of the memory available on the system and the level of output. The approach is standard, but is parallelized over emission scenarios.

2.10.5 Population exposure

After solving substances fate and transport ($\rightarrow\mathbf{M}_{ss}^{v}$, $\mathbf{C}_{ss}^{v}$) based on the $\mathbf{K}$ matrix, *Pangea* must compute population exposure. This operation is often performed by matrix multiplication:

$$\mathbf{IN}_{ss}^{v} = \mathbf{XP}\ \mathbf{M}_{ss}^{v} = \mathbf{XP}\ \mathbf{FF}\ \mathbf{S}^{v} \tag{2.12}$$

where $\mathbf{XP}$ is and exposure matrix that account for bioaccumulation in substrates (air, water, food items) and population intake rates of substrates (inhalation, ingestion). Equation 2.12 occasionally appears with matrices of population and inverse volumes factored out of $\mathbf{XP}$, but the approach remains elegantly simple. The computation of exposure is more complex than basic matrix/vector multiplication in *Pangea*, due to the structure of numerical objects (defined for facilitating computations), but the approach is comparable. *Pangea* defines $\mathbf{XP} = \mathbf{XP}^{v} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ as:

$$(\mathbf{XP})_{ijk} = (\vec{\mathrm{p}}^{v})_{i} \times (\mathbf{IR}^{v})_{ijk} \times (\mathbf{BAF}^{v})_{ijk} \times (\vec{\mathrm{v}}^{v})_{i}^{-1} \tag{2.13}$$

where $\mathbf{XP} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ with n^{ep} as the number of exposure pathways (inhalation, drinking, ingestion per food item category), $\mathbf{BAF}^{v} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ is an array of generalized bioaccumulation factors (that can be interpreted as an application from environmental concentrations to concentrations in air, water, and food items), $\mathbf{IR}^{v} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ is an array of generalized individual intake rates (that can be interpreted as an application from concentrations in air, water, and food items, to masses taken in), $\vec{\mathrm{p}}^{v}$ is a vector of population counts, and $\vec{\mathrm{v}}^{v}$ is a vector of compartment volumes.

While $\mathbf{BAF}^{v}$ is substance dependent, non-spatial, and relatively easy to build, $\mathbf{IR}^{v}$ is substance independent, but is spatial and more complex to build. The reason is that ingestion components of $\mathbf{IR}^{v}$ are based on local food production, the computation/evaluation of which is not direct in a multi-scale setup. To define it, *Pangea* uses an FAO[36] statistic of food production per country and distributes it over the surface area of each country according to the land-cover data set (i.e. national fish production is distributed according to the proportion of fresh water in each cell, above-ground national production is distributed according to the proportion of agricultural land, etc.).

2.10.6 Computation of blocks of the matrix of fate factors

The structure of typical $\mathbf{K}$ matrices, as shown in Figure 2.28, questions the existence of some sparsity in the matrix of fate factors $\mathbf{FF} = -\mathbf{K}^{-1}$ that could be exploited. However, in practice, a numerically quick test based on the *Dulmage-Mendelsohn* decomposition indicates otherwise. Elements of $\mathbf{FF}$ can nonetheless be extracted by iterating through computations of the steady-state corresponding to specific

[36] Food and Agriculture Organization of the United Nations.

sources. Looking at equation $\mathbf{M}_{ss}^{v} = \mathbf{FF}\,\mathbf{S}^{v}$ we see that setting a single non-0 component of $\mathbf{S}^{v}$ per column to 1 will define $\mathbf{M}_{ss}^{v}$ as a selection of columns of **FF**. This approach can be parallelized using Equation 2.10, which makes the computation of large blocks of **FF** technically feasible. The limitation arises from the amount of RAM that is need for storing the dense blocks.

2.11 Wrappers and Parsers

Wrappers are tools that "wrap" external resources and provide an interface layer that facilitates communications with *Pangea*. External resources can be data sets (e.g. databases) or models (e.g. USEtox). Parsers are processing tools that parse files' content, extract and filter data, and store them in a *Pangea*-friendly format. Files can be text data files or even web pages/sites.[37] This section succinctly lists a few of the main wrappers and parsers in order to provide a general understanding of the tools available in *Pangea*, in libraries external to the core.

2.11.1 USEtox wrapper

Per the official website,[38] "**USEtox** is a **scientific consensus** model endorsed by the UNEP/SETAC Life Cycle Initiative for characterizing human and ecotoxicological **impacts of chemicals**". Main output is a database of recommended and interim **characterization factors** including fate, exposure, and effect parameters. It is distributed as an Excel macro-enabled file. The *Pangea* USEtox wrapper uses a COM server to open and communicate with USEtox. This provides access to the most relevant features and parameters of USEtox, and to run it from *Pangea*. It also allows for the addition of substances to USEtox. *Pangea* uses this wrapper to automatize the run and comparison with USEtox (model/model).

2.11.2 Main parsers

Pangea provides parsers for processing ChemIDplus, ChemSpider, EPIsuite, and many other physical, chemical, and toxicological online databases, such as ECHA. This enables, for example, the parsing of EPIsuite batch output within seconds. We do not extend further on parsers, as they are usually operational for a limited amount of time. This is especially true for parsers for web pages, which need to be updated each time web sites are refactored.

[37] The boundary between wrapper and parser is not absolute. The tool for accessing TOXNET ChemIDplus (https://chem.nlm.nih.gov/chemidplus/) is both: it wraps ChemIDplus into an interface layer that allows *Pangea* to send requests to ChemIDplus, but the main part remains parsing web pages.
[38] http://www.usetox.org/

2.12 Outputs: files, maps, reports

The user defines the level of output of simulations, because outputting large files or maps can be extremely time consuming. Standard output file types are MAT-Files, NetCDF files, Excel files, MATLAB figures, images (for maps), shapefiles (GIS, vector), and rasters (GIS, ~geo-referenced grid of pixels).

The full output, including figures analyzing model internals, can involve thousands of files. To structure this output in a user-friendly fashion, *Pangea* can output simulation reports as fully functional web sites, exemplified in Figure 2.30, that can be opened locally (no server-side technology, only HTML, CSS, JavaScript) or published online[39].

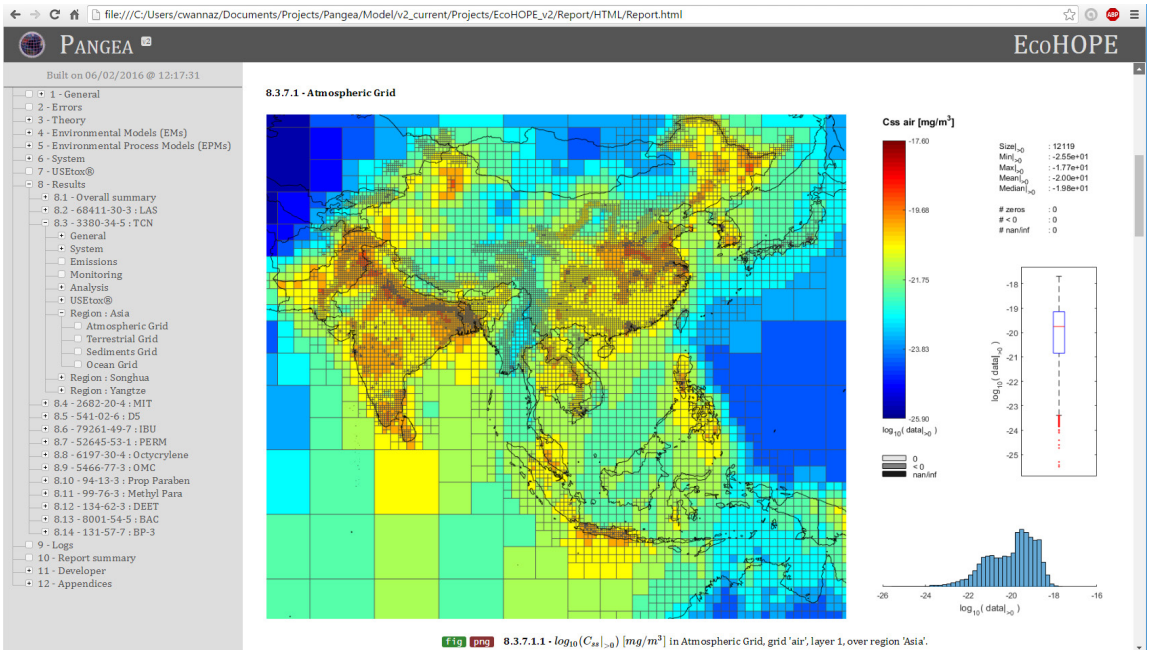


Figure 2.30 – Example of output report as website.

[39] This is possible provided the user has access to a server and enough free space (200Mb to 1GB per project full output).

2.13 Discussion of key, challenging concepts

2.13.1 Instantaneous full mixing

The assumption of instantaneous full mixing is mandatory as compartmental models characterize systems with homogeneous compartment contents. Warren et al. (2009) have discussed the minimal resolution necessary for this assumption to be valid, and it can be addressed through the RP mechanism (Section 2.5.3), i.e. *Pangea* can build atmospheric grids that respect Warren's 25% removal criterion. It is more complicated for other grids and media, typically for the hydrological network, as terrestrial grid refinement is only weakly constrained by the RP and is limited by the native maximum resolution of the hydrological network.

2.13.2 Decoupled EMs and EPMs

The structure of *Pangea* and the structure of the differential equation that describes the evolution of the system are based on decoupled EMs and EPMs. The underlying assumption is that the presence of substances in media involved in transport does not change the dynamics of these media, i.e. the presence of substances in air has no impact on the winds. This allows for using single runs of complex models that describe the dynamics of media (with all their internal complexity), e.g. GEOS-Chem for winds, and to overlay a simpler, first-order dynamics of substances at run time.

2.13.3 Direct versus production-based iFs

Two categories of iFs are present in *Pangea*: the direct iF, which is the inhalation iF, and the production-based iF, which requires more explanations. The output of the fate and transport computation is the distribution of concentrations of substance in the environment (in air, fresh water, sediments, natural land, agricultural land, and sea water). Using BAFs from USEtox, and food production based on national statistics of food production (FAO stat) spatialized using the land cover as a proxy, *Pangea* computes a spatial distribution of the concentration of substance in food items (meat, fish, milk, eggs, above-ground produces, and below-ground produces, and we also include fresh water for consumption as a production good). The concentration in food items of a given cell is therefore the concentration in the food that is produced locally, and the concentration in food that is consumed locally.

Pangea computes and retains direct and production-based iFs. These cannot be mapped and compared, as the direct iF in a cell is the iF through inhalation of the local population, but the production-based iF is the iF due to local production, wherever it is consumed.

The link between production and consumption is *trade*. While a trade model was implemented in the fixed grid/resolution IMPACT World by Shaked et al. (2011), it is difficult to make trade models multi-scale, and there is therefore no trade model in *Pangea*.

2.13.4 Singular regions defined by the hydrological model

Large regions of the world are characterized by a singular hydrology (low precipitations, aridity, no discharge), which translates into singular hydrological parameters in the hydrological model. This leads to situations such as that presented in Figure 2.31, in which the volume of water associated with a large terrestrial grid cell could be held in "a glass". Moreover, as the surface area covered by the cell is large, the deposition may be large as well, which leads to a massive concentration of the substance in fresh water. Finally, the population over a large cell may also be large, which may lead to a situation in which the model computes an exposure based on the assumption that a large city pumps water for drinking in a glass of water that captures all the deposition over a surface area of the size of a country.

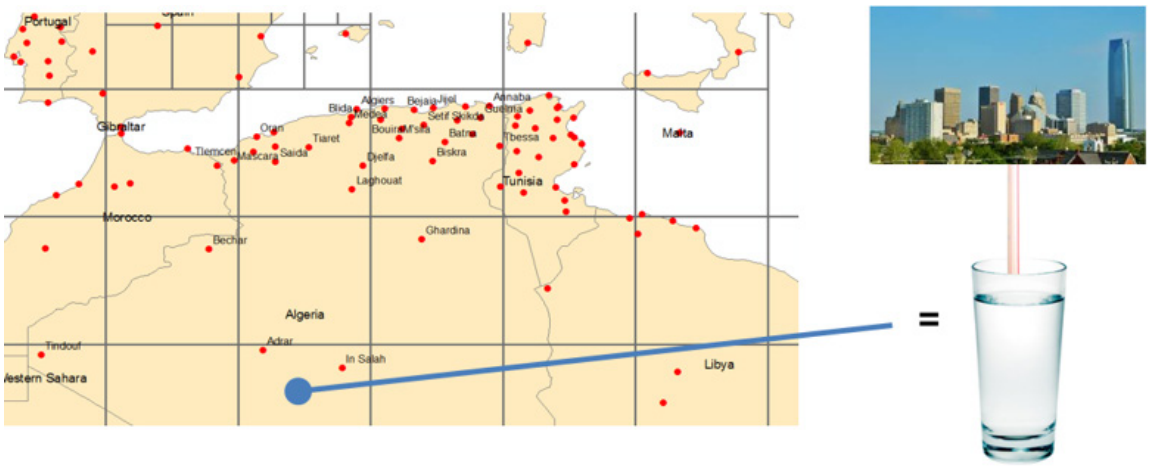


Figure 2.31 – Large arid regions with the equivalent volume of "a glass" of fresh water, concentrate potentially large depositions (over their large area). Large populations may pump and drink from this glass, if not modeled properly.

2.13.5 Exposure as an elimination process in the fate and transport computations

While population exposure is rarely included as an elimination process from the environment, it had to be included in *Pangea*. The reason for this is illustrated by the next section and relates to the use of BAFs. When a substance has a substantial BAF, for example *fresh water to fish*, the extraction of fish must be accounted for as an elimination process, otherwise the concentration of substance in fresh water remains constant while fish production extracts significant amounts of substance (which creates substance that does not exist). In extreme cases, for example with PCB-118 (whose BAF *fresh water to fish* is defined by USEtox as 4,762,000), in regions where the concentration of PCB-118 in fresh water

is extensive due a singular hydrology (previous section), this can literally lead to having 1 kg fish that "carry" 10 kg of PCB-118.

2.13.6 Discrepancy on coasts due to coarse terrestrial grid geometry

A series of three artifacts associated with the geometry of grids has already been discussed in the introduction, and illustrated in Figure 1.2 through Figure 1.4. This section presents a discrepancy due to the coarse geometry of the WWDRII-based terrestrial grid, and evidences a consequence in coastal regions. Figure 2.32 depicts a raster of populations counts (the darker, the more populated) that follows a coastal delineation. A terrestrial grid based on the WWDRII native grid is overlaid.

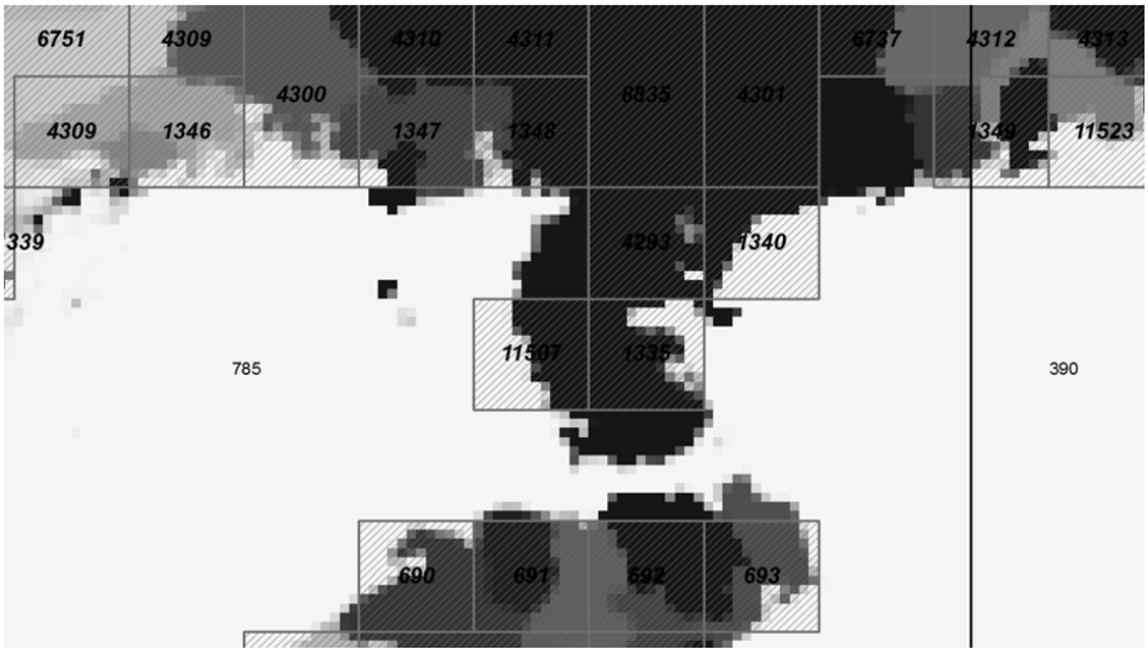


Figure 2.32 – Discrepancy in coastal regions. The background raster in shades of grey represents population counts, and the grid is the terrestrial based on WWDRII.

For computing the population exposure in a terrestrial grid, *Pangea* integrates (zonal statistics) population counts on the grid. In doing this, all people/pixels that are not overlaid by the terrestrial grid are lost. This is a consequence of the mismatch between the geometry of the grid and the geometry of the coast, which loses 300 million people (~ the population of the USA) globally on coasts, which are regions with high activity.

Pangea currently addresses this problem by capturing the loss using the first layer of the atmospheric grid (a partition of the globe ⇒ no loss) and redistributing the people to neighboring terrestrial cells. Using the HydroBASINS geometry significantly reduces this issue, as it has an accurate coastal geometry.

2.13.7 The spatial resolution of *Pangea*

Pangea can technically create grids with arbitrary high resolution (sub-pixel). We do not limit the resolution because we want the model to exploit all available geo-referenced data. Yet, one must keep in mind that spatial data are summarized or interpolated as necessary, depending their nature[40]). Thus, *Pangea* does not have a resolution per se. The technical resolution is defined by refinement parameters (e.g. refinement maximum depth) and is project-specific.

A limitation arises from the linearity of EPMs, which restricts the domain of validity of *Pangea* to situations and scales for which non-linear and second order phenomena (saturation, turbulence) are negligible. The validity of the scales and resolutions must be discussed on a per-study basis.

Finally, in practice, we usually build grids with a higher resolution than we can justify (due to linearity), but we ultimately analyze and discuss an integrated value either spatially (over a zone of intermediate resolution), temporally (yearly average), or both.

2.13.8 Analysis of the dynamic solution

Current literature provides multiple discussions and warnings concerning the importance of a dynamic resolution. For example, McLachlan has written in his review of EEA (2007), "*Non-steady state models should have a key role (for persistent chemicals) in higher tier risk assessment, as monitoring will only give a snapshot. The future developments of the environmental concentrations can only be evaluated with a non-steady state model.*" Studies have reported on and discussed the importance of seasonal events: snowpack and snow melt (Gouin et al., 2002; Meyer et al., 2009a, 2009b), forest scavenging or forest filtering (Wania and McLachlan, 2001), and also changes in temperature that impact a chemical's Henry's law constant or diurnal and nocturnal variations that induce fluctuations in the OH radical concentration profile, and fluctuations of the atmospheric mixing height (Shaked, 2011). Less regular events, such as intermittent rain, have been noted to play a crucial role (Hertwich, 2001; Jolliet and Hauschild, 2005). More generally, the meteorology is constantly changing, and so is the hydrology.

When we are interested in medium- to long-term impacts, some of the aforementioned dynamical events can be reduced to their averages, or algebraic solutions can be used to account for the dynamics in a steady-state model. Jolliet and Hauschild (2005), for example, have reduced the intermittent rain dynamic complexity to first-order deposition coefficients that account for periodic rain events. However,

[40] Raster data can be summarized on grids by zonal statistics, or resampled, and vector data can be interpolated and e.g. integrated on grids cells "walls".

not all the dynamic events and parameters varying with time (a list is given by Hollander et al., 2007) can be simplified in this way.[41]

Pangea implements a numeric solver for the dynamics that allows for the computation of $\mathbf{M}_{ss}^{v}(t)$ and further population exposure and iFs. An analysis of the dynamics was out of the scope of this thesis, but we implemented solvers and various mechanisms (e.g. reduction of spatial extent, time-period archetypes for summarizing time series) to evaluate the feasibility and perform a first set of experiments. However, *Pangea* currently lacks the tools to analyze the dynamics properly and efficiently (e.g. model reduction).

The lack of global, spatial time series for parameterizing *Pangea* is one of the major reasons for omitting studying the dynamics from this thesis. Without global time series, we could only solve the dynamics of a system with constant coefficients, which limits the relevance of the results. The main interest of such an operation is the evaluation of the time necessary for the system to stabilize (close to steady-state), which enables a discussion of the relevance of the steady-state solution for supporting decision-making.

The recent availability (to *Pangea*) of spatial meteorological time series from GEOS-FP motivates the future implementation of more efficient dynamic solvers. The major limitation results from the size of the system and the computation time required for solving the dynamics over long time spans. A promising approach relies on time-period archetypes that reduce data-intensive time series to small sets of parameters that characterize specific time archetypes, e.g. 16 pre-computed sets of parameters that characterize two diurnal and two nocturnal points in time for four seasons.

2.14 Model limitations

The two most important limitations that are discussed first are first-order EPMs, which are fixed by the design of the framework, and the lack of data for parameterizing ocean coastal zones (marine currents), for which we do not foresee any satisfying solution emerging in a near future (either technical or based on other data sets as proxy). Other limitations are due to a current lack of global spatial data sets or to technical issues that can be addressed with a reasonable effort.

[41] We can mention two examples. First, the time duration between pesticide application to crops and their harvest is known to be a crucial parameter for population exposure (Charles, 2004), and harvests are seasonal or yearly events that can see dramatic changes between each pesticide application. Second, since pollutant concentrations do not depend linearly on wind speeds and mixing heights, these parameters are not easily reducible to a representative average value (Shaked, 2011).

2.14.1 First-order EPMs

Environmental process models are required to describe first-order processes. This strong constraint is a modeling choice that ensures structural and mathematical compatibility between EPMs. It allows *Pangea* to couple models that describe a variety of environmental processes in a simple, unified framework. This sets a strong boundary as to what *Pangea* can accomplish, as second-order reaction rate constants cannot be taken into account. *Pangea* was not designed to target problems for which second-order processes are important, even in the long term. Models suited for these problems belong to another, "multi-physics" category of models that do not allow for fast evaluation of emission scenarios. To our knowledge, there is currently no operational global model in this category.

2.14.2 Lack of data for parameterizing ocean coastal zones

The ocean coastal zone is a zone in which most sea-related human activity happens (e.g. fishing), as well as most of the mixing of fresh and sea water and the coupling of the sea and land. Figure 2.33 depicts a coastal zone delineated by brown lines, as well as a data set of ocean marine currents (Ocean Surface Current Analyses Real-time, OSCAR) at locations defined by dots. We observe that there are almost no current data in coastal zones. This makes *Pangea* unable to properly couple its ocean model with the freshwater hydrology or effectively describe what happens in the marine coastal area (concentration in sea water, in marine fauna and flora, and exchanges with air). While this limitation is due to spatial data availability, we present it as a major limitation because we do not foresee timely improvements in data availability, and because technical options for compensating are scarce, complex, and approximative.

Figure 2.33 – The oceanic coastal zone is delineated by the brown line. Ocean surface currents (from OSCAR) are available at dotted locations.

2.14.3 Linear differential equation

The linearity of the differential equation that describes the evolution environmental masses is project-specific. By default, *Pangea* uses the constant coefficients Equation 2.4 and linear EPMs for building the **K** matrix. This allows for computing the steady-state solution (Eq. 2.9 and Eq. 2.10), which is the most relevant solution for studies tackled in this thesis. In all other cases, the differential equation is solved numerically, which enables treatment of the more general case:

$$\frac{\mathrm{d}\mathbf{M}^{v}}{\mathrm{d}t} = f(\mathbf{M}^{v}, t) \tag{2.4}$$

where the function f can be non-linear. One of the simplest implementations of a non-linear behavior is through a differential equation with the same structure as Equation. 2.4, but with $\mathbf{K} = \mathbf{K}(\mathbf{M}^{v})$, where one or more EPMs implement saturation based on the environmental mass (→ concentration).

2.14.4 Framework and model evaluation

As a framework that can incorporate any type of first-order EPM, parameterized using a variety of EMs, and simulate a large number of substances with physical and chemical properties that can be defined based on most chemical databases, it is not possible to fully evaluate *Pangea,* and even less possible to validate it.

The semi-automatic evaluation of *Pangea* is limited to the technical verification of model components and the evaluation of trivial situations (e.g. no wind, homogeneous winds, no discharge of fresh water).

Users must implement the evaluation on a per-project basis, e.g. by comparing it with measurements. It is specific to the model that is ultimately built, i.e. to the choice of EMs, EPMs, and physical and chemical databases, to the choice of other flexible framework components, and to the project definition.

In addition, *Pangea* incorporates external model wrappers, and a USEtox wrapper in particular that allows for systematic model-model comparison with USEtox. A set of tools also permits the evaluation of EPMs by users, i.e. facilitates the analysis of EPMs outputs based on variations of their inputs, but leaves it to users to interpret results.

It is currently not possible to reduce *Pangea* to a simpler, e.g. “unit world,” or to have it fully emulate the 13 compartments of USEtox. While it seems that this feature should have been implemented first, the problem was that the reduction of *Pangea* to USEtox’s 13 compartments for comparison would technically require the elimination (or bypass) of many components and mechanisms of *Pangea*. Doing this would eliminate a large part of *Pangea* from the test. Moreover, the current set of EPMs is transitional (based on IMPACT 2002 and USEtox v1) and does not fully match the current distribution of USEtox (v2), the documentation for which was just released in 2017. This would prevent strict emulation and direct comparison.

The first evaluation that could be achieved was therefore the systematic comparison of the output of *Pangea,* summarized on the geometry underlying USEtox with the output of USEtox run by *Pangea* (automatically, through the wrapper – see Section 3.5.2). The next refactoring of *Pangea* includes an

update to USEtox v2 EPMs. This operation will advance toward a more direct compatibility between the models and a future reduction of Pangea to 13 compartments matching USEtox.

2.14.5 Input data and parameters

Physical and chemical properties – *Pangea* implements a cascaded database of chemical properties, which selects values based on a list of priorities defined per parameter (including, but not limited to selection from literature, ECHA, USEtox, EPIsuite experimental, and EPIsuite modeled). While this is an advantage in that it allows us to select data sources that we think are the most reliable, it creates situations in which chemical properties for a given substance originate from multiple databases (e.g. K_{ow} from ECHA and $DT50_{water}$ from USEtox). This also enables us to observe a large variability (sometimes three orders of magnitude) in published degradation half-lives, which play a crucial role in the parameterization of the fate.

Non-spatial parameters – The use of generic values is also a limitation. Total suspended solids (TSS) are set at 15mg/L in USEtox, as a global average, and this is the default value in *Pangea* as well, as there is no global spatial data set of TSS. Yet, the literature has reported a range of 12 mg/L to 23,900 mg/L for Asia, with peaks of 48,000 mg/L in the Yellow River.[42] These parameters will be made spatial when global spatial data sets become available, but this is a slow process. It is therefore necessary to discuss non-spatial parameters on a per-project basis, and to refine them for specific regions or contexts when relevant.

Spatial parameters – Most spatial data sets (except for the land cover and population counts) are not available at high resolution. The highest resolution in GEOS-Chem meteorological fields is currently 0.25° ×0.3125°, for example. This limits the scope of the model to situations and scales for which the details of the terrain (or buildings) are not important. Some spatial data sets are not global; for example, the precipitation data set from TRMM is limited to a band of 60° N-S latitudes. In such cases, *Pangea* can use spatial data sets where available and e.g. USEtox global averages elsewhere, but the discussion of the compatibility is left to users. Finally, limited global spatial data availability often yields situations in which projects are evaluated based on combinations of data sets from different years, or combinations of spatial data and global/generic averages. This must be discussed on a per-project basis.

[42] Thus, a higher value of 100 mg/L was selected specifically for the EcoHOPE project (Chapter 3), which covers Asia.

2.15 Conclusion

Pangea is a unique functional spatial multi-scale framework for fast evaluation of multimedia fate and transport, and subsequent multi-pathway exposure of populations. It solves the *technical* problem of spatialization of fate and exposure models on a project-specific basis.

The main limitations are discussed in Section 2.14; they are primarily the restriction of fate and transport models to first-order processes, which is a modeling choice, and data quality, especially the limited availability of global spatial data sets.

The flexibility of the framework, its evolutivity (standard interfaces to EMs, EPMs, and data sets that can be replaced quickly), and its incorporation of tools for processing large geo-referenced data sets (which were once a time-consuming manual task for GIS specialists) have opened new perspectives for performing studies tailored to any region of the globe, for re-analyzing projects following the evolution of geo-referenced data sets, and for performing sensitivity studies against multiple data sources (e.g. comparing the influence of GEOS-Chem and ECMWF wind fields on the outcome of studies). This flexibility and implementation of external model wrappers are key strengths of the framework, as they allow for systematic re-analyses, as well as systematic model-model comparison.

Finally, the computational efficiency makes it possible to address large numbers of contexts and substances. *Pangea* can, for example, easily simulate fate, transport, and population exposure associated with emissions of the ~3,100 substances defined in the USEtox v2 substances database on a global geometry that is refined around 100,000 points of interest (e.g. location of waste treatment plants).

Chapter 3 Model Applicability and Evaluation (fate)

3.1 Foreword

This chapter will be submitted to the journal *Environment International*, under the title "A Global Framework to Model Spatial Ecosystems Exposure to Home and Personal Care Chemicals in Asia", by:

Cedric Wannaz[a,*], Antonio Franco[b], John Kilgallon[b], Juliet Hodges[b], Olivier Jolliet[a]

[a] Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109, United States

[b] Unilever, Safety & Environmental Assurance Centre, Colworth Science Park, Sharnbrook, MK441LQ, United Kingdom

3.2 Abstract

This paper analyzes ecosystems spatial exposure to home and personal care (HPC) chemicals, accounting for market data, environmental processes in hydrological water basins networks, including multi-media fate and transport. We present a global modeling framework built on ScenAT (spatial scenarios of emission), SimpleTreat (sludge treatment plants), and *Pangea* (spatial multimedia fate and transport of chemicals), that we apply over Asia to four chemicals selected to cover a variety of applications, volumes of production and emission, and physico-chemical and environmental fate properties: the anionic surfactant Linear alkylbenzene sulphonate (LAS), the antimicrobial triclosan (TCS), the personal care preservative methyl paraben, and the emollient decamethylcyclopentasiloxane (D5). We present maps of predicted environmental concentrations (PECs), and compare them with monitored values, over all of Asia and per river. LAS emission levels and PECs are two to three orders of magnitude greater than for other substances, yet the literature about monitored levels of LAS in Asia is very limited. We observe a good agreement for TCS in fresh water (Pearson r = 0.82, for 253 monitored values covering 12 streams), a moderate agreement in general, and a significant disagreement for methyl paraben in sediments. This points to the need for more specific/spatial information on both chemical/hydrological parameters (DT50water, DT50sediments, Koc, foc, TSS) and monitoring sites (e.g. spatial/temporal design, site medium parameters), especially for evaluating PECs in sediments in

Asian streams. We illustrate the relevance of local evaluations for short-lived substances in fresh water (LAS, methyl paraben), and their inadequacy for longer lived-substances (TCS, D5). This framework constitutes a milestone towards higher tier modeling approaches for identifying hotspots, and linking large-scale fate modeling with (sub) catchment-scale ecological scenarios; a major limitation encountered comes from the discrepancy between streams routed on a gridded, 0.5°×0.5° global hydrological network and actual locations of streams and monitoring sites.

Keywords: Ecosystem exposure, Multi-scale spatial models, Multimedia fate models, Home and Personal Care Chemicals, Global hydrological network, Asia

3.3 Introduction

3.3.1 Background

Exposure modeling plays an essential role in exposure assessment under regulatory frameworks for environmental risk assessment of chemicals. Non-spatial models, such as generic "box" multimedia fate models (SimpleBox, USEtox), are widely used for screening exposure and risk assessment. Versatility, easiness of use, minimal input data requirement and simplicity of outputs are key ingredients of their widespread use in both academic and regulatory applications. Such models, however, lack environmental realism and are unable to reflect large spatial variations in both chemical releases (Aronson et al., 2012) and chemical fate in the environment (Keller, 2006; Pennington et al., 2005, Figure 6), in particular due to spatial variations in population densities, connectivity to waste water treatment plants (WWTP, Hodges et al., 2012), and in local characteristics of the hydrological network. This is especially true for household-use chemicals such as ingredients of HPC products and pharmaceuticals that can be released into and transported via multimedia pathways including water, sediments, agricultural soil, and air, and can present highly variable spatial exposure profiles.

The recognized lack of environmental realism of non-spatial approaches has stimulated the development of several models based on or incorporating a Geographical Information System (GIS), which accounts for specific spatial characteristics (e.g. the river network and WWTP locations), for use in higher-tier realistic assessments (Franco et al., 2016). However, the use of these spatial models in regulatory contexts remains scarce for consumer use chemicals. Limited versatility and high data requirements are key limitations for widespread use and regulatory acceptance. Existing large-scale spatial exposure models are actually often limited to specific countries (Keller et al., 2014; Suzuki et al., 2004), subcontinental catchments (Kehrein et al., 2015; Lindim et al., 2016; Wang et al., 2005), and/or fail to

account for multimedia exposure pathways (Keller et al.; 2012, Aronson et al., 2012) of chemicals that are transported to media different from media of emission.

In consequence, there is a need for a versatile, medium- to high-resolution, global-scale, multimedia model platform adapted for chemicals with a multimedia behavior. This platform can take advantage of recent developments in spatial emission models of HPC ingredients that enable accurate estimates of ingredients use based on market data and environmental and socio-economic parameters, such as the ScenAT model (Hodges et al., 2012), which can be used as input for global environmental fate models. Targeting chemicals that are primarily emitted via domestic wastewater, this platform can also integrate global spatial hydrological models that couple both rivers and lakes, and define relevant hydrological parameters at a regional catchment scale. The World Water Development Report II (WWDRII, Vörösmarty, 2002) global gridded hydrological model, available at a typical resolution of 0.5°, updated by (Helmes et al., 2012), for example, provides long-term annual averages of hydrological parameters.

This paper aims therefore to develop and test, for all of Asia, a modeling platform combining spatialized emissions and a global, spatial multi-media fate and transport model. It specifically aims to: 1. Develop a spatial exposure modeling framework for HPC products, from market to environmental concentrations. 2. Apply the framework to estimate spatial multimedia environmental concentrations (PECs) of selected HPC chemicals across Asia. 3. Evaluate the framework by comparing PECs with USEtox results (model-model comparison with consensus model), and with monitored concentrations in fresh water and in sediments collected from literature. 4. Analyze spatial variations of PECs across Asia and identify key parameters/factors that are associated with high environmental concentrations.

This work presents a global, spatial model framework to estimate environmental exposure of HPC ingredients by coupling the ScenAT spatial emission model, the wastewater treatment plant model SimpleTreat (Struijs, 2014), and the multi-scale multimedia fate and transport model *Pangea*. We present maps of predicted environmental concentrations (PECs) for four selected HPC chemicals, a model-model comparison with USEtox, and a model-monitoring comparison. We discuss key parameters, discrepancies, limitations, and opportunities for future developments aiming at large-scale, high-resolution modeling of freshwater ecological scenarios.

3.4 Methods

3.4.1 Area and substances of interest

The study analyzes environmental concentrations of a selection of HPC chemicals over 14 countries across South and East Asia: Bangladesh, Cambodia, China, Vietnam, India, Indonesia, Japan, Laos, Malaysia, Pakistan, Philippines, South Korea, Sri Lanka, Thailand, and Vietnam. The following four

case study chemicals were selected to cover a variety of applications, volumes of production and emission, environmental persistence and affinities to water and air (Table S2.1, SI): the anionic surfactant linear alkylbenzene sulphonate (LAS), the antimicrobial triclosan (TCS), the personal care preservative methyl paraben, and the emollient decamethylcyclopentasiloxane (D5).

3.4.2 Model framework

3.4.2.1 General structure

The framework (Figure 3.1) is built by combining the ScenAT spatial emission model with the SimpleTreat wastewater treatment plant model (Struijs, 2014) to define spatial environmental emissions. These emissions are used as input to the *Pangea* global spatial multimedia fate and transport model, that computes and maps predicted environmental concentrations (PECs). Modeled PECs are then compared with the non-spatial output of the consensus model USEtox and with a data set of monitoring values. Components of the framework are described in detail in the following sections.

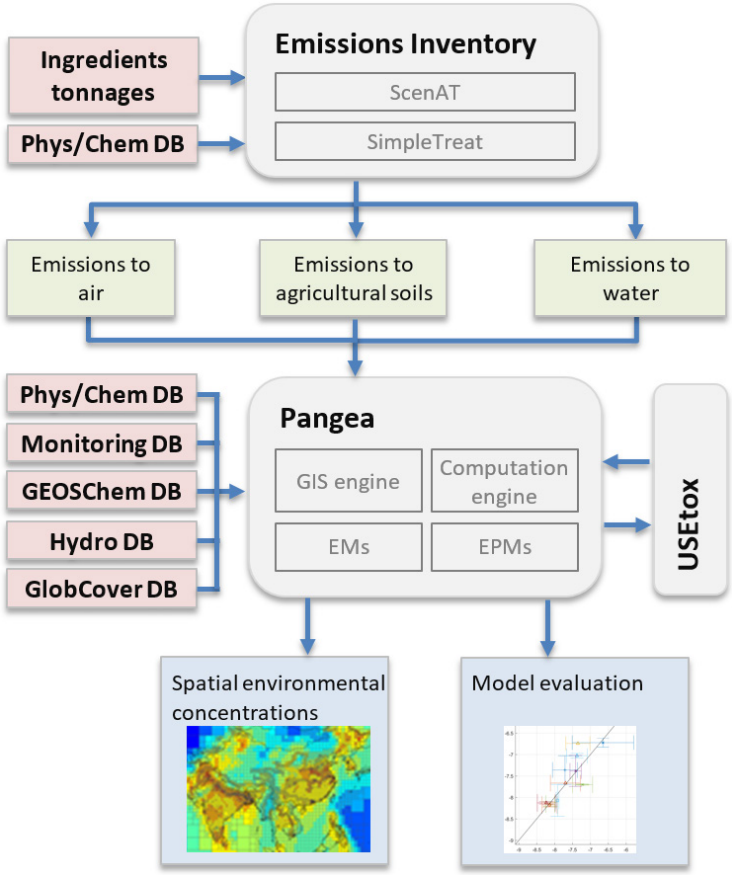


Figure 3.1 – Model framework coupling ScenAT (spatial emission inventories), SimpleTreat (sewage treatment plants), and Pangea (spatial, multimedia fate and transport). Input data in red, emission inventories in green, outputs in blue.

3.4.2.2 From market to spatial emissions

Market tonnages - To derive spatial emissions for the selected case study chemicals, national market sales data were used following the methods in (Zhu et al., 2016). In brief, products sale volumes of each product category per country were taken from Euromonitor International databases (Euromonitor, 2016). The market penetration of a particular ingredient in a given product category was derived using the Mintel Global New Products (GNDP; Mintel, 2016) information on market activity and product variant data. Finally, ingredients inclusion level in products were derived from regulatory dossiers (i.e. TCS – European Commission, LAS – HERA, Methyl Paraben – Cosmetics Ingredient Review, Personal Care Products Council – D5) (SI). Multiplying product sales data with ingredients market penetration and inclusion levels provides a total tonnage at the national scale for the select chemical. As LAS is not labeled on home care products the GNPD cannot be used to determine market penetration, therefore Euromonitor sales data and inclusion levels were applied to determine LAS market tonnage following (Whelan et al., 2012).

SimpleTreat - To account for fate and elimination in wastewater treatment plants (WWTPs) the SimpleTreat model was used (Struijs 2014). SimpleTreat models a conventional activated sludge treatment plant and calculates the fraction degraded and the fractions emitted to freshwater, to air, to sludge based on a substance physico-chemical and biodegradability properties. Results from previously published simulations were used for LAS, TCS and D5 (Franco et al. 2013); a similar approach was taken to obtain estimates for methyl paraben (readily biodegradable), using the same chemical properties as for the present study (Table S2.1).

ScenAT - ScenAT is a global environmental exposure model which can be used to derive spatially explicit emission inventories from country-level tonnages and spatial socioeconomic and environmental variables (Hodges et al., 2014, 2012). Discharge pathways for substances used in HPC products depend on a country's infrastructure (i.e. STP connectivity) and a substance physico-chemical properties (i.e. K_{aw}). Figure S2.4.1 (SI) is a conceptual model of the discharge scenarios pathways used, and shows the inputs required for *Pangea*. Upon consumer use HPC substances can be emitted directly by volatilization (*Pangea* input to air), or disposed via domestic wastewater. Wastewater in turn can be treated in septic tanks (fraction considered removed and not modelled here), collected via a non-confined sewer (no removal), collected via confined sewer and possibly treated in WWTPs, or disposed untreated to receiving water bodies (direct discharge). Underlying data on discharge and treatment scenarios is held in ScenAT (Hodges et al., 2014, 2012). The sum of the fractions emitted treated and untreated to surface water contribute to the freshwater emission input into *Pangea*. Emissions to agricultural soil are

calculated from the fraction removed via sludge, estimated by SimpleTreat, and the extent of sludge-to-land application per country (Table S2, SI).

The methods for the emission inventories to air, water and soil of the four case study chemicals are summarized in SI. D5 is the only chemical that volatilize during consumer use. (Mackay et al., 2015) provide evidence that 90% of D5 will be emitted directly to the air compartment. For the remaining three chemicals, no volatilization is assumed and 100% is be disposed via wastewater to WWTP, to septic tank or directly to surface water. Evidence of in-sewer removal (elimination before entry into the STP) is only reported in the literature for LAS (about 50%, HERA, 2013; Matthijs et al., 1999). The value is applied to confined sewers either to WWTPs or direct discharge. The sludge-to-land pathway was most relevant for TCS in South Korea due to a high Log K_{ow}, high STP connectivity (SI), and sludge application to land practices.

3.4.2.3 From emissions to environmental fate and exposure

The *Pangea* Model is used to model fate and transport of substances. *Pangea* is a spatial multimedia compartmental model. It is built upon a computational engine, a Geographic Information System (GIS) engine, a set of Environmental Models (EMs), and a set of Environmental Processes Models (EPMs). The GIS engine (MATLAB/Python/ArcGIS) makes *Pangea* able to create multi-scale grids – global with local resolution where relevant – and to project spatial data quickly on a project-specific basis. EMs provide data for parameterizing media and transport models; they are based on external models, e.g. GEOS-Chem (Bey et al., 2001) for wind velocities, WWDRII (Vörösmarty, 2002) for the hydrology combining lakes and rivers in a global hydrological network (Helmes et al., 2012), and GlobCover (ESA GlobCover 2009 Project) for terrestrial composition. EPMs model fate and transport of chemical substances within (degradation) and between (advective and diffusive transport) media. *Pangea* creates an abstraction layer which allows to keep EPMs generic, i.e. they describe fate within and transport between abstract compartments. This allows to use EPMs from well established, multimedia models. *Pangea* currently implements USEtox (Hauschild et al., 2008) and IMPACT2002 (Margni et al., 2004; Pennington et al., 2005) first order EPMs. Finally, the computation engine builds a mathematical compartmental system, which can be solved for the dynamics and for the steady-state of environmental concentrations resulting from a set of emissions.

The media considered in *Pangea* for this study are air, fresh water, freshwater sediments, sea water, fresh water, natural land, agricultural land, the latter three resulting from a *Pangea*-specific aggregation of the 22 GlobCover types of land cover. The geometry of the system is defined by the following grids:

- A 3D atmospheric grid of 17 layers, approximately covering altitudes up to 15 km.

- A grid of clusters delineated by watersheds delivering to a global ocean. Watersheds composition includes freshwater, natural land, and agricultural land.
- A freshwater sediments grid compatible with the terrestrial grid.

***Pangea*: Refinement Potential and grids** – The mechanism for specifying characteristics of multi-scale grids is called Refinement Potential (RP) in *Pangea*. It is a scalar field (a raster in practice) whose value defines the "need for high resolution" at each point of the globe. This RP is integrated in an iterative procedure that starts from a low-resolution background grid, refines all grids cells whose integrated RP is above a user-defined threshold, and repeats the procedure until a desired depth/resolution is reached. The RP for this study (Fig. S2.5.1, SI) is a sum of weighted components that include population counts, proximity to rivers, intensity of emissions, and the selected a region of interest.

The grid obtained by refinement is called the *Results grid*. The atmospheric grid (Fig. S2.5.2, SI) is built upon this *Results* grid (as first layer) and all intermediary grids from the refinement process (stacked vertically, with deceasing resolution). The terrestrial grid (Fig. S2.5.3, SI) is obtained by clustering WWDRII native 0.5°×0.5° grid (by watershed, weakly constrained by the RP). We ultimately un-cluster all clusters over the region of interest, to enforce maximum/native resolution. Finally, the sediments grid is built to match the terrestrial grid, but limited to cells that contain freshwater.

3.4.2.4 Evaluation and model comparison

Comparison with USEtox – *Pangea* EPMs are currently based on USEtox and IMPACT2002. *Pangea* can parameterize and run USEtox automatically; we use this feature for performing a systematic model-model comparison between *Pangea* and USEtox, comparing a statistic (boxplot) of spatial PECs computed by *Pangea* with a single continental-scale PEC computed by USEtox (Southeast Asia, region 29).

Monitoring: data collection and comparison – To compare *Pangea* outputs with monitoring data, a literature review for georeferenced data was done for the four case study chemicals in the selected geographies in freshwater and sediments. When only maps were available with monitoring locations, georeferenced locations were manually identified and extracted using google maps. In some cases concentrations were read off graphs as the raw data was not available (i.e. Ramaswamy et al., 2011). Each article was reviewed and information on location names, sampling sites, sample time, season, flow, sampling procedure, analytical method, reported concentration and uncertainty (i.e. standard deviation

or range) were recorded. In many cases only mean/median, ranges and/or boxplots of measured concentrations were available so these were excluded from the final monitoring data set.

For TCS, 442 measured monitoring data points (265 fresh water; 176 sediments) were extracted across China, India, South Korea and Japan (Chen et al., 2014; Kimura et al., 2014; Liu et al., 2015; Ramaswamy et al., 2011; Ryu et al., 2014; Wang et al., 2012; X. K. Wang et al., 2014; Yu et al., 2011; Zhang et al., 2015). For methyl paraben, 303 measured monitoring data point (171 fresh water; 132 sediments) were extracted across China and Japan (Chen et al., 2014; Kimura et al., 2014; Liu et al., 2015; Yu et al., 2011; Zhang et al., 2015). For LAS, 30 measured monitoring data points (21 fresh water; 9 sediments), specific to urban areas, were extracted across China and Philippines (Eichhorn et al., 2001; Mu and Wen, 2013). No georeferenced data could be identified for D5 in the literature, therefore non-georeferenced data has been used for comparison. Full details of the monitoring data used in the comparison are available in Section S2.6 (SI).

Monitoring data are imported in *Pangea* and associated with relevant grid cells (location) and compartments (medium: freshwater, sediments). The time-average is computed at each monitoring site and for each medium. Modeled and monitored means and standard deviations are then computed per main stream (e.g. Yangtze, Songhua) for comparison. Pearson correlation, Spearman's rank correlation, root mean square error (RMSE), and bias are computed per main stream.

3.4.2.5 Spatial analysis and identification of key parameters

To analyze the relationship between emission sources, local hydrological conditions, and predicted environmental concentrations (PECs), we identify influential parameters that affect PECs, and characterize the relative importance between local sources and transfers from upstream cells in PECs predictions. For this purpose, we express modeled PECs at steady-state in a way that shows how it relates to equivalent dilution flows and to emissions. The mass balance for a given freshwater compartment *j* can be written as:

$$\frac{d(\mathrm{c}_j \mathrm{v}_j)}{dt} = \mathrm{s}_j^{tot} - \mathrm{k}_j^{tot} \mathrm{c}_j \mathrm{v}_j \tag{3.1}$$

where c_j [mg/m^3] is the concentration of a substance in the compartment, v_j [m^3] is the volume of fresh water, $\mathrm{k}_j^{tot} = \mathrm{k}_j^{adv} + \mathrm{k}_j^{sed} + \mathrm{k}_j^{vol} + \mathrm{k}_j^{deg}$ [1/s] is the overall removal rate coefficient (sum of the removal rate coefficients by advection, sedimentation and diffusion, volatilization, and degradation), and s_j^{tot} [mg/s] is the cumulative inputs, i.e. the sum of the local emission to freshwater compartment *j* plus the sum of all other inputs from neighboring, upstream cells or other media. At steady state, we get

$$0 = s_j^{tot} - k_j^{tot} c_j^{ss} v_j \quad \Rightarrow \quad c_j^{ss} = \frac{1}{k_j^{tot} v_j} \, s_j^{tot} \tag{3.2}$$

The predicted environmental concentration is therefore proportional to the *cumulative* inputs and to the inverse of the equivalent dilution flow $\dot{v}_j^{equ} \coloneqq k_j^{tot} v_j$ [m^3/s].

The contribution of the local emission source to the local concentration can be evaluated by performing a local, single-medium assessment, restricting Eq. 3.2 to:

$$c_j^{ss,local} = \frac{1}{k_j^{tot} v_j} \, s_j^{local} \tag{3.3}$$

where s_j^{local} is the local emission only (not accounting for contributions from upstream and from other media). This corresponds to a situation where no spatial, multimedia model is available, and only local sources and outputs to other media (losses) are accounted for. The locally evaluated environmental concentration is therefore proportional to the *local* emission and to the inverse of the equivalent dilution flow.

Equations 3.2 and 3.3 show that in both cases, both the source term and the inverse of the equivalent dilution flow can be interpreted as potentials for high PECs. It is therefore worth mapping them and comparing their spatial patterns with maps of PECs. It is also of interest to compare local estimates $c_j^{ss,local}$ and PECs c_j^{ss} through their ratio

$$\frac{c_j^{ss}}{c_j^{ss,local}} = \frac{m_j^{ss}}{s_j^{local}/k_j^{tot}} \geq 1 \tag{3.1}$$

where m_j^{ss} is the mass of substance in freshwater compartment *j*, and to determine for which substances and locations local estimates ($c_j^{ss,local}$) approximate well PECs (ratio close to 1).

3.5 Results

3.5.1 Predicted environmental concentrations

Figure 2 presents the predicted environmental concentrations (PECs) in air, fresh water, sediments, agricultural and natural soils, for all considered substances. Regarding the fractions of emissions to air, fresh water, and agricultural soils (top pie charts), LAS and TCS are mostly emitted to freshwater and to a lesser extent to agricultural soil via sludge-to-land application, while D5 is primarily emitted to air, with only 6% emitted to water and 1% emitted to agricultural soil, thus the importance of a multi-media approach. For D5, TCS and LAS, steady-state masses in the environment are primarily found in sediments (52% to 84% of the total mass), which act as a sink for hydrophobic substances and are in our simulations the medium with the lowest removal rate. This fraction of mass in sediments tends to be nevertheless lower for short-lived substances in freshwater such as LAS ($DT50_{water}$ = 0.3 d) for which a substantial fraction of the substance is degraded before being able to sediment. For methyl paraben ($DT50_{water}$ = 0.83 d), its relatively low predicted K_{oc} (8.6×10^{1} L kg^{-1}, EPIsuite) and short half-lives in both sediments and water lead to a low predicted mass in sediments.

Figure 2.C-E. maps a selection of PECs for the four substances. Freshwater median PEC for LAS ($\tilde{c} = 2.0$ mg/m^3) is two orders of magnitude higher than freshwater median PEC for TCS ($\tilde{c} = 7.8 \times 10^{-3}$ mg/m^3), methyl paraben ($\tilde{c} = 3.1 \times 10^{-3}$ mg/m^3), and D5 ($\tilde{c} = 1.1 \times 10^{-2}$ mg/m^3). The four maps of freshwater PECs show grey areas where the hydrological model is singular (no discharge); the hydrology in these regions is currently not accounted for in *Pangea* since the average estimated discharge by WWDRII is zero. This points to necessary improvements of *Pangea* hydrological model, as discussed below. Predicted concentration in agricultural land are higher in South Korea and China than in India (Figure 2D for LAS) reflecting differences in connectivity to WWTPs and sludge management practice, reflecting the higher emissions in these regions. Finally, we observe that D5 is the only substance for which a non-negligible fraction of the mass is found in air (12%), the highest air concentrations being observed downwind from highly populated areas of India, China and South Korea (Figure 2E). While 94% of emissions of D5 are atmospheric (90% consumer use-phase (Mackay et al. 2015), 4% from STP), the maximum mass is nevertheless found in sediments due to the substantially longer half-lives in water (60 days) and sediments (90 days) than in air (0.45 day).

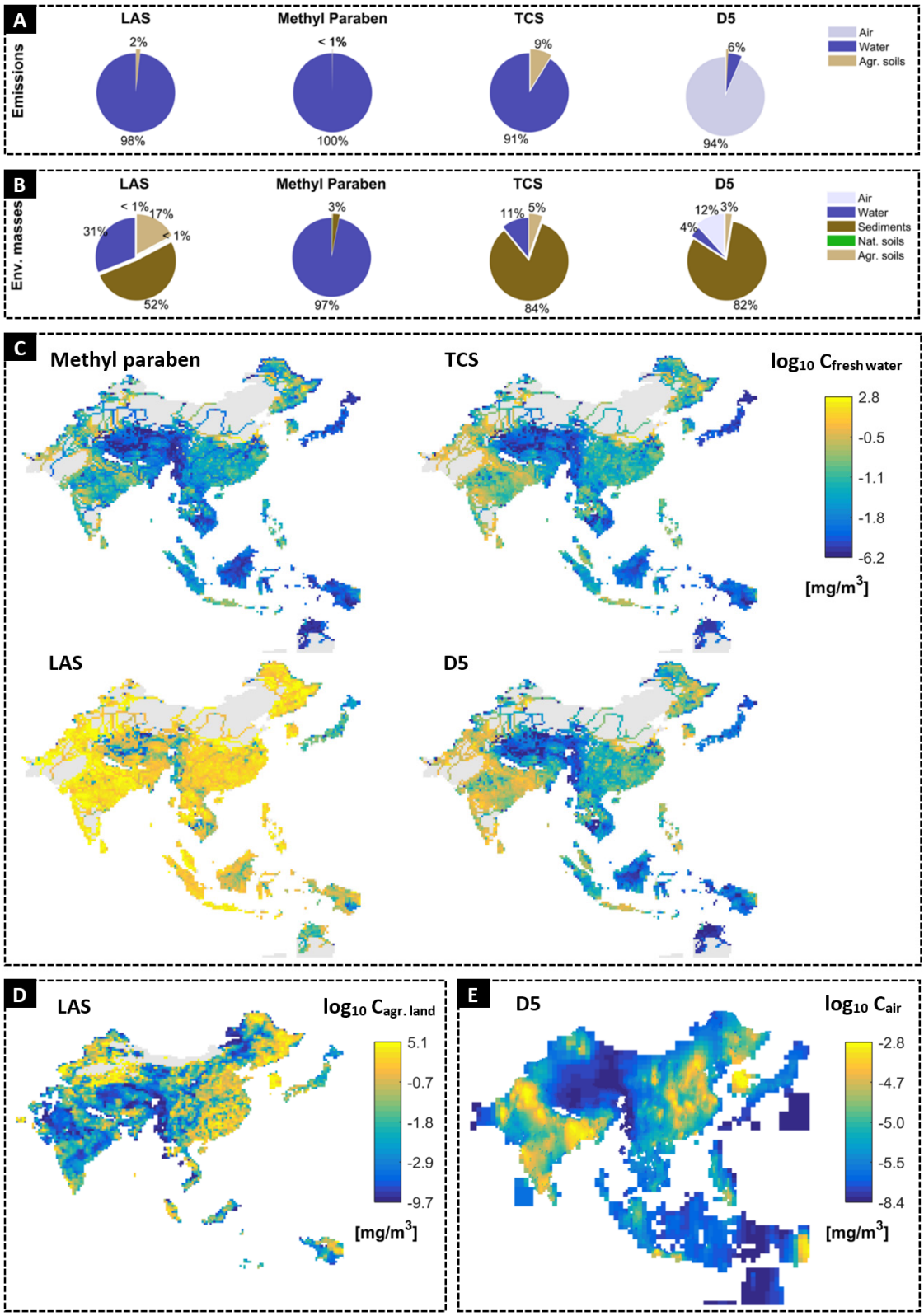


Figure 3.2 – Fractions of emissions to air, fresh water, and agricultural soils (A), fractions of environmental masses at steady-state in each medium (B) and maps of Pangea Predicted Environmental Concentrations in freshwater for all substances (C), in agricultural soils for LAS (D), and in air for D5 (E).

3.5.2 Comparison with USEtox and with measurements

Boxplots comparison modeled vs monitored – Figure 3 shows the comparison of modeled *Pangea* vs modeled USEtox (green bars) vs monitored concentrations in fresh water and in sediments, for the four selected substances. USEtox has no explicit sediment compartment, which is why there is no green bar for this medium, and we had no monitoring data available for D5. We observe a good agreement

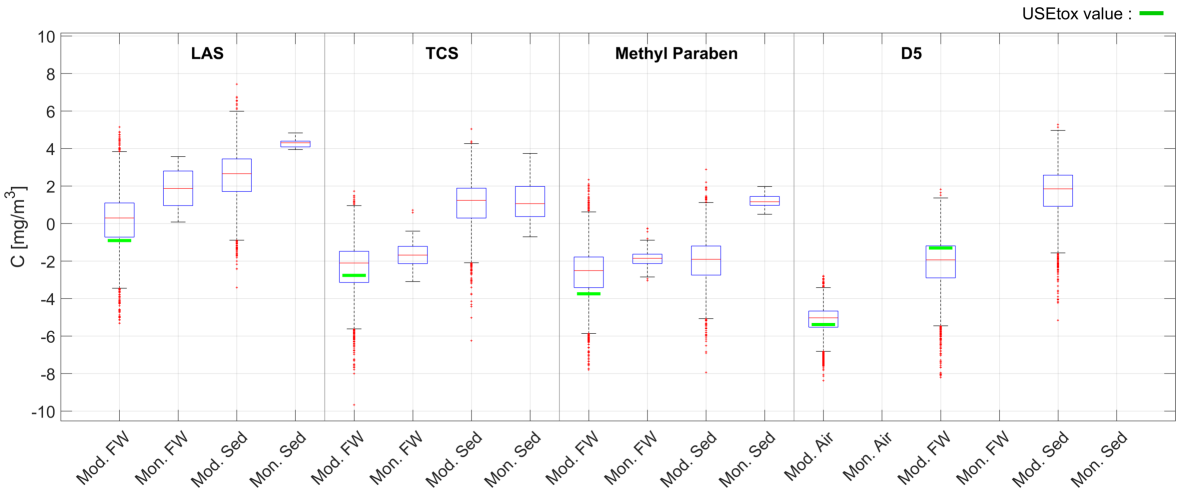


Figure 3.3 – Boxplot of air (D5 only), fresh water (FW) and sediments (Sed) PECs (Mod.) from *Pangea* and USEtox (air and fresh water), and monitoring data (Mon.). The central read mark in boxplots indicates the median (50th percentile), the bottom and top edges of boxes indicate the 25th and 75th percentiles respectively. Whiskers correspond to approximately ±2.7σ and 99.3 percent coverage for normally distributed data.

(maximum one order of magnitude difference) between *Pangea* and the average continental freshwater concentrations calculated by USEtox for all substances. The comparison between monitored and modeled values shows good agreement for TCS, with less than an order of magnitude difference between monitored and modeled medians in both media. For methyl paraben, there is also good agreement for concentrations in fresh water, but the modeled concentration in sediments are strongly underestimated compared to monitored values. Based on the present physical properties (K_{oc}, non-spatial f_{oc} and TSS, half-lives), the present model and separate runs made with SimpleBox (for verification, in addition to *Pangea* and USEtox) cannot explain these high sediment concentrations with methyl paraben. Further investigations are required to determine the causes for this behaviour specific to parabens.

Predicted LAS environmental concentrations ($\log_{10}\tilde{c} = 0.3$) are two to three orders of magnitude higher than the concentrations of the other three substances, which have comparable concentrations in fresh water ($\log_{10}\tilde{c}$ values of -2.1, -2.5, and -1.9, for TCS, methyl paraben, and D5 respectively). Yet, only a few measured LAS concentrations are available in Asia. The only monitored LAS concentrations found in the literature are urban values from (Eichhorn et al., 2001; Mu and Wen, 2013); the statistic is based

on 30 values overall (against 442 for TCS), and they are in the higher ranges, which is consistent with the high range of the values predicted by the model.

Comparison modeled vs monitored per stream – Figure 4 compares predicted concentrations with *Pangea* to monitored concentrations in fresh water and sediments for TCS and methyl paraben in each individual stream. Very good agreement between modeled and monitored values can be observed for TCS in freshwater (Figure 4.A), reflecting well the variation in freshwater concentrations between different rivers (Pearson correlation r = 0.82, Spearman rank correlation r_s = 0.84, RMSE = 0.31). River Tamiraparani in south India is the one with highest predicted and measured concentration, while the

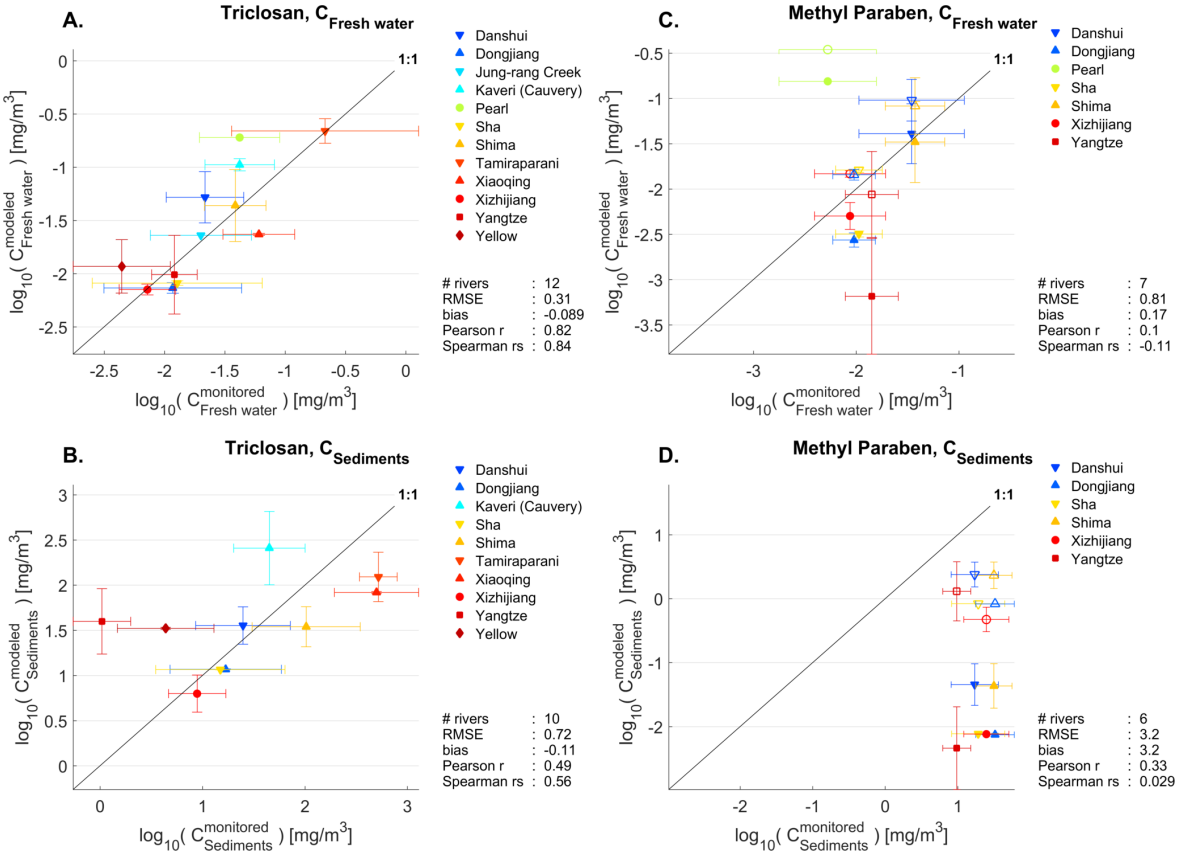


Figure 3.4 – Comparison of Pangea predicted fresh water and sediment PECs with monitored concentrations for TCS and methyl paraben in major water basins across Asia. Markers and bars indicate the mean and standard deviation of predicted and monitored values for the monitored sites along each river. Dashed error bars in plots C and D are based on modeled values obtain with a tenfold increase in half-lives in fresh water and sediments and in K_{oc}.

biggest rivers (Yellow, Yangtze) generally display lower concentrations due to higher dilutions. TCS is the substance with the largest number of available measures, with 253 exploitable values for fresh water, which reduces the influence of outlier and the potential impact of sampling locations that may fall in singular regions not covered by the model.

Figure 4.B shows a fairly good agreement between average monitored and predicted TCS in sediments that was reflected in the boxplot of Figure 3, but with a higher discrepancy than for freshwater and a seeming tendency for underestimation (RMSE = 0.72, bias = -0.11 log units).

Plain line/symbols error bars in Figure 4.C shows a moderate agreement for methyl paraben in freshwater when comparing the individual rivers (RMSE = 0.81). Modeled values are less than half an order of magnitude off monitored values, except for the Pearl (south of China, very seasonal) and for the Yangtze rivers that are ~1.8 orders of magnitude off.

Plain lines/symbols error bars in Figure 4.D show a clear, constant discrepancy for methyl paraben in sediments (~3.5 orders of magnitude).

3.5.3 Spatial variations and key parameters

Analysis of spatial variations in freshwater concentrations – For the four selected substances in the Yangtze River basin, Figure 5 maps spatial distributions of emissions, of inverse equivalent dilution

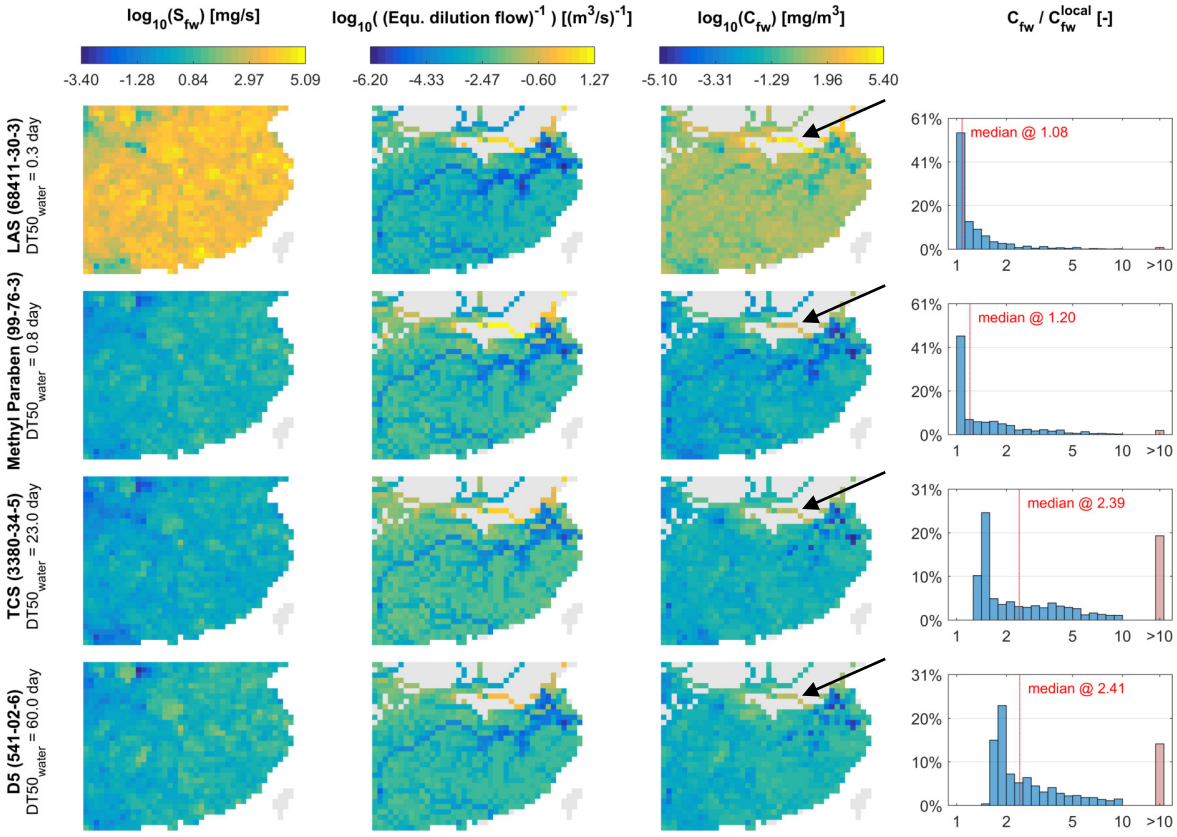


Figure 3.5 – Maps of emissions to fresh water [mg/s], inverse equivalent dilution flows [$(m^3/s)^{-1}$], concentrations in fresh water [mg/m^3], and histograms of mass ratios due to "total vs local" emissions, for the four selected substances in the Yangtze basin.

flows, and of concentrations in fresh water. We observe two to three orders of magnitude higher emission levels and environmental concentrations for LAS than for the three other substances. For the short-lived substances (LAS, methyl paraben), spatial variations in freshwater concentrations primarily reflect variations in the equivalent volume of dilution, with higher concentrations in regions with a low volume of dilution. A high concentration is particularly noticeable in the northern part of the maps (arrows), which is a region where the hydrological model provides extremely low to no flow values. It corresponds to the Ying River, a tributary of the Hao He River, a river characterized at the highest level of pollution 5+ (Report on the State of the Environment in China, 2006). In contrast, a "flushing effect" is also noticeable for shorter-lived substances in water (LAS and methyl paraben, $DT50_{water}$ = 0.29 and 0.83 days respectively): concentrations are lower along the main stream of the Yangtze River, with larger flows, than over their tributaries and water basins (maps in column 3). This relates directly to high dilution flows observed in the second column, along the main stream. This effect is reduced or disappears for more persistent substances such as TCS and D5 ($DT50_{water}$ = 23 and 60 days respectively), leading to more uniform concentrations even along the main stream of the Yangtze river and the rest of the catchments: these substances travel farther, and there is no flushing because there is a cumulative effect of emissions in upstream cells.

Regions with the highest freshwater concentrations, such as the Ying River, are not necessarily regions with largest emissions: they generally correspond to regions with medium to high levels of emission and with low volume of dilution. The consequence of the flushing effect is that higher concentrations occur in tributaries rather than in the largest rivers, which points to the need for diversifying the selection of monitoring sites that have often primarily focused on the main streams.

Identification of key parameters – To further investigate the importance of the contributions of upstream versus local emissions to the PEC in each cell, Figure 5 presents in its last column histograms of the ratio of equation 2.4, the PEC divided by the concentration derived from local emissions only (% of total cells count in mapped region vertically, versus ratio horizontally). For short-lived substances, we observe that local evaluations are comparable to PECs (ratios close to 1 for LAS and methyl paraben). For these substances, transport from upstream does not play a significant role and concentrations are dominated by local emissions. On the contrary, for longer-lived substances such as TCS and D5, PECs are substantially higher than concentrations evaluated based on local emissions only, by a median factor 2.4, and by a factor greater than 10 for 14% to 20% of compartments (red bars). For these substances, transport from upstream regions plays a significant role. Maps in columns two and three show graphically this discrepancy: given the relative homogeneity of emissions to fresh water over the Yangtze basin, for short-lived substances (LAS, methyl paraben) environmental concentrations are

directly proportional to the inverse of the dilution flow, and we can distinguish the pattern of inverse dilution flows in the map of concentrations (flushing effect); for longer-lived substances (TCS and D5), contributions of upstream sources matter (in the s_j^{tot} term of Eq. 2.2), which reduces the pure dilution or flushing effect along the Yangtze River.

3.6 Discussion

3.6.1 Limitations and insights of the comparison between modeled and monitored data

Comparability between modelled and monitored data – One important preliminary consideration and insights from this effort are worth mentioning. The routing of the hydrological network on WWDRII 0.5°×0.5° native grid may lead to important spatial discrepancies and misassociations between PECs (expressed on the grid) and locations of monitoring sites. Figure 6 shows the approximate location of the Yangtze River stream as the curvy, light blue line, which matches well the geographical location of monitoring locations (red points) along the river. It also displays the WWDRII native grid with thick grey lines and the river network as dark blue lines. The diagonal connection between cells 16255 and 16384 represents the Yangtze River "as seen/routed by WWDRII". We observe that the diagonal routing creates a mismatch between the hydrological network and monitoring locations, because it associates the monitoring location of cell 16256, which represents a set of 20 monitoring values on the Yangtze, with another stream routed in cell 16256. The monitoring/routing discrepancy was corrected manually (for 23 streams), but we are generally reaching the limit of what can be achieved with a hydrological model based on a 0.5°x0.5° square grid. To systematically solve this problem, *Pangea* needs to transition towards a higher resolution global hydrological data set. Figure 3.6 shows that HydroBASINS (Lehner and Grill, 2013), represented by polygons with thin grey boundaries, meets this requirement. The color of polygons provides an indication of the discharge, which shows that the Yangtze River "as seen by HydroBASINS" matches well the stream delineation. This data set has significant advantages: fine

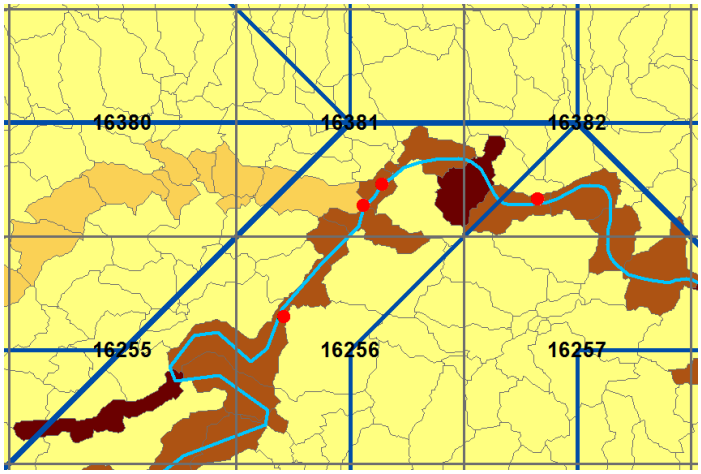


Figure 3.6 – Routing and monitoring mismatch.

delineation of watersheds, higher resolution, consistent parameterization, consistent scale transition mechanisms, and it offers a much better concordance with monitoring locations.

Modeling uncertainty. Broadly, there are two main modeling steps involved (calculation of emission inventories and of environmental fate and transport) although this study primarily focuses on the latter. Limitations are inherent to the length and complexity of the full modelling chain, introducing assumptions and associated uncertainties at each step. Sales and products data inventories are distributed spatially using night time light as a proxy for gross domestic product (GDP) (Hodges et al., 2014). The down-the-drain model uses a single standard sewage treatment plant (STP) model to estimate removal efficiency (Struijs, 2014) and other simple assumption on the wastewater and sludge management scenarios that determine compartmental emission inventories (SI).

Pangea determines a steady-state solution that defines a single cell-averaged concentration assuming uniform emissions and well mixed compartments within each cell. This contrasts with monitoring data that can be based on a few or even a single point in time for substances with limited monitoring data, and are measured at specific sites. Sampling is moreover often performed close to river banks, and often with a single monitoring site per *Pangea* cell, which may have been selected at a location close to effluent discharges. Lateral effluent discharges need long distances to reach complete cross mixing. A cross-sectional composite sample taken where the river crosses adjacent cells would more appropriate for comparison with modeled concentrations.

Finally, the global multimedia fate and transport model incorporates first order EPMs and significantly simplifies complex phenomena such as sedimentation, resuspension and burial, in rivers that are subject to high temporal and spatial variability.

Uncertainty and variability in input parameters. The next level of uncertainty is introduced at the selection of input data and parameters. Sources of emissions other than those related to HPC products are not accounted for. While HPC products are the main sources for the four selected chemicals, industrial sources could be important for LAS and D5 (10-20%; Brooke et al., 2009; HERA, 2013). There is a very large variability in total suspended solids (TSS) in Asian rivers, that is not accounted for in the present model: the USEtox default global TSS average is 15mg/L, and was set to 100mg/L in *Pangea* for this study, to reflect the higher average TSS content in Asian rivers, ranging typically from 12 mg/L to 23,900 mg/L, and up to 48,000 mg/L for the Yellow River (Likens, 2010). There is therefore a need to systematically collect TSS concentrations (and organic content) for relevant rivers, especially for tributaries for which spatial data availability remains sparse. In addition to influencing sedimentation

rates, this is an important factor of influence for substances that degrade in water by photolysis (such as TCS), since photodegradation is reduced with increasing turbidity.

Model estimates for methyl paraben in sediments, presented in Figure 4.D, are about 4 orders of magnitudes lower than reported monitored values. Observed values may therefore challenge the predicted low K_{oc} and half-lives in sediments. Under these conditions, they may also question the domain of validity of the statement made in TOXNET HSDB (2017, Environmental Fate & Exposure section) that "*methyl paraben is not expected to adsorb to suspended solids and sediment*". A sensitivity study was carried out for methyl paraben to study the influence of chemical properties. Dashed line error bars with non-filled symbols in Figure 4.C and 4.D correspond to a tenfold increase of half-lives and K_{oc}: $DT50_{water}$=8.3d, $DT50_{sediments}$=83d, and K_{oc}=896L/Kg, which brings freshwater PECs associated with the Yangtze river close to monitored level, and reduces a 4 orders of magnitude underestimation of sediment PECs to 2 orders of magnitude. This suggests that the discrepancy between observed and predicted concentrations can be substantially reduced by using half-lives, Koc, and TSS values greater (but not unrealistically) than predicted by EPIsuite. This emphasizes the need to have measured K_{oc} and degradation rates under local conditions, and to analyse in further detail causes and mechanisms for higher persistence than expected.

3.6.2 Opportunities for large-scale, high-resolution modelling

Model functionality and performance – Coupling ScenAT/Simple Treat and *Pangea* provides a functional multimedia framework for estimating PECs over all of Asia. The overall comparison of modeled vs monitored data shows good agreement for PECs in freshwater for both TCS and methyl parabens, and an order of magnitude difference for PECs in sediments for TCS only. PECs in sediments for parabens are consistently underestimated by *Pangea*, USEtox (back calculated), and SimpleBox when using the estimated chemical properties. This underlines the importance to further investigate the main processes that are responsible for removal, and to measure in-situ degradation rate and partition coefficients. Considering the nature and the length and complexity of the full modelling chain, the agreement at the level of individual rivers for TCS, the substance with the largest number of measured concentrations available for Asian streams, is very promising.

PECs for LAS are at least two orders of magnitude higher than PECs for D5, in both fresh water and sediments. While a considerable effort is undertaken to analyze TCS and D5, we found very little environmental data of LAS and generally of high volume surfactants in Asia in the literature. Surfactants were among the first down-the-drain chemicals to be prioritized and investigated in Europe and North America. Possibly, the emphasis of current monitoring programs in Asia on so-called emerging chemicals reflects a global scientific trend rather than sound prioritization.

Towards higher-tier modeling of ecological scenarios – *Pangea* can currently help in identifying areas of high exposure, e.g. the region of the Ying River, north of the Yangtze River, which is characterized by extremely low to no flow values, and corresponds to an area characterized as level 5+ pollution. Yet, the WWDRII-based *Pangea* Hydrological Model (PHM) has limitations inherent to the underlying hydrological model, which provides pathological volumes of water or flows in regions that are hydrologically singular (typically around the Ying River). These regions appear in grey in Figures 2 and 5. *Pangea* does not create fresh water compartments in these regions; it is therefore not able to provide predicted concentrations in regions of low to no flow, and hence of potentially high concern. This indicates that we need a finer underlying hydrological model, that includes regions with lower average yearly flows (even with a high uncertainty), to allow *Pangea* to define fresh water and sediments compartments, and to provide screening level concentration over these intrinsically difficult regions to model.

The key advantage of *Pangea* is that, while being global, multimedia, and spatial, it is parsimonious from a user perspective: it does not require to define more than emissions levels, physical/chemical properties, and criteria for building multi-scale grids. It uses the parsimonious USEtox/SimpleBox type of parameters and environmental process models, while offering the possibility to spatially differentiate the input data when spatial data become available.

Our framework constitutes therefore a milestone towards higher tier modeling approaches for ecological risk assessment. The identification of worst case (sub)catchments in combination with source apportionment analysis, for a given chemical, can point to worst case (sub)catchments and to the required scale of exposure assessment for further investigations.

3.7 Acknowledgements

Funding for the University of Michigan work was provided by the Unilever sponsored agreement CH-2012-1262 on “Ecological fate and assessment of Home and Personal care products in Asia (Ecohope)”. We thank Lei Huang (University of Michigan) and Peter Fantke (Technical University Denmark) for their support in parametrizing the model, and Bernhard Lehner (McGill University, Montreal) for providing the HydroBASINS data set.

Chapter 4 Radial Intake Fraction (fate and exposure)

4.1 Foreword

This chapter will be submitted to the journal *Environmental Science & Technology*, under the title "Multi-scale spatial modeling of human exposure from local sources to global scale", by:

Cedric Wannaz[a,*], Peter Fantke[b], Olivier Jolliet[a]

[a] School of Public Health (SPH), University of Michigan, 6622 SPH Tower, 1415 Washington Heights, Ann Arbor, Michigan 48109-2029, United States

[b] Quantitative Sustainability Assessment, Department of Management Engineering, Technical University of Denmark, Produktionstorvet 426, 2800 Kgs. Lyngby, Denmark

4.2 Abstract

Exposure studies, used as a basis for human health risk and impact assessments of chemical emissions, are largely performed locally or regionally. It is usually not known how global impacts resulting from exposure to a point source emission compare to local impacts. In this paper, we introduce *Pangea*, an innovative multi-scale, spatial multimedia fate and exposure assessment model. We first study local to global population exposure associated with emissions from 126 point sources matching locations of waste-to-energy plants across France. Results for three chemical substances with distinct physicochemical properties are expressed as the evolution of the population intake fraction through inhalation and ingestion as a function of the distance from the emission source. Less than 20% of the global population intake through inhalation (median of 126 emission scenarios) can occur within a 100 km radius from the source. This suggests that, by neglecting distant exposure to low exposure levels, local assessments might only account for a small fraction of global cumulative intake. We finally evaluate ~30,000 emission scenarios (~10,000 locations across France for the three studied substances) covering France more densely, and we plot maps of global intakes through inhalation and ingestion associated with each emission location. These maps are relevant for determining which locations minimize global intakes, for each substance and exposure route.

Keywords: spatial multimedia model; multi-pathway exposure assessment; population intake fraction; environmental fate of chemicals; local to global modeling

4.3 Introduction

4.3.1 Background

For human health risk and impact assessments of chemical emissions, exposure studies either focus on local or regional scales, or assess larger areas at the expense of spatial resolution. How impacts on human health at the global scale compare to local impacts for the same emission source has not been systematically studied. We therefore focus on assessing the spatial distribution of the *intake fraction* – the fraction of an emission that is ultimately taken in by an exposed human population – from local to global scale.

Exposure studies for risk or health impact assessments are generally focused on regions with high pollutant concentrations and model fate and transport of pollutants within close proximity of emission sources, typically within a radius of 5 km to 30 km (EEA, 2007; Silverman et al., 2007; Zou, 2010). Such studies provide estimates of human exposure and subsequent intake of a specific pollutant in a local area, but neglect potentially important intakes associated with populations exposed on a larger scale. Lohman and Seigneur (2001) show for example that up to 90% of polychlorinated dibenzo-*p*-dioxins and polychlorinated dibenzofurans (PCDD/F) emitted from stacks might be deposited farther than 100 km away from sources. Other studies analyzing long range transport (Breivik et al., 2016; Hageman et al., 2015; Hollander et al., 2008; Kuramochi et al., 2014; Mackay et al., 2015; Scheringer et al., 2012, 2009; Zarfl et al., 2011) also suggest that local assessments might only account for a fraction of global exposure and related impacts, especially for chemicals that are both mobile and persistent.

In contrast, exposure studies in Life Cycle Impact Assessment (LCIA) generally covers a larger scale (100 km to global), is screening oriented (fast evaluation of a large number of chemicals), and uses multimedia, multi-pathway exposure (EEA, 2007; McKone and MacLeod, 2003; Scheringer and Wania, 2003). The lower spatial resolution of LCIA exposure studies limits their ability to account for important differences in the spatial distribution of parameters exposure.

To better evaluate exposure to chemical emissions for informing decision makers, we need the advantages of LCIA approaches combined with encompassing relevant spatial details from local to global scales. Both fine resolution at relevant places (e.g. around emissions sources or highly populated areas) and flexibility to easily change the region of focus are therefore essential.

To meet these requirements, we consider three categories of spatial models as previously analyzed and compared (EEA, 2007; Hansen et al., 2006; Hollander et al., 2008; Lammel et al., 2007; Scheringer and Wania, 2003). The first category, *chemical transport models* (CTMs), consists of models with generally high spatial or temporal resolution, such as the atmospheric chemical transport and air quality models AERMOD (Cimorelli et al., 2005; Perry et al., 2005), ADMS (Cambridge Environmental Research Consultants) and GEOS-Chem (Bey et al., 2001). While being essentially focused on the atmospheric dynamics of chemicals, some CTMs were extended to take other media into account, e.g. CMAQ (CMAS), ECHAM5/6 (Roeckner et al., 2003; Stevens et al., 2013), and EMEP-MSCE. Models from this category are generally slow at runtime (a few minutes to several hours or days per pollutant/scenario), require many input parameters and data, are rarely global, and often model other media only as reservoirs, sometimes with no reverse flows (e.g. deposition without re-volatilization) or exchanges between reservoirs. Moreover, none of these models covers sufficiently or at all population-level multi-pathway exposure.

The second category of models consists of *spatialized multimedia box models*, accounting for intra- and inter-media phenomena generally described by first order differential equations, such as ChemRange (Scheringer, 1996), Globo-POP (Wania and Mackay, 2000), ClimoChem (Scheringer et al., 2000), G-CIEMS (Suzuki et al., 2004), NIAES-MMM-Global (Wei et al., 2008), BETR variants (MacLeod et al., 2005, 2001, Prevedouros et al., 2004a, 2004b; Woodfine et al., 2001), and IMPACT variants (Humbert et al., 2009; Jolliet et al., 2008; Margni et al., 2004; Pennington et al., 2005). These models were often parameterized using the output of models from the first category (e.g. wind fields from GEOS-Chem for parameterizing IMPACT). While being well suited for screening large numbers of chemicals, i.e. computationally light and fast at runtime (seconds to hours for multiple chemicals), these models have other limitations in common. They are essentially based on fixed grids defined during their development phase that lead to specific parameterizations. This makes it difficult and time consuming to change the region of focus or even grids resolution and geometry, which is essential for capturing the spatial variability of parameters relevant to exposure. Large and especially global scale multimedia box models are generally low resolution. This introduces artificial dilution artifacts, lack of discrimination power between sources and receptors, and prevents an accurate characterization of regions of interest (typically highly populated regions and vicinity of emissions sources). This lower resolution is in particular problematic for studying exposure in areas with sharp spatial variations in population densities. In addition, most of the global spatial multimedia models do not cover multi-pathway human exposure.

A third category of GIS-based models recently emerged (Pistocchi, 2014), which can achieve very high spatial resolution. When based on raster algebra, however, they can become data and computation

intensive, especially in contexts where both high resolution at specific locations and global scale coverage are required. While multi-media coupling (with feedback) is possible, it remains rather complicated to achieve, especially if one needs to tailor their parameterization on a per study basis.

In conclusion, there is a need for an innovative multimedia model that combines the advantages of the three aforementioned model categories. This model must enable high resolution at locations of interest and contrast local to global impacts of pollutant emissions based on multimedia transport and fate, as well as multi-pathway exposure. In addition, this model must automatize GIS processing to be flexible enough and readily adapted to specific locations or grid geometries, allowing for wider applicability.

4.3.2 Objectives

To address these needs, we propose a flexible multi-scale, spatially explicit multimedia model that meets the aforementioned criteria – local to global scale, pollutant fate and population exposure, flexible grids – well-suited for comparing human intake from local to global scales. More specifically, we aim to: (1) present the structure and elements of the newly developed multi-scale *Pangea* model, (2) evaluate the local to global fate and exposure associated with emissions of selected pollutants, analyzing the magnitude of the intake as a function of chemical properties and distance from a set of 126 point sources, for relevant exposure pathways, and (3) assess ~30,000 scenarios to build maps of global intake fractions associated with emission points that cover densely France.

4.4 Methods

4.4.1 The *Pangea* Model

Conceptually, *Pangea* is a compartmental model (2nd model category) that incorporates a GIS engine (3rd category), and aims at making a step towards CTMs (1st category) spatial resolutions, while maintaining the ability to perform fast evaluations.

Model structure – Multimedia "box" models consist of a set of compartments representing relevant environmental media (e.g. atmosphere, freshwater, specific land covers), associated with a set of environmental processes models (EPMs) (Bachmann, 2006; EEA, 2007; Mackay et al., 2001; MacLeod et al., 2010; Rosenbaum et al., 2007; Scheringer and Wania, 2003). EPMs describe interactions between media, i.e. diffusive and advective flows, as well as internal processes, such as degradation. *Pangea* extends this structure with the ability to create multi-scale grids and to project spatial data onto these grids at runtime. This allows building grids adapted to the specific characteristics and aims of each study or scenario and, in our context, enables the comparison of local and global impacts of pollutant emissions within the same, consistent framework.

The structure of *Pangea* is outlined in Figure 4.1. Geo-referenced data describing geographic features in the *natural system* (e.g. atmosphere and terrestrial coverage) are gridded by a GIS engine based on MATLAB Mapping Toolbox and Python/ArcGIS, which outputs multi-scale grids, projected data, and geometric and topological information. This process yields a *geometric system* of grid cells with homogeneous and inhomogeneous content; terrestrial cells, for example, may be composed of several land covers and freshwater. This system is transformed into a *virtual system* of homogeneous compartments using a "re-indexing" engine that splits inhomogeneous cells into their homogeneous media components according to the proportion of media types in each cell. The virtual system is well suited for defining a compartmental system (Jacquez and Simon, 1993) by a set of first order linear differential equations that describe the evolution of the mass of a pollutant associated with each compartment. While *Pangea* can also solve the dynamics of such system, we focus this study on the steady-state solution of a system with constant coefficients. The solution is combined with exposure

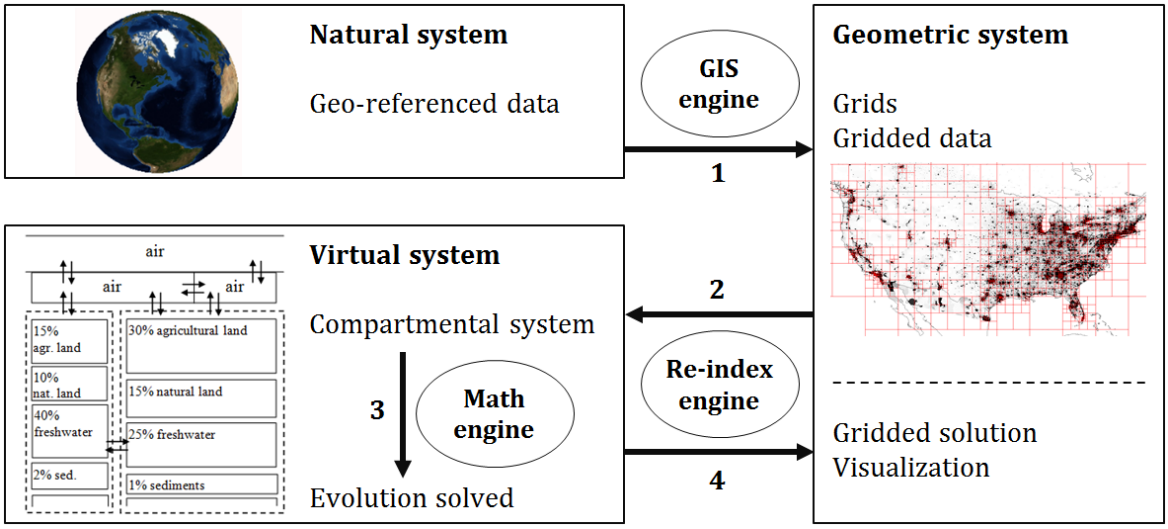


Figure 4.1 – Structure of *Pangea* illustrating the sequence of operations and utilized engines.

data to yield a distribution of human intake. Ultimately, all relevant quantities (e.g. masses, concentrations, intakes, intake fractions) in the virtual system are re-indexed back to obtain spatial distributions in the initial geometric system, and can finally be visualized as maps.

The media considered in this study are air, ocean water, sediments, freshwater, natural land, and agricultural land, the latter three resulting from a *Pangea*-specific aggregation of the 22 GlobCover (ESA and GlobCover, 2009) types of land cover. The geometric system consists in the following grids:

- A 3D atmospheric grid of 17 layers, approximately covering altitudes from 100 m to 15 km.

- A terrestrial grid of clusters delineated by watersheds delivering to a global ocean. Watersheds composition includes freshwater, natural land, and agricultural land.
- A freshwater sediments grid compatible with the terrestrial grid.

An additional grid, called *results grid*, is built for expressing final results on a common grid by projection, for comparison.

Chemical fate and transport modeling – Each spatial parameter describing a grid cell property (e.g. volume) or a projected dataset (e.g. population per grid cell) is expressed as a vector whose size equals the number of cells of the associated grid. Similarly, each parameter that describes interaction between two grids (e.g. advection between atmospheric grid cells or deposition from atmosphere to freshwater grid cells) is expressed as a matrix whose size is defined by the size of the grid(s) involved (e.g. *n* atmosphere cells × *m* terrestrial cells). These grid-specific vectors and matrices characterize the geometric system. To efficiently parameterize a global mass balance equation based on generic EPMs, we perform a re-indexing operation that splits vectors and matrices from the geometric system according to the proportion of each medium in each grid cell, and transforms them into block vectors/matrices structured by medium type. Re-indexed vectors and matrices, whose elements represent homogeneous compartments, define the *virtual system*. Its size is noted n^v and corresponds to the total number of compartments in the system ($n^v = n^v_{air} + n^v_{fresh\ water} + n^v_{sediments} + ..$). We note n^{es} the number of emission scenarios for a given system, where an emission scenario defines a given distribution of emission sources. We restrict the description of the virtual system dynamics to a system of autonomous, first order differential equations, with constant coefficients:

$$\frac{\mathrm{d}\mathbf{M}^v(t)}{\mathrm{d}t} = \mathbf{K}\,\mathbf{M}^v(t) + \mathbf{S}^v \tag{4.1}$$

with, $\mathbf{S}^v \in \mathbb{R}^{n^v \times n^{es}}$ an array of vectors of constant emission scenarios [$\mathrm{kg\,s^{-1}}$] written in column, $\mathbf{M}^v(t) \in \mathbb{R}^{n^v \times n^{es}}$ the corresponding array of vectors of masses of chemical [kg] at time t [s] written in column, and $\mathbf{K} \in \mathbb{R}^{n^v \times n^v}$ a matrix of transfer and elimination rate coefficients [$\mathrm{s^{-1}}$]. **K** is a sparse matrix, with dimensions typically in the range of 70,000×70,000 to 500,000×500,000; it can be factorized and the solving of Eq. 4.1 can be parallelized in cases where n^{es} is large (e.g. 10,000 as in the case study described further below). While being mathematically simple and easy to solve numerically, Eq. 4.1 is complex to build, because components of **K** depend on spatial parameters (e.g. contact surface areas between mismatching cells geometries, volumes, projected spatial datasets), which are specific to each configuration of grids. The consequent work (several person-months when done manually) is usually performed only once for models with fixed grids. The GIS and computation engines integrated in

Pangea, instead, provide a solution for addressing this challenging task dynamically, at runtime, with a re-projection of all spatial data (e.g. land cover, demographic data, atmospheric flows, flows through the freshwater network) each time a new grid is generated. Based on spatial parameters and projected data, environmental non-spatial parameters, and physicochemical parameters, a set of EPMs is used to compute the components of the **K** matrix. EPMs are specific to each medium for elimination processes (degradation), to each pair of media for transfer processes (advection and diffusion), and to each exposure pathway (inhalation and all ingestion pathways). EPMs are currently based on IMPACT 2002(Margni et al., 2004; Pennington et al., 2005) and USEtox,(Rosenbaum et al., 2008) adapted for taking spatial data into account. A description of media, processes, and EPMs is given in the Supporting Information (SI), Section D.2.1.

The solution of Eq. 4.1 is the distribution of pollutant mass as a function of time, expressed in the virtual system, $\mathbf{M}^{v}(t)$. *Pangea* can provide steady-state and dynamic solutions; for the scope of this study, we focus on the steady-state solution, $\mathbf{M}_{ss}^{v}$. This solution of the fate and transport is back re-indexed to relevant grids to get e.g. $\mathbf{M}_{ss}^{atm,L1}$, the spatial distribution of masses in the first layer of the atmospheric grid, or $\mathbf{M}_{ss}^{fw}$, the spatial distribution of masses in the global fresh water network.

Human exposure modeling – Exposure pathways considered in this study are inhalation of air, and ingestion of freshwater and food (fish, meat, milk, below-ground produce, and above-ground produce). The population intake through both inhalation and ingestion is computed and defined as:

$$\mathbf{IN}_{ss}^{v} = f(\mathbf{p}^{v}, \mathbf{IR}^{v}, \mathbf{BAF}^{v}, \mathbf{C}_{ss}^{v}) \tag{4.2}$$

where $\mathbf{C}_{ss}^{v}$ is the array of environmental concentrations corresponding to $\mathbf{M}_{ss}^{v}$, $\mathbf{BAF}^{v}$ is an array of generalized bioaccumulation factors (that can be interpreted as an application from environmental concentrations to concentrations in air, water, and food items), $\mathbf{IR}^{v}$ is an array of generalized individual intake rates (that can be interpreted as an application from concentrations in air, water, and food items, to masses taken in), $\mathbf{p}^{v}$ is a vector of population counts, f is an appropriate product between its arguments, and $\mathbf{IN}_{ss}^{v} \in \mathbb{R}^{n^{v} \times n^{es} \times n^{ep}}$ is a 3D array of population intakes [kg s^{-1}] with n^{ep} the number of exposure pathways (inhalation of air, ingestion of drinking water, ingestion of different food items). Finally, the population intake fraction (iF) is defined as:

$$(\mathbf{iF}_{ss}^{v})_{ijk} = \left(\sum_{r=1}^{n^{v}} \mathbf{S}_{rj}^{v} \right)^{-1} (\mathbf{IN}_{ss}^{v})_{ijk} \tag{4.3}$$

Fixing the emission scenario (index j, 2nd dimension), this formulation allows to aggregate population intake fractions over compartments (sum over index i, 1st dimension) and/or exposure pathways (sum

over index k, 3rd dimension). *Pangea* considers only the exposure through inhalation to be direct. All other exposure pathways, including water consumption), are considered production based. We define the total *production based population intake* and *intake fraction* as the sum of all intakes and intake fractions inducted by production (of fresh water and food items). Direct and production-based intake fractions cannot be summed; the former occurs locally, whereas the latter happens wherever production items are consumed ultimately. The missing link is a trade model such as the one implemented in IMPACT World,(Shaked, 2011) not implemented in Pangea because of the difficulty to parameterize it globally at higher resolution (multi-scale) than country level.

Input datasets – Table 4.1 summarizes datasets involved in the process of grids creation, parameters for the EPMs, and for computing population exposure. The GIS engine performing data projection at runtime makes the selection of datasets flexible; it can be study-specific and will evolve with data availability.

Table 4.1 – Major datasets involved in the process of grids creation, parameters for the EPMs, and for computing population exposure. Section D.2.2.1 (SI) provides a more detailed description of main datasets.

Dataset	***Description***
GEOS-Chem	Yearly bidirectional averages on a six-hourly basis were computed based on a GEOS-Chem run for the reference year 2005.
GlobCover	The 22 land cover categories are aggregated into *Pangea*-specific categories.
WWDRII	World Water Development Report II. Gridded 0.5°×0.5° water network used as a basis for computing flows between watersheds, as adapted by Helmes et al. (2012)
FAOSTAT	Global food production per country in 2009.
USEPA EPI Suite	Physicochemical properties of pollutants.
USEtox	Basis for selected intra- and inter-compartment processes and default parameters.
IMPACT 2002	Basis for selected intra- and inter-compartments processes and default parameters.
CIESIN	Raster of population counts for 2005.
ISWA	Location and capacity of 126 waste-to-energy-plants in France.
ESRI	Data and Maps 9.3 for country boundaries.

4.4.2 Local versus global impacts of pollutants

Case study design – The comparison of local and global impacts of selected chemical substances is based on atmospheric emissions from 126 point sources, representing waste-to-energy plants' stack, located across France, mapped using black dots in Figure 4.2. Locations were determined based on the ISWA report (ISWA, 2012).

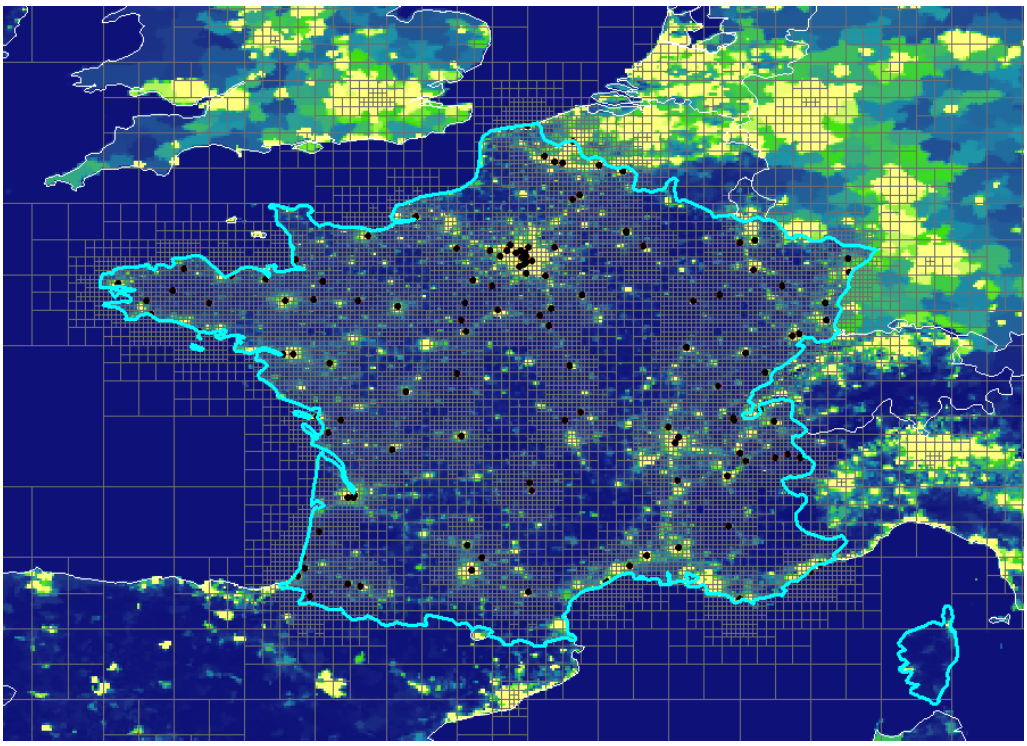


Figure 4.2 – 126 point sources, representing waste-to-energy plants' stack, overlaying a map of population density that shows the spatial distribution of the population.

Three substances were selected to cover distinct dominant impact pathways, and to represent different classes of pollutants: 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) for dioxins and furans, benzene for volatile organic compounds (VOC), and benzo[*a*]pyrene (B[*a*]P) for polycyclic aromatic hydrocarbons (PAHs). Main properties of these substances are given in Section D.2.2.2 (SI), with the highest persistence in air for benzene and the lowest for B[a]P, and the highest bioaccumulation factor for TCDD. We first analyze the spatial distribution of concentrations and intakes for a single test atmospheric source, illustrating via a series of maps how the model performs and determines intake fractions. For each substance and each of the 126 point sources, spatial distributions of the intake and intake fractions through inhalation and total ingestion (aggregated over all ingestion pathways) are then computed and summarized radially.

Key to this study is the creation of a set of multi-scale grids with higher resolution over locations of interest, such as emission sources and regions with large population. The high resolution around the source allows characterizing the local evolution of intake with distance, accounting for specific features in the source vicinity. The mechanism for building the multi-scale grids involves a *refinement potential* (RP): a raster whose pixels quantify the need (potential) for grid refinement. The refinement potential

used for this study results mainly from two components: a raster of inverse distances from source and a raster of population counts. Starting from a background, low resolution grid, the RP is used as a basis for iteratively refining each cell whose integrated potential is above a defined threshold. The outcome of this refinement method is shown with the grid of Figure 4.2. It has a high resolution around points of emission (approximately 7 km × 7 km cells) which decreases with distance from source. The resolution is also constrained by the presence of large populations. The first layer of the atmospheric grid is identical to the results grid; the other 16 atmospheric layers have resolutions decreasing with altitude. The terrestrial grid is made of watersheds obtained from clustering the 0.5° × 0.5° gridded water network defined by WWDRII. The clustering is only weakly constrained by the RP, because the underlying watersheds structure of the hydrological network must be respected.

Radial distribution of iF – The evolution of iF with distance from source for each exposure pathway, is obtained by summarizing radially each iF spatial distribution. Radial summaries are performed on concentric thin rings centered at emission point sources. The integral of each radial summary from sources and the percentage of the global iF that they represent are then expressed as a function of the distance from source.

4.4.3 Map of global iF for large number of emission locations

Maps of global iFs associated with a large number of emission locations in France, are obtained by repeating the simulation for ~10,000 locations, for each of the three substances. Global/total iF through all exposure pathways are computed per exposure route, and associated with each emission location.

4.5 Results

4.5.1 Flexible model parameterization

Using a refinement potential fitting the needs of this study, it takes *Pangea* about 5 minutes on a laptop (Xeon CPU, 32GB RAM) to create a full spatial multimedia model: build grids, connections, and project data. It takes less than 10 seconds per substance to parameterize and run emissions from 126 sources. Exports to data files and maps, can then take from a few minutes to hours depending on the level of output desired, mainly because creating maps with thousands of polygons is time consuming.

4.5.2 Local versus global intakes of pollutants

Spatial distributions of environmental concentrations are computed for the three substances: benzene, 2,3,7,8-TCDD, and B[a]P. To illustrate the approach underlying the 126 and then 10,000 scenarios, Figure 3 shows environmental concentrations [$\mathrm{mg/m^3}$] of the three substances in air (layer 1), fresh water, and agricultural soils, and the population intake fraction surface densities [$\mathrm{1/m^2}$] ([$\mathrm{kg_{taken\ in}/kg_{emitted}/m^2}$]) through inhalation (direct) and total ingestion (production-based), for a unit emission (1 $\mathrm{kg/s}$) at a single emission source located in La Veuve, in the North East of France (white cross in Figure 3).

Figure 4.3.A shows that plumes are stretched in the prevailing wind direction with atmospheric concentrations decreasing with the distance from source. The size of the plume directly reflects the substance specific persistence in air, with long-range transport for benzene and much more localized air concentration for B[a]P which has the shortest half-life in air. The contrast between freshwater concentration (Figure 4.3.B) and concentrations in agricultural soils (Figure 4.3.C) shows that the former is not due to atmospheric transport only, and that transport through the hydrological network is relevant, especially for less volatile substances. The concentrations in freshwater of Figure 3.B reflect the combination of atmospheric transport, deposition, and runoff. The Seine River (medium size with 500 $\mathrm{m^3/s}$ average discharge, flowing through Paris), whose watershed is fully under the plume, drives high concentrations in a direction opposite to the plume. The Rhine River (large with 2000 $\mathrm{m^3/s}$ average discharge, flowing through Germany and Netherlands), whose watershed is only partly under the plume, has more dilution power and leads to concentrations lower than over its surrounding watershed. Due to its high volatility and low air-water partition coefficient, benzene concentrations in water and soil are as expected orders of magnitude inferior to TCDD for the same emission scenario.

Figure 4.3.D/E show population intake fractions as a surface density. While the unit $[1/\mathrm{m}^2]$ ($[\mathrm{kg}_{taken\ in}/\mathrm{kg}_{emitted}/\mathrm{m}^2]$) is more difficult to interpret than a unit-less fraction, it facilitates interpreting maps whose values/colors are independent of cells sizes. Densities can moreover be integrated over any area to compute (or evaluate visually) the intake faction associated with this area.

Figure 4.3.D shows the intake fraction through inhalation; it is direct and hence driven by the plume and the distribution of population. Regions under the plume, even far from the emission source, in Belgium, Netherlands, and Germany, represent a large share of the global intake fraction. Global intake fractions through inhalation for benzene, TCDD, and B[a]P are 1.4×10^{-5}, 4.9×10^{-6}, and 9.5×10^{-7}, respectively, which is consistent with the decreasing persistence of these substances in air.

Figure 4.3.E shows the production-based population intake fraction densities through total ingestion. While the global ingestion iF of a volatile substance like benzene (5.4×10^{-7}) is negligible and two orders of magnitude smaller than the global iF though inhalation, global ingestion iF of TCDD and B[a]F are substantially larger, with 4.9×10^{-3} (main contributions: fish 1.4×10^{-3} and above-ground produces 2.3×10^{-3}) for TCDD and 4.8×10^{-4} (main contribution: above-ground produces 4.2×10^{-4}) for B[a]F. The contribution through the Seine River is more important for TCDD relative to the rest of the map than for B[a]P, which shows the contribution of fish consumption for TCDD and additionally the much lower contribution of fresh water for both. It is interesting to see that the ingestion iF density of TCDD get first reduced as we move away from the source, but increases again over the farther but higher intensive agriculture producing region of the Netherlands.

Spatial distributions of intake fractions can be summarized radially and integrated from the emission source to any given distance from source (radius), to express a cumulative population intake as a function of the distance from source.

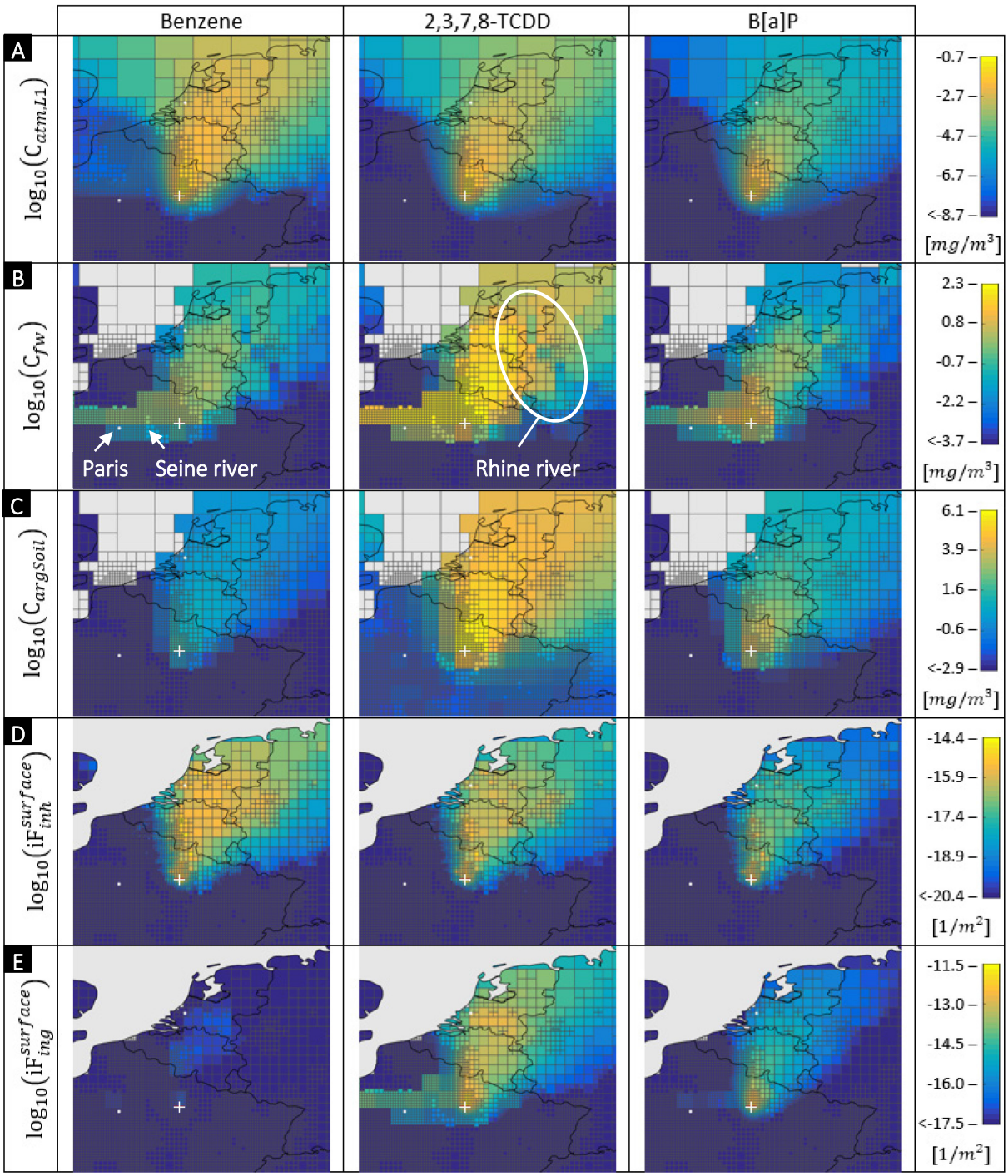


Figure 4.3 – Environmental concentrations $[mg/m^3]$ at steady-state and subsequent population intake fraction surface densities $[1/m^2]$ of selected substances, resulting from a unit (1 kg/s) atmospheric emission at a point source in La Veuve, France (white cross). **A.** Concentration in first layer of atmospheric grid. **B.** Concentration in fresh water. **C.** Concentration in agricultural soils. **D.** Population intake fraction density through inhalation. **E.** Population intake fraction density through total ingestion.

We performed this calculation for the 126 emission point sources shown in Figure 4.2, and built a statistic of all radial summaries. Figure 4.4.A shows the distribution of the evolution of the cumulative population iF by inhalation with the distance from the source, as absolute (A1), and relative (A2, as percent of global iF) values. The top row shows that the 95th percentiles of global inhalation iFs for the three substances are in the range 4.5×10^{-5} to 9×10^{-5}. The spatial variability in absolute inhalation iF between the 5th and the 95th percentiles of all sources amounts to a factor 9 for benzene and up to a factor 64 for B[a]P that show higher spatial variations related to the lower persistence in air and more condensed plume around the source. Analyzing the relative iF plot for benzene, we see that the median curve indicates ~20% of global iF over a 100 km radius from point sources. We further see that the median distance for getting 90% of the global iF is ~600 km from point sources. The corresponding relative values are ~54% at 100 km and ~400 km for 90% for TCDD, and ~90% at 100km for B[a]P, demonstrating well the differences in travel distances between these substances.

Figure 4.4 B shows the distribution of the evolution of the cumulative ingestion iF, in absolute (B1) and relative (B2, as percent of global ingestion iF) values. Figure 4.4 B1 shows that the ingestion iFs span a wider range, from 9×10^{-7} to 6×10^{-3}, than the inhalation iFs. The spatial variability in absolute ingestion iF between the 5th and the 95th percentiles of all sources is less than a factor 6 and therefore lower than the variability of the inhalation iF, reflecting a more uniform spatial distribution of agriculture production, compared to population distribution that is highly concentrated in the large cities. Analyzing the relative iF plot for benzene, we see that the median curve indicates only ~10% of global ingestion iF over a 100 km radius from point sources (Figure 4.4 B2). We also see that the median distance for getting 90% of the global iF is ~600 km from point sources. More importantly, because total ingestion is the dominant exposure pathway for these substances, respectively ~47% and ~85% of the global ingestion iFs for TCDD and B[a]P are generated by food/water production within a 100 km radius from point sources, with respectively ~460 km and ~120 km needed to capture 90% of their global ingestion iFs.

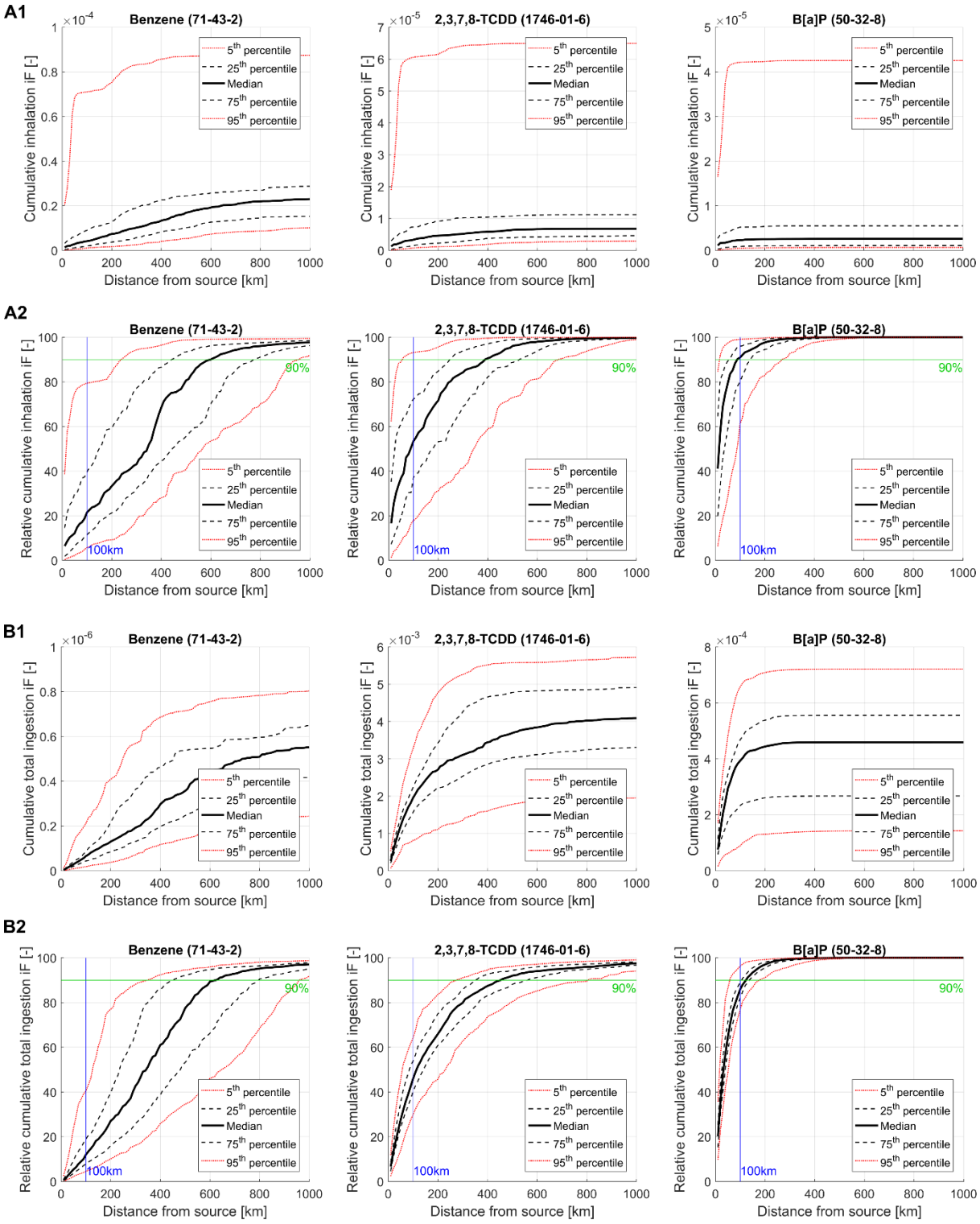


Figure 4.4 – Absolute and relative cumulative population intake fractions for three substances, as a function of distance from source. Inhalation absolute (A1) and relative (A2), production-based ingestion absolute (B1) and relative (B2). The vertical blue line identifies the relative value at 100km from the source (e.g. 20% for benzene/inhalation), the horizontal green line identifies the distance for accounting for 90% of the intake (e.g. 600 km from source for benzene/inhalation).

4.5.3 Maps of global iF for large number of emission locations

Figure 4.5 shows maps of global inhalation iF (4.5 A) and global ingestion iF (4.5 B) associated with ~10,000 emission locations distributed over France. More explicitely, the value of a given pixel is the value of the global cumulated iF associated with an emission source from this pixel, yielding iF maps by source location.

Comparing Figures 4.5 A and 4.5 B, the dominant route is inhalation for benzene, and total ingestion for TCDD and B[a]P. Analyzing the map of global inhalation iF for benzene, we observe high values in regions up-wind from Paris, Lyon, Geneva, and more generally north of France upwind from Belgium, Germany, and the Netherlands. Maps of global ingestion iF for both TCDD and B[a]P match well the distribution of large crop (maize, wheat, etc) and animal-based agricultural regions in France (internet search “carte production agricole France”, that we cannot reproduce here).

Figure 4.5 C shows distributions of inhalation and ingestion iF, as well as the continental rural and urban iF archetypes obtained from USEtox (European region 19). There is a good match between distributions of global inhalation iFs for the three substances and the boundaries defined by USEtox continental and urban regions, considering that the 10,000 locations modeled with Pangea include cases of extreme urban exposure (emission precisely in the middle of large population). Pangea yields respectively one order of magnitude lower total ingestion of benzene and higher total ingestion of TCDD and B[a]P compared to the generic model USEtox.

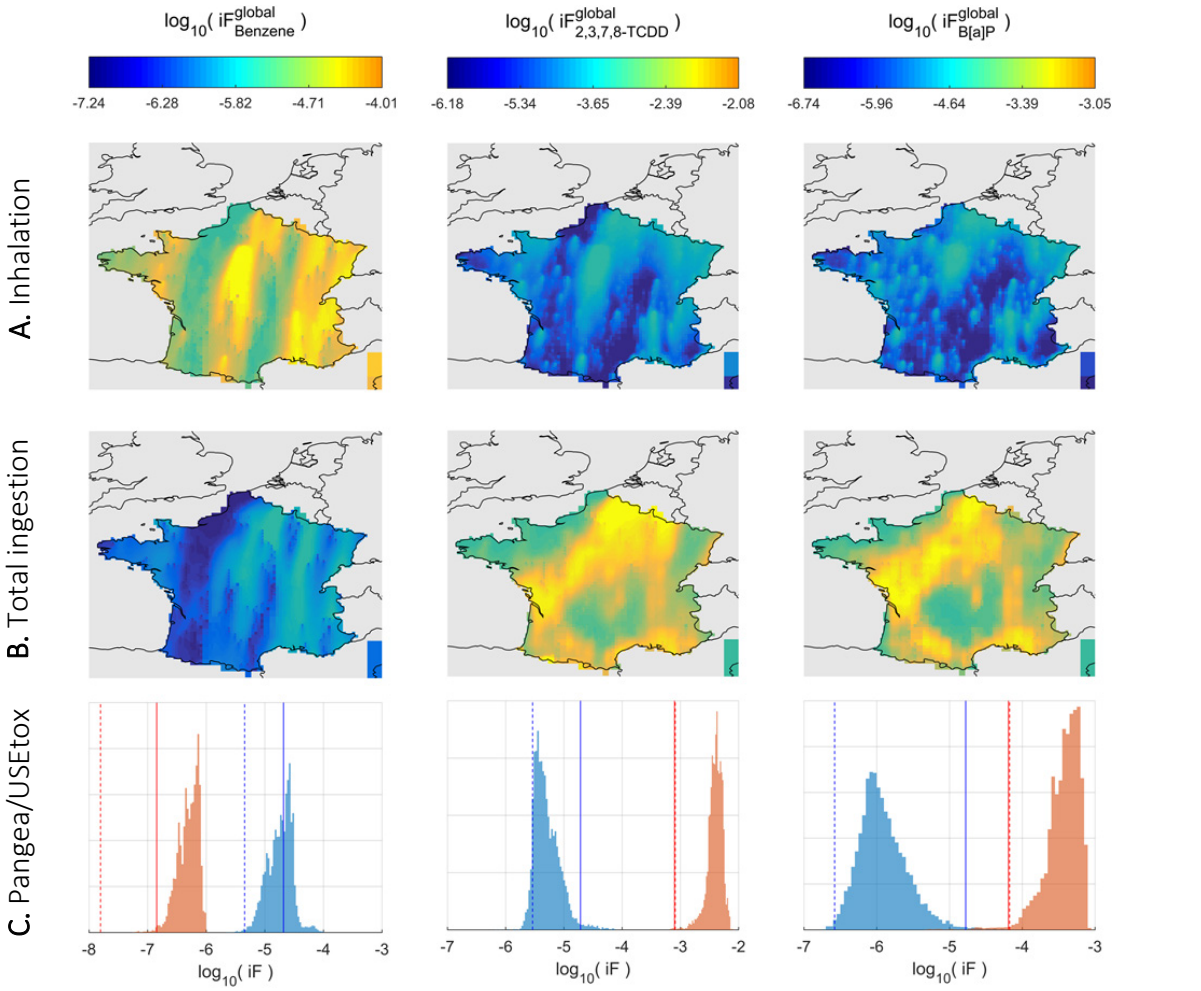


Figure 4.5 – Maps of global iF through inhalation and total ingestion associated with ~10,000 emission locations covering France.

4.6 Sensitivity study

We performed a re-analysis of the studies presented in this chapter, using meteorological fields from GEOS-FP. The objective was to evaluate how sensitive the radial studies were to the GEOS-Chem grid resolution, and to the year. GEOS-FP has a native resolution of 0.25°×0.3125° globally on 72 vertical layers, but in practice it can be used on a 2°×2.5° global grid, with higher resolution 0.25°×0.3125° nested grids that cover specific continents. Figure 4.7 shows both grids over France: the global 2°×2.5° grid with red nodes, and the European 0.25°×0.3125° grid with green nodes. It also shows a circle with a 100 km radius, centered on a solid waste treatment plant in the North East of France. This represents an average domain for local/regional specialized atmospheric dispersion models. Figure 4.6 shows that the 2°×2.5° grid cells are lower resolution than the 100 km boundary, but that the higher resolution grid has ~30 nodes within the 100 km radius around the source. Figure 4.6 shows the V+ component as defined on both grids, which informs about what each resolution captures at the scale of France.

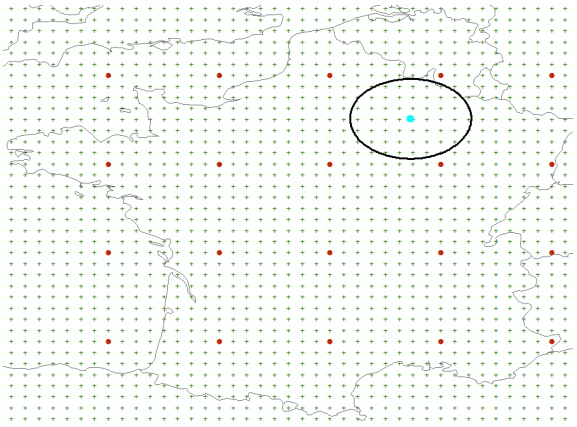


Figure 4.6 – GEOS-FP global 2°×2.5° grid (red), and nested 0.25°×0.3125° grid for Europe (green). The blue dot is the location of a solid waste treatment plant in the North East of France. The circle with a 100 km radius allows a better comparison of the two grid resolutions.

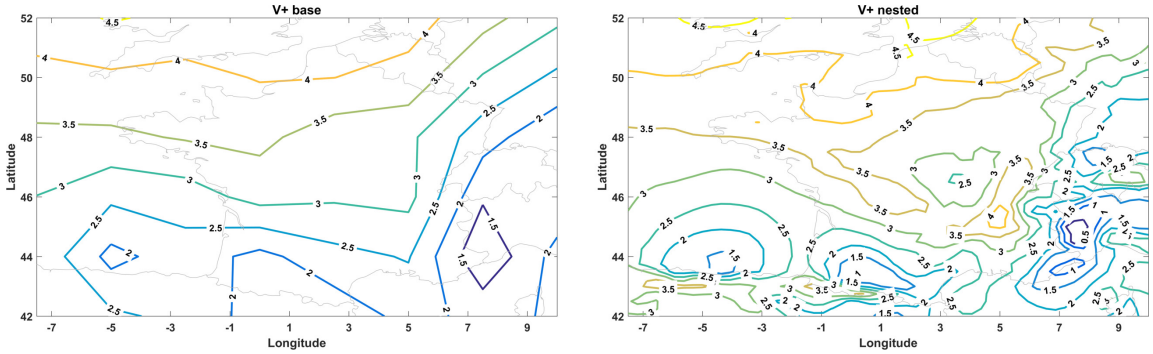


Figure 4.7 – V+ component of the wind field defined by GEOS-FP global 2°×2.5° on the left, and the nested 0.25°×0.3125° European version on the right.

Finally, Figure 4.8 shows the evolution with the version of the meteorological field of the intake fractions at 100 km and the global intake fraction, for the inhalation and the total ingestion routes, and for the three studied substances. A series of versions of meteorological fields were tested, alternating

between 2°×2.5° only grids and nested 0.25°×0.3125° grids, for years 2013 to 2016. Year 2005 is based on GEOS-4 and is only available on the 2°×2.5° grid.

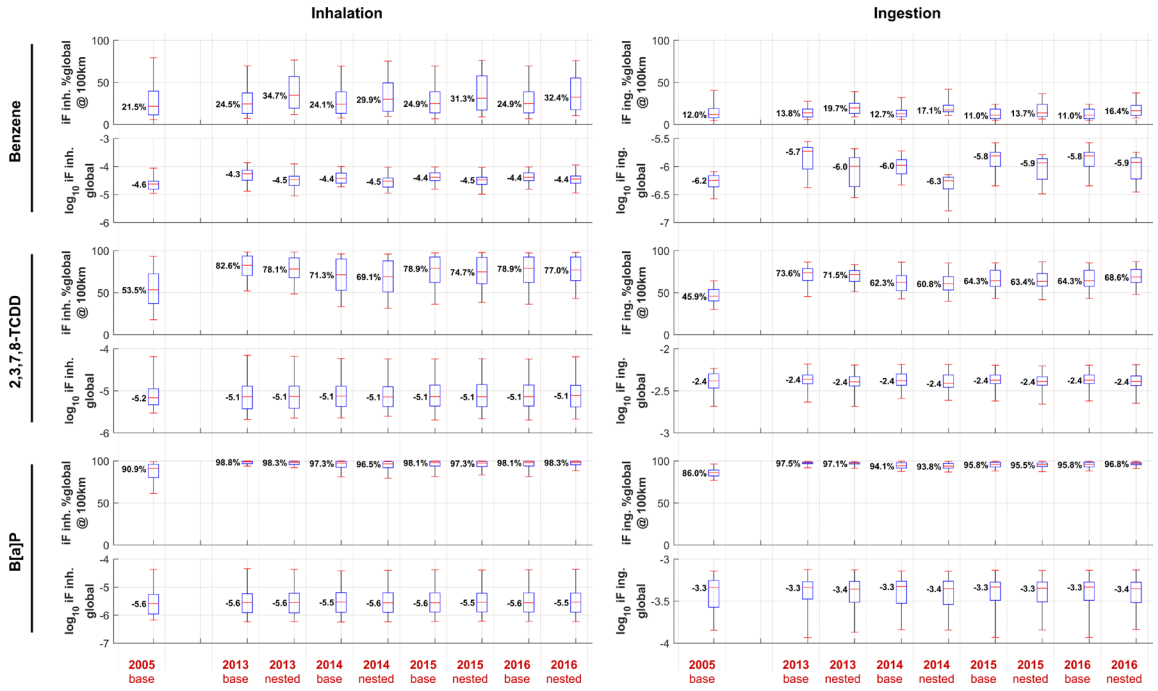


Figure 4.8 – Evolution of iF global through inhalation and total ingestion, and the percentage of global iF at 100km, with the version and year of GEOS-Chem. Boxplots represent a statistic over the 126 SWTP point sources studied in the article. Year 2005 was based on GEOS-4 and later years on GEOS-FP. The base resolution is a global 2°×2.5° grid. The nested resolution refers to a 0.25°×0.3125° grid over Europe, nested in a global base grid.

Absolute values of global intake fractions are stable across the years and resolutions of the meteorological field. Percentage of global intake fractions at 100 km shows a slightly lower fraction in 2005, especially for 2,3,7,8-TCDD, but there is no substantial variation with the resolution.

4.7 Discussion

We compared local with global population inhalation and production-based intake fractions for a set of 126 point sources located at the position of 126 solid waste treatment plants across France. Median curves of evolution of the cumulative iF with the distance from source show that for selected volatile and persistent pollutants, modeling local pollutant fate and subsequent population exposure within a 100 km radius from the source may only account for 20% of the global population intake through inhalation, while capturing about 85% of the global intake through total ingestion for the less persistent B[a]P. Two competing factors define the evolution of the population intake radially from the source, namely decreasing environmental concentrations, and increasing exposed population count. Individual intakes decrease in the same fashion as environmental concentrations; therefore, locations distant from the source are characterized by a large cumulative exposed population intake, but with small individual intakes. The relevance of this situation relates to the dose-response relationship specific to each pollutant (linear/non-linear) and endpoint (e.g. cancer), i.e. the probability of developing a specific adverse effect expressed as a function of the intake dose for a specific pollutant. Of special interest are carcinogenic pollutants, in particular when considered under the EPA default assumption of a linear mutagenic mechanism, and situations where doses due to background concentrations are already above thresholds. In such cases, a large population exposed to low doses increments may lead to as many cases of e.g. cancer as a smaller population exposed to higher doses. This suggests that, by discarding distant population exposure to low doses increments, local and regional assessments might only account for a small fraction of global impacts, in particular for mutagenic substances.

More generally, intake fractions may spatially vary by several orders of magnitude, especially for short-lived substances. In such a context, *Pangea* offers the possibility to draw high resolution iF maps for tenths of thousands of locations, accounting for source to population typical distances and to provide iF calculations of multiple exposure scenarios without having to reparametrize the model for each scenario.

Several limitations apply to *Pangea*. All input datasets and grids have different resolutions; datasets with lower resolution than the finest multi-scale grid cells are interpolated, and datasets with higher resolution are summarized by zonal statistics. It is therefore not possible to define a single resolution for *Pangea*. Simulation durations compatible with pollutant screening can only be achieved with *Pangea* setup for computing a steady-state solution of a linear model with constant coefficients. Despite the technical ability of the model to build study-specific grids with arbitrarily high resolution, an absolute limit of resolution is defined by the scale at which dynamic and non-linear effects become significant.

In conclusion, *Pangea* is an innovative global spatial multimedia fate and exposure model able to create study-specific multi-scale grids and project data at runtime. The mechanism for grids generation based on a user-defined refinement potential allows focusing on multiple locations and sources of various natures, reducing averaging artifacts and allowing for a better characterization of emitters and receptors at relevant locations, which can be relevant in risk and impact assessment related decision contexts.

Chapter 5 Exposure to pollutants in Australia, from sources to receptors

5.1 Introduction

The previous chapters have studied source(s) to intake(s) from an emitter perspective, looking in particular at the spatial distribution of intake fractions (iFs), the fraction of a unit emission that is taken in by the overall population. Studies were considering multiple single source locations, multiple sources (or diffuse sources), and multiple substances. To complement the spatial analysis of the source to intake/receptor relationship, there is a need to study the case of multiple sources of varying emission intensities, and to look at these sources both from an emitter and from a receptor respective. We are especially interested in identifying, for a given receptor location, the main sources contributing to the atmospheric concentration at this location, and further to the exposure for this receptor. In other words, we aim to perform a source apportionment.

Australia represents an interesting case study for multiple reasons: it offers high contrasts in the population density between the populated coastal cities, in particular on the East Coast, and large desert inland zones. As a continental island, concentrations in Australia are likely to be primarily due to "local" emissions. At the same time, Australia is in relative proximity to much more densely populated areas from Indonesia, enabling to further study the role of long-range transport of relatively persistent compounds in the atmosphere. In addition, Australia maintains a pollutant emission inventory of more than 80 substances (for the period 2014-2015) and 4,101 sources (reports) spread in the entire country. The inventory associates the reports with 166 industrial sectors, which enables us to compare the respective contributions of specific sectors.

This chapter aims thus to study the source to receptor relationship of industrial sources of organic compounds in Australia. More specifically, it aims to: **1.** Identify the radial spatial distribution of population intakes for contrasted source locations in urban, rural, desert, and sea areas. **2.** Determine the population exposure resulting from the combined emissions of 4,101 point sources spread across the entire Australia, identifying the main contributing sectors and the magnitude of their impact on human

health. **3.** Study a receptor perspective for a contrasted set of receptors, and perform a source apportionment, identifying the main sources contributing to exposure at a given location.

5.2 Methods

Inventory of emissions: Australian Pollutant Inventory – We use the Australian National Pollutant Inventor (NPI) to define spatial emissions. *Pangea* implements a parser/wrapper for NPI, which processes NPI original XML files and builds databases that allow to access emissions per year, per report, per substance, and per sector.

Substances and parameters – On the 83 substances defined by NPI for the period 2014-2015, we keep the 43 substances present in the USEtox substances database. Physico-chemical parameters are hence fully based on USEtox. Bioaccumulation factors, human toxicity effect factors, and severity factors for computing DALYs are also extracted from USEtox. The wind field is defined by GEOS-Chem (Bey et al., 2001), more specifically GEOS-FP for the year 2014. All other data sets and parameters are *Pangea* defaults.

Modeling pollutants fate and populations exposure – The *Pangea* framework is used to model the spatial fate and transport of substances, and the subsequent populations exposure. USEtox is run by *Pangea* to perform a systematic model-model comparison.

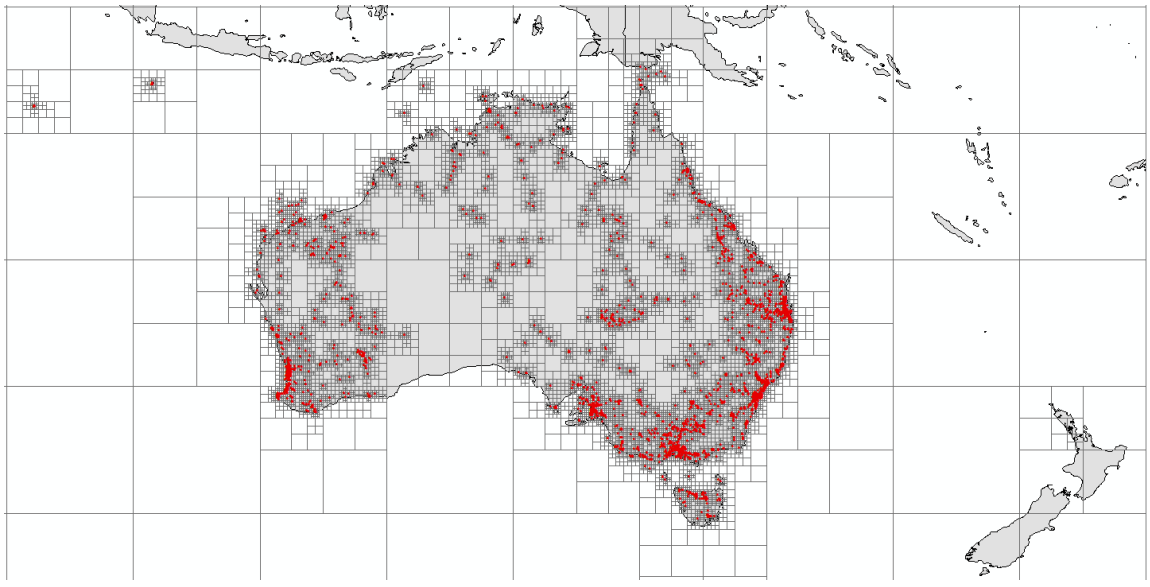


Figure 5.1 – First layer of the atmospheric grid and locations of 4,101 sources of emission defined by NPI.

Grids – Grid refinement is based on the coordinates of all emission sources for the year 2014 (refinement decreases with the radius from each source), the population distribution, and two flag that target lands from a specific region of interest (including Australia, New Zealand, Tasmania, Christmas Island, and an offshore platform North of Australia). The outcome of the refinement procedure is shown in Figure 5.1. Red dots mark the location of the 4,101 emission point sources. This grid defines the first layer of the atmospheric grid (17 layers with decreasing resolution). The terrestrial grid is defined by the Word Water Development Report II (WWDRII) native 0.5°×0.5° grid over the region of interest, and by larger clusters elsewhere.

Receptor perspective and source apportionment – For simplifying the source apportionment, we computed the block of **FF**, the matrix of fate factors (section 2.11.6), that corresponds to transfers from the first layer of the atmospheric grid to itself (Figure 5.2). This block defines the contribution of each cell of this layer to the mass at steady state in each other cell.

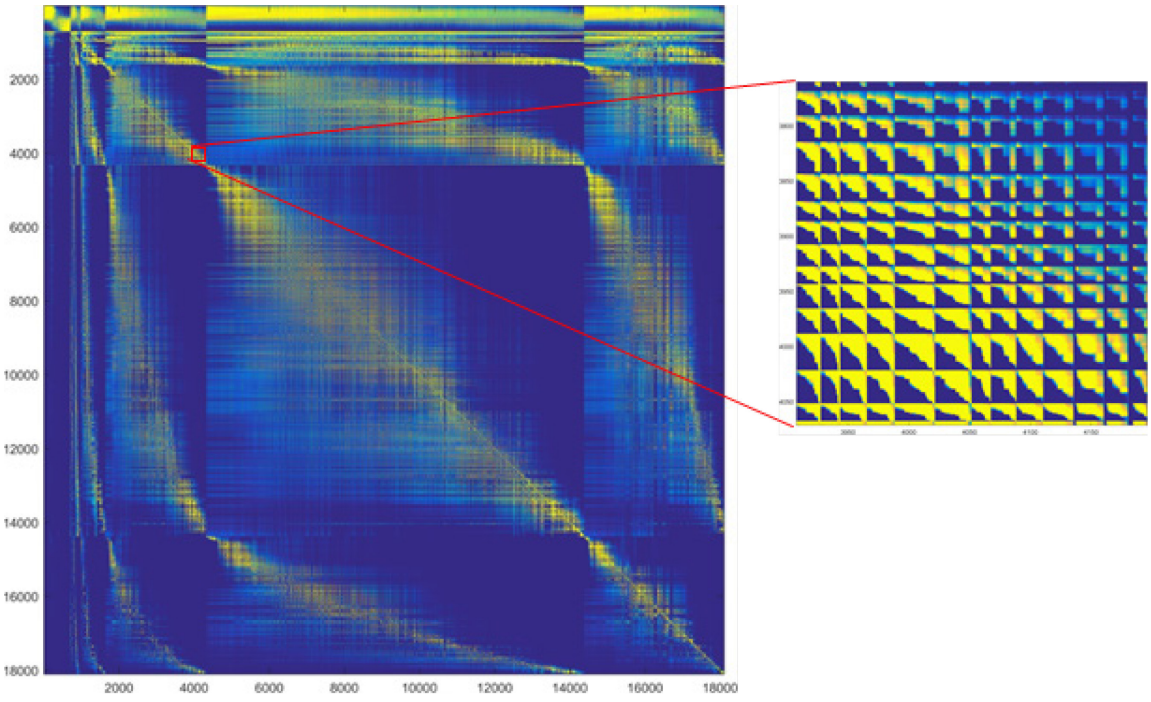


Figure 5.2 – **Left**: 18107×18107 dense block of a **FF** matrix (for benzene and the geometry of the Australian project), associated with transfers from the atmospheric layer #1 to itself. **Right**: zoom on a small part, detail of fine pattern. The full 109766 × 109766 **FF** matrix cannot be computed fully/directly.

Case study – Based on the spatialized inventories of emissions, we designed the case study in four steps: **1.** We simulate unit emissions at a set of five locations that represent archetypical situations (urban, rural, offshore, etc.), to illustrate our approach from an emitter perspective. We build maps of concentrations, iFs through inhalation, and cumulative radial statistic of the iFs. The radial statistic

shows at what distances each emission source reaches a highly populated exposed region. **2.** We simulate all 43 substances and we show maps of total concentrations and iFs for benzene and formaldehyde. **3.** Using USEtox effect factors and factors for converting cases of adverse effects to DALYs, we compute total DALYs and rank substances accordingly. **4.** We select four top ranking substances, namely formaldehyde, styrene, dichloromethane, and benzene, and we analyze them per sector. **5.** We perform a source apportionment at 5 relevant locations, to analyze a receptor perspective. **6.** Finally, we perform a source apportionment for a specific location, and we trace back the most relevant contributor using meta-information from NPI.

5.3 Results and discussion

5.3.1 Emitter perspective, intake fractions and radial analysis of individual sources

We first compute the spatial distributions of population iFs for contrasted source locations – urban (Sydney), rural (Orange depot, 200 km North-West of Sydney), desert (Alice Springs), and sea areas (Oil platform North-West Australia) – and their cumulative radial statistics. Figure 5.3 and Figure 5.4 show maps of atmospheric concentrations, inhalation iFs per square meter, and plots of radial statistics, that corresponds to emissions of benzene and formaldehyde in these five locations.

For emission at Sydney airport, the plume is primarily to the East over the ocean (A1) and the quasi-entire intake takes place within 50 km of the Sydney agglomeration (A2), with high urban inhalation iFs (A3) of 33 ppm for benzene and 9.5 ppm for formaldehyde (OH atmospheric degradation half-life of 1.1 day), which is a factor 7 less persistent than benzene (OH degradation half-life of 8.7 days) and therefore has a substantially shorter atmospheric travel distance. For such urban emission, the formaldehyde intake by ingestion is small (0.06 ppm) and the ingestion of volatile benzene is anyway negligible in all cases.

When emitting 150 km West of Sydney at the Orange Depot, located on the other side of the blue mountains, we observe a plume to the South-East towards Sydney, and another plume towards the desert area of the North-West of Orange Depot (B1). The local iF is limited to 1 ppm for benzene, and most of the intake takes place when benzene reaches large populations in Sydney (~200 km of Orange Depot, B2), with a cumulative iF of 7.3 ppm for benzene (B3). The size of the formaldehyde plume is limited around the emission source, and does not reach Sydney, so that all the intake takes place locally, with a low iF of 0.25 ppm (Figure 5.4, B1-B3). Interestingly, for formaldehyde emissions in this rural area, the formaldehyde intake by ingestion is higher than by inhalation, with an iF by ingestion of approximately 4 ppm.

For emissions in the primarily desert area around Alice Spring airport, dominant winds direct the plume to the North-West (C1), leading to a very low local iF of 0.25 ppm for benzene (C2-C3).

Figure 5.5 shows a map of locations and concentric circles around each location, at distances that correspond to the steps in the radial statistics for benzene (Figure 5.3, column C). The first step in the radial statistic for emissions from Alice Springs happens around 2,100 km from Alice Springs (Figure 5.3, C3), which corresponds to the inner violet circle (2,100 km) that passes through most large cities on the East coast (Figure 5.5). Further steps in the range 3,000 km to 3,300 km correspond to the next two circles, that cross highly populated regions of Indonesia, yielding a limited cumulative iF by inhalation of 0.7 ppm for benzene and only 0.17 ppm for formaldehyde, with only local intakes in Alice Springs itself since formaldehyde is removed from the atmosphere before reaching other inhabited areas (Figure 5.4, C3).

One could expect a very low iF for emissions at the Montara Field oil platform, which is in the middle of the Timor Sea, with no population within 300 km and a very limited population within 1400 km (Figure 5.5). This is indeed the case for the short-lived formaldehyde in air, with a negligible intake fraction of 0.0005 ppm (Figure 5.4, D3). However, this is not the case for benzene with a relatively high cumulative iF of 15 ppm when the plume reaches the highly-populated Indonesia, between 1,400 km and 4,000 km from the source.

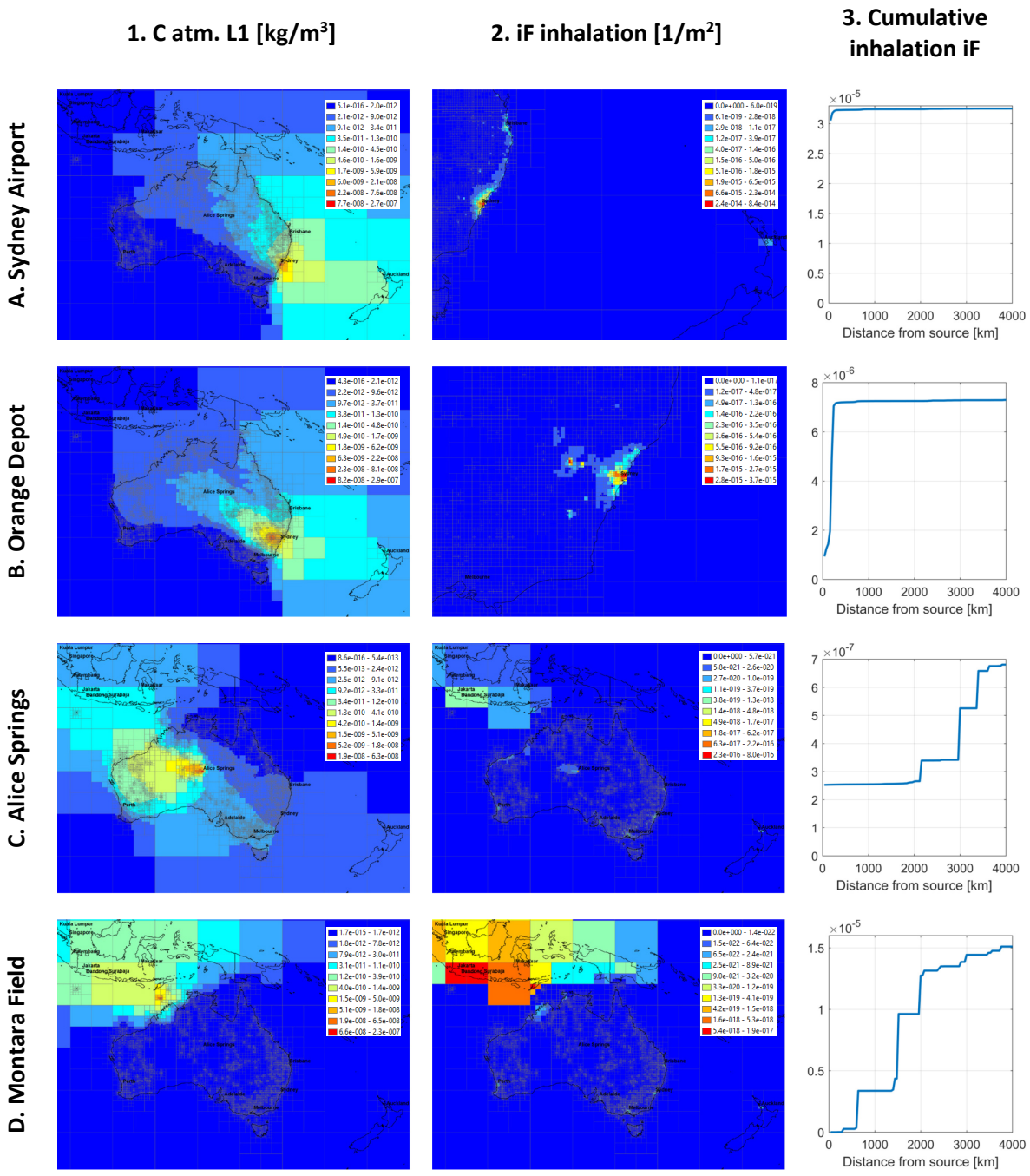


Figure 5.3 – **1.** Atmospheric concentration in layer #1, **2.** Inhalation iF per square meter, and **3.** Cumulative radial statistics of inhalation iF, for a unit emission flow of 1 kg/s of benzene in **A.** Sydney airport (urban), **B.** Orange depot, 200 km North-West of Sydney (rural), **C.** Alice Spring, desert, and **D.** the Montara Field oil platform (remote, sea).

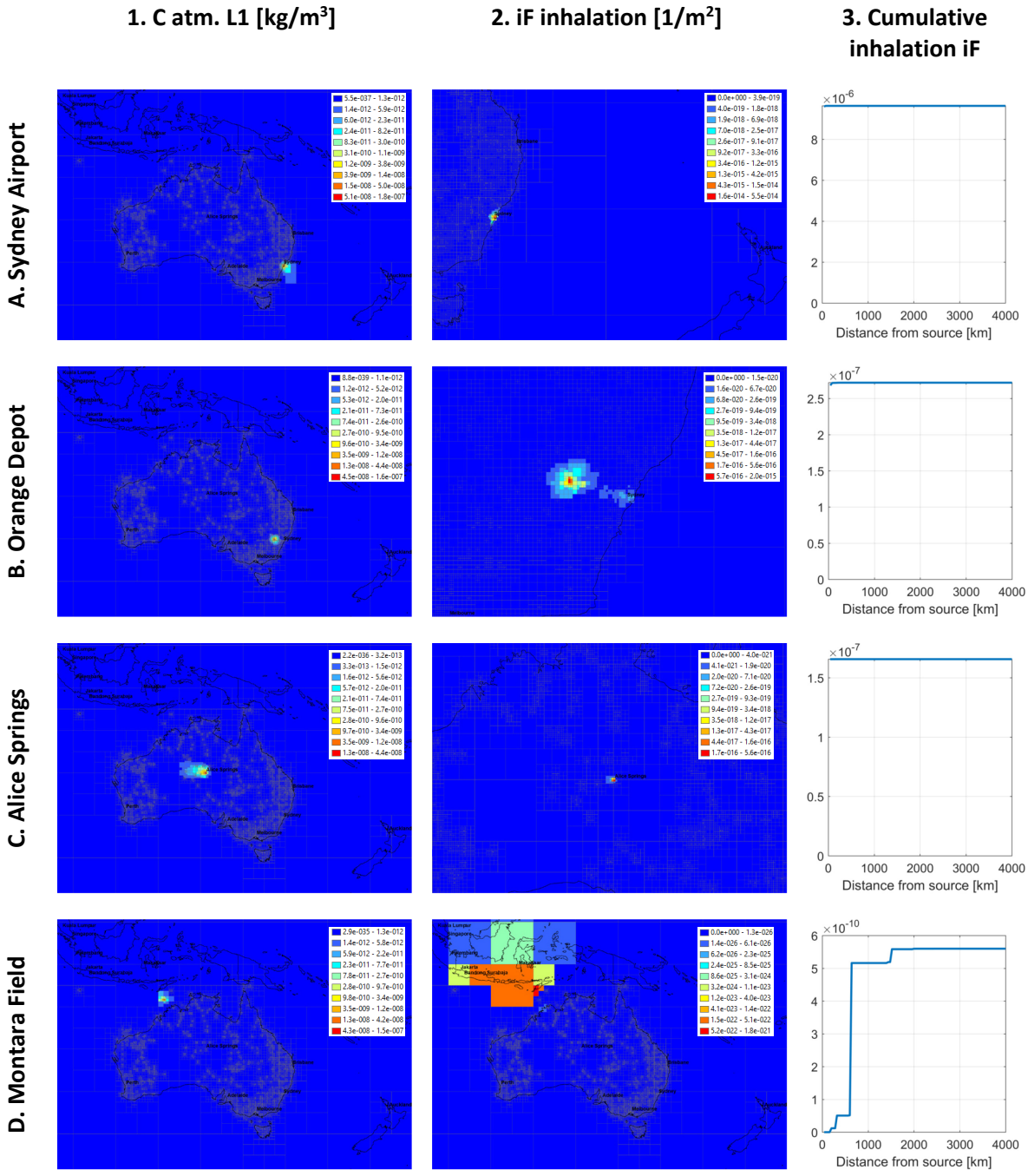


Figure 5.4 – **1.** Atmospheric concentration in layer #1, **2.** Inhalation iF per square meter, and **3.** Cumulative radial statistics of inhalation iF, for a unit emission flow of 1 kg/s of formaldehyde in **A.** Sydney airport (urban), **B.** Orange depot, 200 km North-West of Sydney (rural), **C.** Alice Spring, desert, and **D.** the Montara Field oil platform (remote, sea).

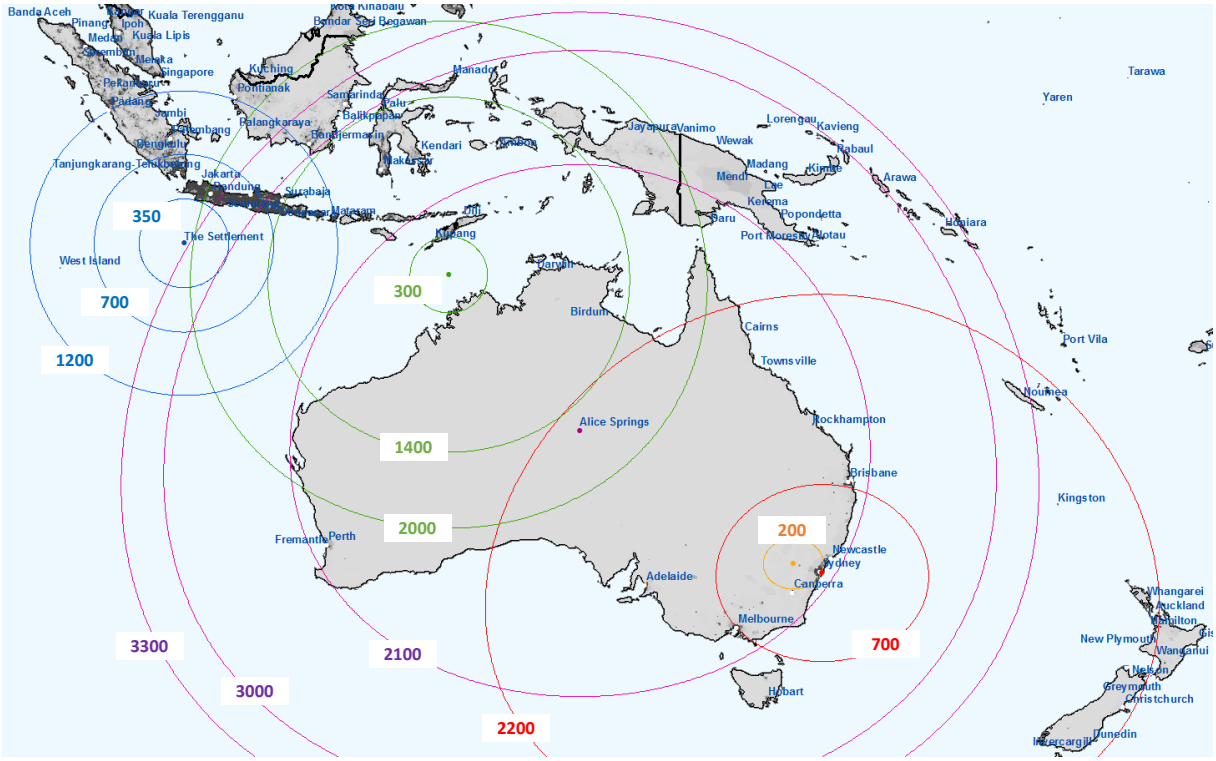


Figure 5.5 – Circles around emission source, approximately at the distances of the steps in the cumulative radial statistics of inhalation iF of Figure 5.3. Sydney airport, red circles at {700, 2,000} km. Alice Springs, violet circles at {2,100, 3,000, 3,000} km. Montara Field, green circles at {300, 1,400, 2,000} km. Christmas Island, blue circles at {350, 700, 1,200} km. Orange Depot: orange circle at 200 km.

5.3.2 Emitter perspective, comparison of global intakes with USEtox

We generalize the study of iFs through inhalation and ingestion to all substances and all emission points, and we compare distributions of global iFs computed by *Pangea*, with pairs of iFs computed with USEtox and associated with the urban and continental (rural) archetypes. To illustrate this comparison with benzene, the 755 emissions (reports) of benzene defined by NPI are located in 552 atmospheric cells; we compute global iFs associated with these cells, and we get a distribution. We then extract urban and continental iFs for Australia (zone 10) from USEtox, and we compare. Figure 5.6 shows results for formaldehyde, benzene, and styrene. We observe a good match for inhalation iFs (blue), where USEtox values for urban emissions (plain blue line) and for continental emissions (dashed blue line) are well within *Pangea*'s distributions (median value marked by dotted blue line with value). Intake fractions through total ingestion are generally an order of magnitude (max) higher in *Pangea* than in USEtox[43].

[43] This will be further investigated after the next refactoring of *Pangea* that aims to build a set of EPMs for *Pangea* that matches strictly USEtox EPMs.

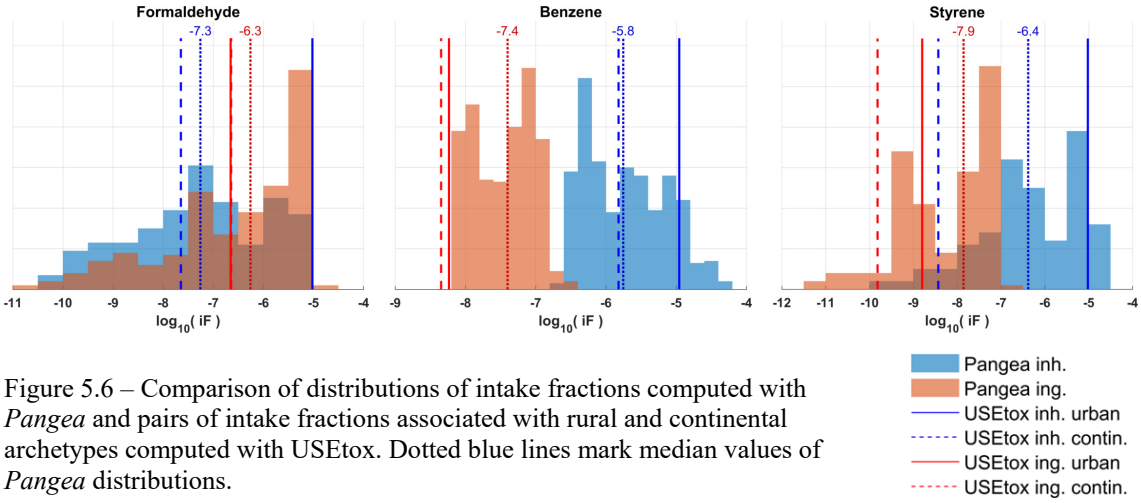


Figure 5.6 – Comparison of distributions of intake fractions computed with *Pangea* and pairs of intake fractions associated with rural and continental archetypes computed with USEtox. Dotted blue lines mark median values of *Pangea* distributions.

Figure 5.7 shows a comparison for all substances. Points are defined by USEtox mean value (x axis) and *Pangea* mean value (y), bars are defined by urban/continental values for USEtox, and one standard deviation for Pangea. We observe a relatively good match between *Pangea* and USEtox for inhalation iFs (Pearson r = 0.5, RMSE = 0.61) with generally less than an order of magnitude difference between the two models, and a good match for ingestion iFs, where *Pangea* provides consistently larger values than USEtox (Pearson r = 0.94, RMSE = 0.89). The inhalation iF of the USEtox urban archetype is constant ($\log_{10}$iF = -5, iF = 10 ppm) for most substances, unless persistence in air is high, whereas *Pangea* is able to account for the effective population density around sources.

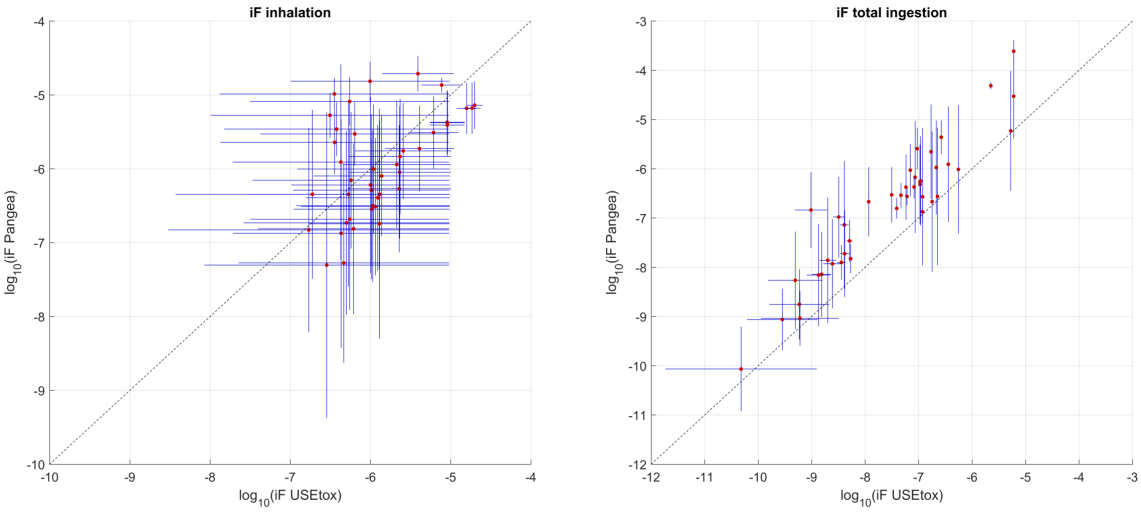


Figure 5.7 – Comparison of inhalation and ingestion intake fractions from USEtox (x axis, mean, continental, urban) and *Pangea* (y axis, one standard deviation), for the 43 substances.

5.3.3 Overall population exposure to the 4101 sources of the NPI inventory

Overall intake fractions and intakes – We determine the population exposure resulting from the combined emissions of 4,101 point sources spread in all of Australia, per substance and per industrial sector (each sector is an emission scenario). For each substance, we compute the steady-state solution of the fate and transport, and subsequent population exposure. This yields distributions of environmental concentrations and population iFs by inhalation and by ingestion, per scenario/sector. We finally aggregate over sectors and build maps of total concentrations and iFs.

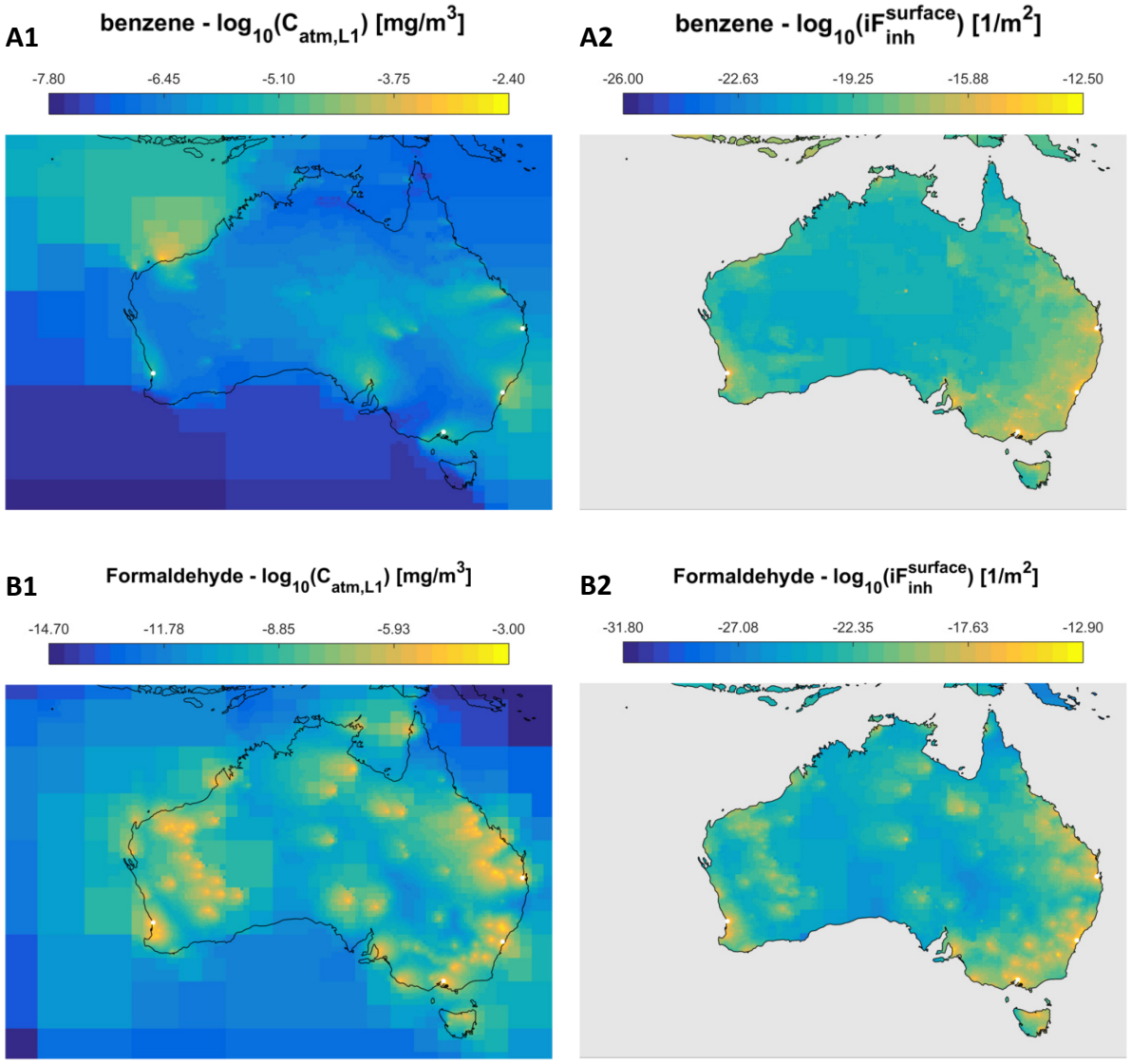


Figure 5.8 – Atmospheric concentrations and inhalation iFs per square meter, of benzene (A1, A2) and formaldehyde (B1, B2), for the annual average emissions flows of the 4101 source of the 2014-2015 NPI inventory.

Figure 5.8 presents the resulting atmospheric concentrations and inhalation iFs per square meter of benzene and formaldehyde, showing the more diffuse and extended exposure to benzene that travels over longer distances, versus the more local exposure for formaldehyde. The total emission of benzene defined by the NPI (0.033 kg/s) is a factor three lower than the total emission of formaldehyde (0.094 kg/s). The corresponding inhalation intakes are, however, higher for benzene (1.87×10^{-7} kg/s) than for formaldehyde (9.0×10^{-8} kg/s), reflecting the higher persistence and average iF of 5.7 ppm for benzene, against 0.96 ppm for formaldehyde. The average intake by ingestion of benzene is, as expected, negligible compared to inhalation ($3{,}0\times10^{-9}$ kg/s), but the intake through ingestion is substantial for formaldehyde (1.5×10^{-7} kg/s), which corresponds to an ingestion iF of 1.6 ppm. This behavior reflects the classification of benzene as a volatile compound and formaldehyde as a multi-pathway compound, based on their respective air-water and octanol-water partition coefficients (Bennett et al., 2002a).

Decomposition per industrial sector – Figure 5.9 analyzes the sector-specific contributions to the total ingestion and inhalation intakes, selecting the four neutral organic compounds with the highest potential health impact from Figure 5.10. For benzene, the highest contributing sectors are, as could be expected, mostly from the petroleum and the iron and steel industry, whereas the electricity sector leads to the highest exposure for formaldehyde, just higher than the also important contribution of petroleum refining. For styrene and dichloromethane, most contributing sectors are from the chemical and pharmaceutical industry. In term of exposure pathway for the ingestion route, none of these substances tends to bio-concentrate, so that the direct water ingestion is dominant for all substances but styrene, for which the ingestion via above ground agricultural products is dominant but nevertheless low compared to inhalation.

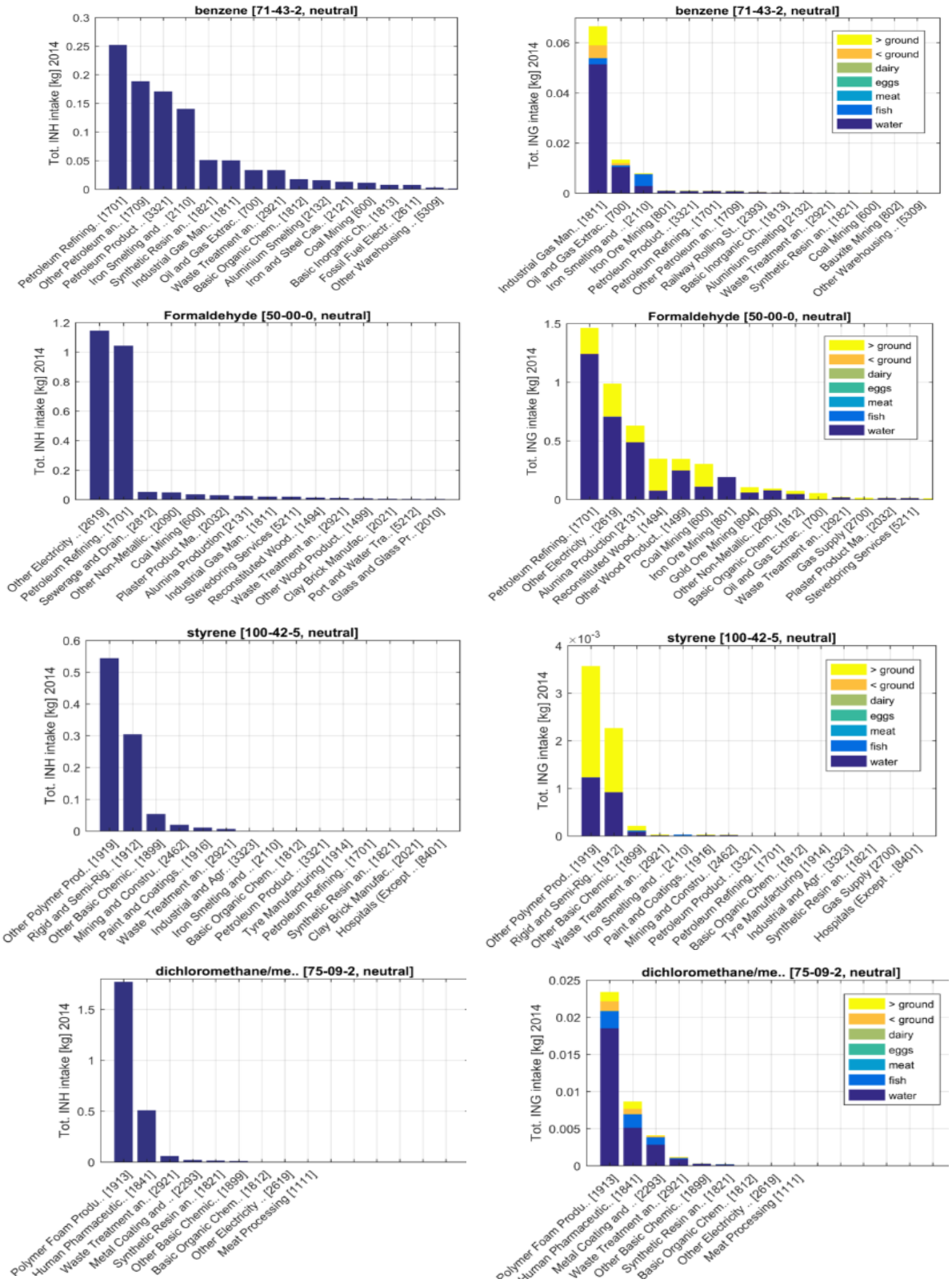


Figure 5.9 – Total inhalation and ingestion intakes per NPI Australian sector, for benzene, formaldehyde, styrene and dichloromethane.

Exposure to severity – Finally, using USEtox human effect and dose-response factors (in DALYs/kg_{in}), we compute cancer and non-cancer effects, and then cancer and non-cancer Disability Adjusted Life years (DALYs). Figure 5.10 shows all 43 simulated substances, ranked based on total DALYs. It suggests that the nation-wide human health footprint of the considered point source emissions of organic compounds is restricted, with a maximum of 30 DALYs per year for formaldehyde. Even considering the high uncertainty associated with the USEtox effect factors – typically a geometric squared deviation of a factor 60 on the dose response of case of non-cancer per kg_{in}, and a factor 13 for the severity factor of DALY per case of non-cancer according to Fantke et al., (2012) – this level of health impact associated with the considered NPI source emission of organic compounds is much lower than the 38,000 DALY/year estimated in the Global Burden of Disease for fine particulate matter impacts in Australia[44]. It is however important to emphasize that the NPI emission inventory only considers emissions related to industry, and does not account for most of e.g. traffic related emissions[45].

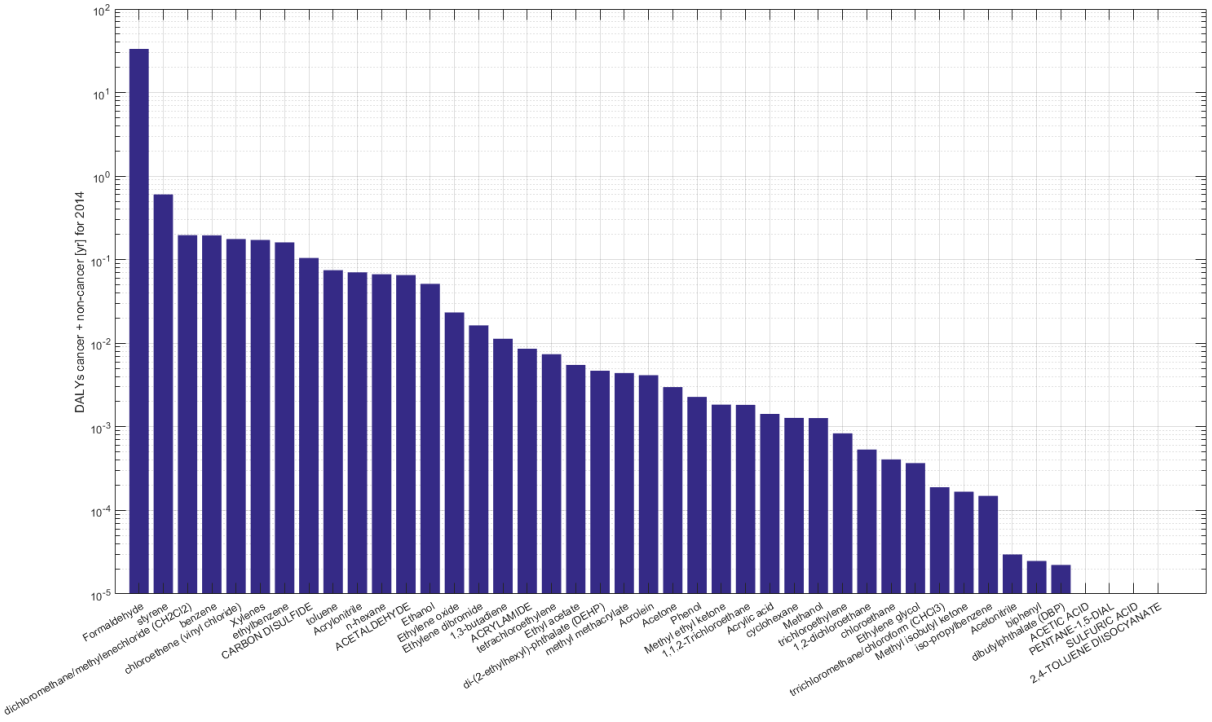


Figure 5.10 – Estimates of total cancer and non-cancer DALYs associated with the overall emissions of 43 organic compounds of the NPI inventory, ranked from highest to lowest total DALY.

[44] https://vizhub.healthdata.org/gbd-compare/

[45] It provides estimates of emissions related to motor vehicles for a limited set of air sheds that were not included in this study.

5.3.4 Receptor analysis - atmospheric source apportionment

In this section, we switch to a receptor perspective, and we aim to identify the main sources that contribute to the exposure at specific locations. We selected contrasted receptor locations in Australia for urban (Sydney opera), rural (Orange depot, 200 km North-West of Sydney), desert (Uluru Rock), island (George Town, Tasmania) conditions, as well as a receptor in Indonesia, to determine the intake in this country due to Australian emissions (emissions in Indonesia are not considered).

Figure 5.11 and Figure 5.12 show maps of fate factors (column 1) and radial statistics (column 2) for the five selected receptor locations. A map of fate factors represents the contribution of (potential) unit emissions (1 kg/s) in all cells, to the concentration at a given receptor location. Cumulative radial statistics integrate contributions from all cells and actual sources define by the NPI database. These curves converge towards the concentration at each specific location, obtained when simulating the whole inventory.

Figure 5.13 shows locations of emission sources that contribute the most to the concentration at the Sydney opera. Most of the contributing sources are in the Sydney agglomeration itself, especially for formaldehyde, with very limited contributions from sources outside of Sydney. For benzene, however, we observe substantial contributions from more distant sources.

It is only in the case of Uluru rock, with no important close sources, that contributions from distant sources represent an important share. Contrasting the radial statistics of intakes from a receptor perspective of Figure 5.11 and Figure 5.12 with the emitter perspective of Figure 5.3 and Figure 5.4 is insightful. From a receptor perspective, the contribution of local sources substantially contributes to the total concentration even for a persistent substance such as benzene, unless we are in a really desert area (Uluru rock) without local source or if we disregard the local emissions (Indonesia).

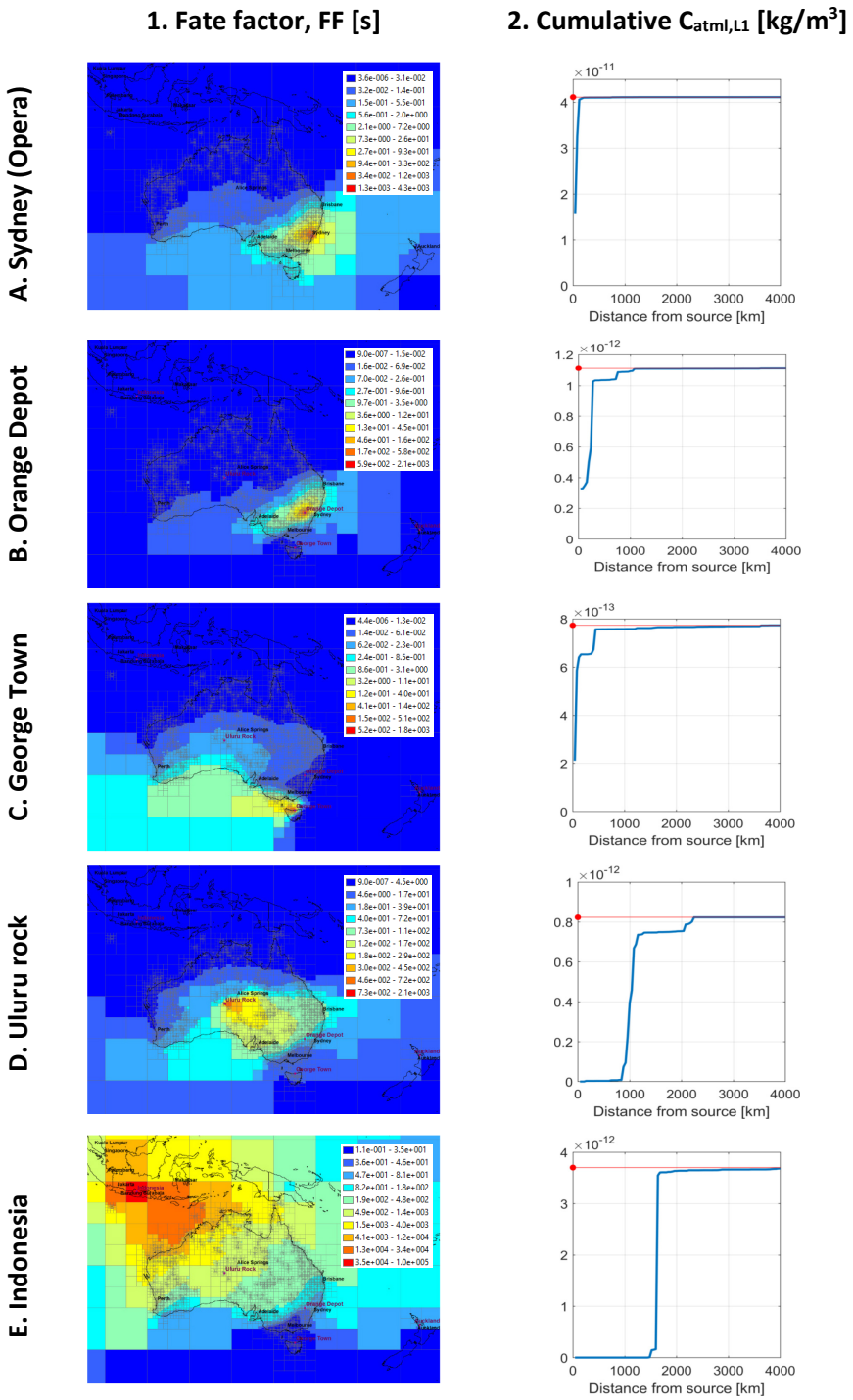


Figure 5.11 – **1.** Maps of fate factors (increase in receptor concentration due to emission of 1 kg/s in the considered cell) and **2.** radial statistic of contributing Australian sources for benzene, for receptors in **A.** Sydney opera, **B.** Orange depot, 200 km North-West of Sydney, **C.** George Town Tasmania, **D.** Uluru Rock, and **E.** Indonesia. The sum of remote contributions to the local concentration, computed with source apportionment, converges towards the concentration computed based on total emissions.

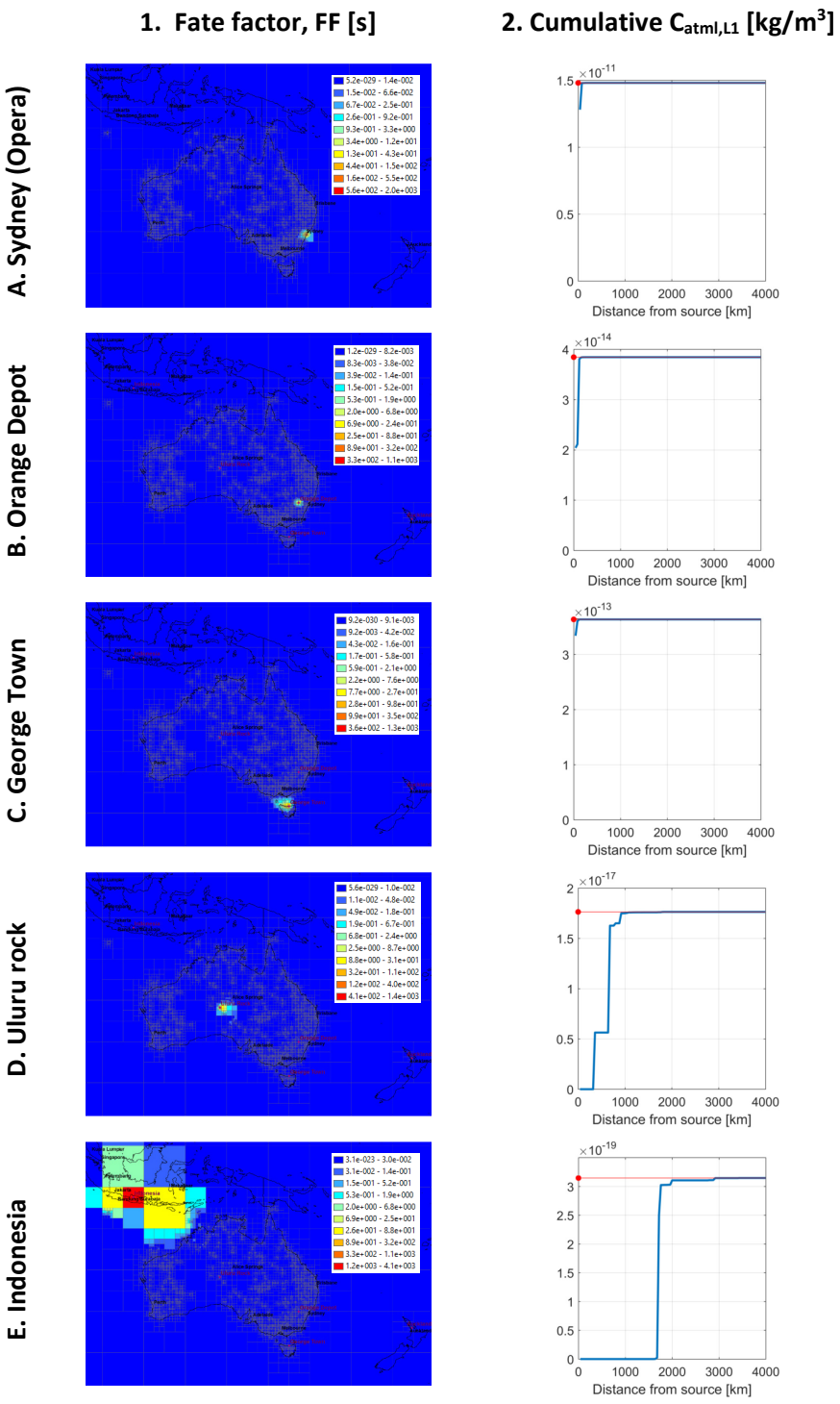


Figure 5.12 – **1.** Maps of fate factors (increase in receptor concentration due to emission of 1 kg/s in the considered cell) and **2.** radial statistic of contributing Australian sources for benzene, for receptors in **A.** Sydney opera, **B.** Orange depot, 200 km North-West of Sydney, **C.** George Town, Tasmania, **D.** Uluru Rock, and **E.** Indonesia. The sum of remote contributions to the local concentration, computed with source apportionment, converges towards the concentration computed based on total emissions.

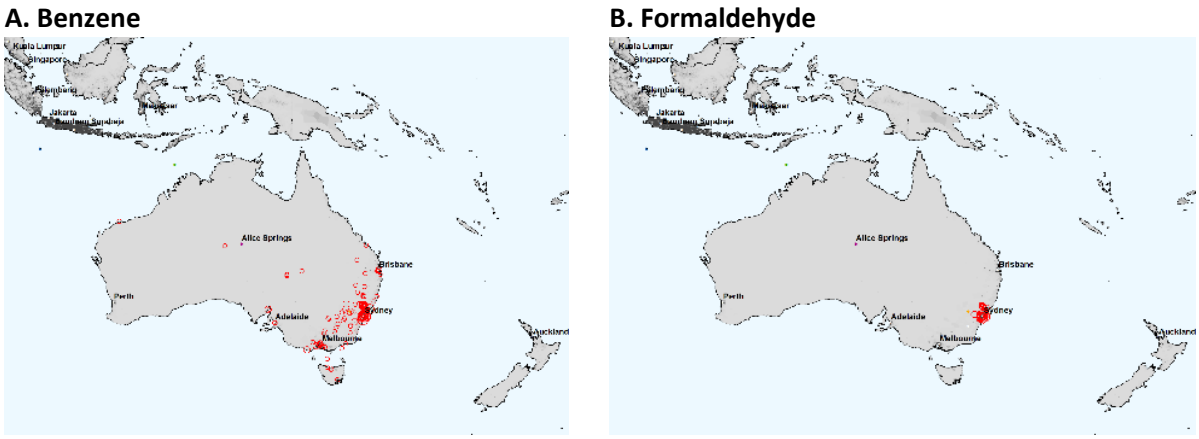


Figure 5.13 – Maps of locations of main sources contributing to concentration of **A.** benzene and **B.** formaldehyde at the Sydney Opera.

Tracing further major contributors – Using lists of largest contributing cells, we can extract NPI reports and rank them by emission. Taking the Uluru rock, for example, we find that the major contribution to the local atmospheric concentration of benzene is cell 7183 (Figure 5.14), North-West of Adelaide, which encompasses emissions from NPI reports {2469, 2946, 3752}. Table 5.1 shows the ranking of these reports by total emission of Benzene during year 2014, differentiated by industry sector to identify the main contributing industries.

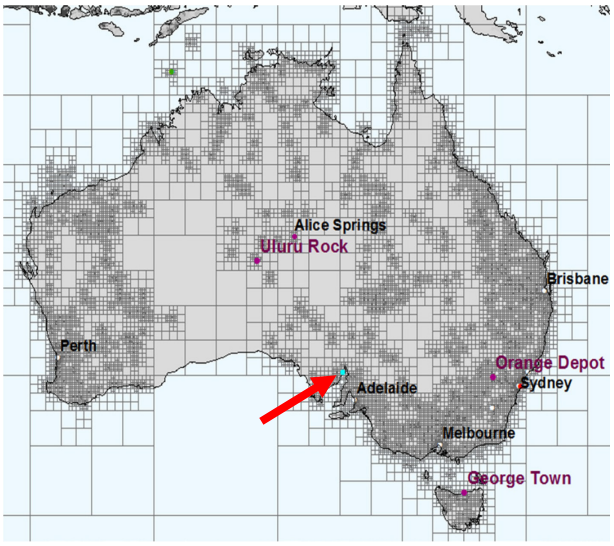


Figure 5.14 – Identification of the cell that contributes the most to the atmospheric concentration of benzene at Uluru Rock.

Table 5.1 – Ranking of reports associated with cell 7183, major contributor to the intake of benzene at the Uluru rock.

Rank	Report ID	Company sector	Total emission [kg] for 2014
1	3752	Iron Smelting and Steel Manufacturing	4.7×10^4
2	2946	Rail Freight Transport	0
2	2469	Other Non-Metallic Mineral Mining and Quarrying	0

5.4 Discussion and conclusion

Emitter perspective – This analysis shows that for a continental region such as Australia, with a high spatial variation in population density, iFs and local exposures vary substantially (0.68 ppm to 33 ppm for benzene, and 5.6 ppt to 9.5 ppm for formaldehyde) depending on the source and population densities, especially for short-lived chemicals.

Comparison with USEtox – The comparison of global iFs through inhalation and total ingestion associated with all emission cells for the 43 substances considered, shows that results computed by *Pangea* and by USEtox are generally within an order of magnitude. This is an interesting result, because it shows that if we are only interested in global iFs, USEtox urban and continental (rural) archetypes capture well a large fraction of the spatial variability that Pangea accounts for. *Pangea* primarily enables the user to further refine the calculation of each source-specific iF.

Emitter versus receptor perspectives – The emitter perspective focuses on where a substance is transported after it is emitted, and the subsequent population exposure (intake or iF). The receptor perspective addresses the reasons for the environmental concentrations and for the population exposure at a given location. Most analyses (e.g. radial statistic) can be performed from both perspectives. Emitter and receptor perspectives are therefore complementary. Figure 5.15 illustrates this complementarity through two maps.

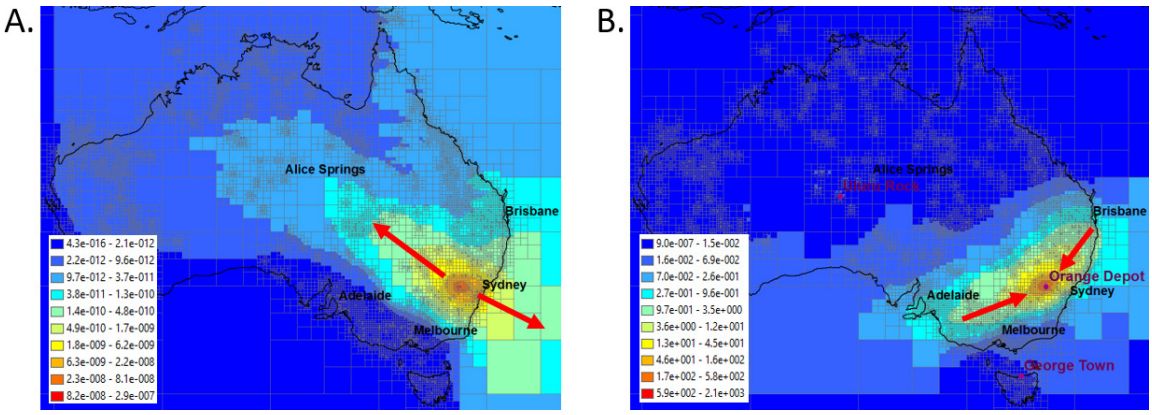


Figure 5.15 – **A.** $C_{atm,L1}$ [kg/m^3] for a unit emission of benzene in Orange Depot, **B.** Fate factor FF [s] for benzene for Orange Depot.

Figure 5.15 A shows the concentration of benzene in the first layer of the atmospheric grid, resulting from a constant unit emission (1 kg/s) at Orange Depot, 150km West of Sydney. We observe two plume directions: North-West and South-East. This spatial distribution of concentrations can be seen as a potential for exposure (everywhere) associated with a point source emission at Orange Depot: its combination with the spatial distribution of population (receptors) defines the spatial distribution of

population exposure. Figure 5.15 B shows fate factors for benzene in Orange Depot, i.e. the substance mass increase in Orange depot per kg/s emission flow in any other cell, which can be interpreted as a potential for exposing Orange Depot (from everywhere). Combining these spatially distributed fate factors with the spatial distribution of emissions (emitters), defines the concentration at Orange Depot and by extension the exposure at Orange Depot. We observe that this "inverted plume" defined by the fate factors and the "emitter plume" from Figure 5.15 A are close to orthogonal. They reflect the average weather pattern for 2014 in this region, with main wind directions aligned with the red arrows.

Figure 5.16 compares the cumulative radial statistics associated with both perspectives. Figure 5.16 A shows the evolution of the cumulative iF by inhalation for an emission of benzene at Orange Depot[46]. The primary step in the curve starts ~100 km away from Orange Depot, when the plume reaches Sydney. The plateau that follows indicates that most of the intake happens in Sydney. Figure 5.16 B shows the evolution of the cumulative contributions of actual emissions of benzene defined by the NPI, to the atmospheric concentration at Orange Depot. The first step is again associated with the region of Sydney, and subsequent steps in the range 700 km to 1000 km, given the pattern of fate factors from Figure 5.15 B, are likely associated with sources in Brisbane, Melbourne, or Adelaide. Finally, we observe that the curve from Figure 5.16 B converges towards the absolute atmospheric concentration at Orange Depot, indicated by the horizontal red line.

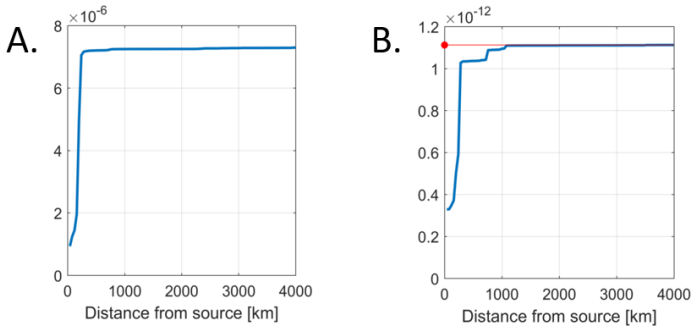


Figure 5.16 – **A.** Cumulative inhalation iF [-] for an emission of benzene in Orange Depot, **B.** Cumulative contributions to $C_{atm,L1}$ [kg/m^3] at Orange Depot.

Numerically, the two perspectives require working with substantially different matrices and approaches. The emitter perspective involves the **K** matrix, which is sparse and requires less than 20 MB of memory (RAM) to store for typical projects. One the contrary, the receptor perspective involves the **FF** matrix which is dense and can neither be computed nor be stored in memory (if would require ~100 GB of RAM to store a single copy of it for typical projects). *Pangea* can however compute blocks or bands of the **FF** matrix (Figure 5.2), which allows to analyze specific regions, and to perform source apportionment.

[46] Linearity implies that this result, as it is a fraction, is valid for any emission intensity.

Multimedia source apportionment – The algorithm for computing blocks of the **FF** matrix is not limited to specific media or regions, but by the size (dense) block. In this study, we wanted to analyze a receptor perspective for multiple atmospheric cells and we computed the block corresponding to transfers from the first layer of the atmospheric grid to itself. This layer is made of 18,107 cells and the block is hence a 18,107×18,107 dense matrix (Figure 5.2) whose size in memory is ~2.6 GB. Choosing a smaller number of receptors, however, would allow to perform a source apportionment accounting for more sources, even in other media. For the current system with 109,766 compartments, 2.6 GB (which is not a limit but gives an idea) is the size of a block of **FF** that enables to perform a source apportionment between 3,000 receptors an all possible source compartments (in all media).

Main limitations – The primary limitation is that NPI is limited to industrial sources[47] and there is no large-scale consistent repository of environmental levels for the substances present in NPI. This prevents to perform and evaluation against measurements. The study was also limited to atmospheric emissions, because a substantial fraction of emissions defined by NPI are atmospheric, and because the current hydrological model is singular over most of Australia.

Conclusion – The multi-scale approach enables us to assess the spatial distribution of iFs both from an emitter perspective and from a receptor perspective. The two representations can be important depending on the questions addressed: the emitter perspective is well suited to inform product oriented approaches such as Life Cycle Assessment, whereas the receptor perspective is well suited to allocate exposure to emission sources. Prioritization schemes for intervention need to consider consumption, emission sources, and exposed receptors, to determine priorities for impact mitigation.

An important limitation from the receptor perspective is related to the ingestion pathway, which is production rather than consumption oriented. This means that the food produced at one location may be consumed in other locations. It would be of interest to combine, in future research, the source to impact model *Pangea* with the Australian Multi-Regional Input-Output economic model developed at Sydney University, to analyze the environmental health effect and burden of disease of Australian consumption, accounting for food transport and consumption in addition to the atmospheric transport.

[47] Except over specific air sheds, for which emissions from non-industrial sources (e.g. motor vehicles) were estimated, but that we did not include in the study.

Chapter 6 Conclusion

The objective of this thesis was to study the fate and transport of substances on local to global scales and examine subsequent population exposure. In Chapter 1, we defined four specific aims to achieve this objective:

SPECIFIC AIM 1: Develop a flexible multi-scale modeling framework, that models multimedia fate of substances (e.g. pollutants) and the subsequent multi-pathway exposure of populations, and which provides steady-state and dynamic solutions at local to global scales.

SPECIFIC AIM 2: Evaluate the model and apply it to releases of chemicals used in household and personal care (HPC) products for all of Asia.

SPECIFIC AIM 3: Study the radial distribution of intake fractions (iFs) through inhalation and ingestion of a selection of substances emitted from solid waste treatment plants (SWTPs) located in France.

SPECIFIC AIM 4: Study variations in population exposure in Australia based on emissions defined by the National Pollutant Inventory (NPI). Study emitter perspective per sector, as well as receptor perspective (source apportionment).

Sections 6.1 to 6.3 present key achievements and limitations of the three studies performed using *Pangea* (aims 2 to 4). Section 6.4 describes the *current* version of the framework (aim 1) as of May 2017, its strengths, and its limitations. Section 6.5 discusses the novelty of the modeling strategy and implications of this work. Section 6.6 describes future work and demonstrates how this thesis is integrated in a long-term research effort. Finally, Section 6.7 delivers a more personal conclusion and presents the future of the framework as I would like it to evolve.

6.1 Model evaluation, exposure of ecosystems to HPC product chemicals in Asia

We performed this study in collaboration with the Unilever Safety and Environmental Assurance Centre (SEAC) team, who identified *Pangea* as a key platform to enable spatial analysis of freshwater ecosystems exposure at the continental level. The project focused on the *fate and transport* of 16 HPC

chemicals emitted in Asia and their presence in Asian *ecosystems*. It provided us with a unique opportunity to evaluate the framework based on *large-scale* high-resolution *spatial inventories of emissions* and a *large database of more than 1,600 monitored environmental levels*, in both fresh water and sediment. We selected four substances for the publication because they were representative of four categories of HPC products with the most monitored values available. These substances were the anionic surfactant *linear alkylbenzene sulphonate* (LAS), the antimicrobial *triclosan* (TCS), the personal care preservative *methyl paraben*, and the emollient *decamethylcyclopentasiloxane* (D5).

Model evaluation – The evaluation consisted of a model-model comparison between *Pangea* spatial results and USEtox non-spatial estimates, as well as a model-monitoring comparison between *Pangea* and monitoring values for the whole of Asia and at sampling sites. We found good agreement (generally below an order of magnitude) between median concentrations from *Pangea* and USEtox estimates. We also found a good agreement between modeled and monitored concentrations in fresh water, especially given that results are the outcome of forward modeling with no adjustment whatsoever. The best agreement is for TCS in fresh water (Pearson r = 0.82), which is a long-lived substance in fresh water ($DT50_{water}$ = 23 days) and which has the largest number of monitoring values (253 in fresh water). We generally found a moderate agreement for sediments, except for parabens (methyl, ethyl, propyl) for which there was a 3.5 order of magnitude discrepancy. Due to its importance, and although the variability of measurements in sediments can be large on short scales, this discrepancy is highly interesting to us because it constitutes a likely case where the model did not work. Sensitivity analysis suggests that the uncertainty on parameters can explain a maximum of half the discrepancy, which implies that there is a mechanistic issue with environmental process models (EPMs) associated with sediments. Early investigations convey that the burial process may need refinement, because it is the dominant removal mechanism for the model when measures indicate otherwise.

Limited literature about LAS levels in Asia – While we observed emission levels and predicted environmental concentrations for LAS (a historical pollutant) that were two to three orders of magnitude higher than for the three other substances, we found very little literature about measurements of environmental levels of LAS in most of Asia. This contrasts with TCS, an emerging pollutant that is well reported and analyzed in the literature despite relatively small emissions and environmental levels compared to LAS.

Flushing effect and local assessment – We analyzed key parameters in the prediction of environmental concentrations and the possibility of using local assessments rather than modeling the full transport from upstream regions. We found that for the short-lived substances (LAS and methyl paraben, $DT50_{water}$ = 0.29 and 0.83 days respectively), spatial variations in freshwater concentrations primarily reflect

variations in the equivalent volume of dilution, with higher concentrations in regions with a low volume of dilution. A "flushing effect" was noticeable for these substances, with concentrations in fresh water lower along e.g. the main stream of the Yangtze River (large flow) than in its tributaries (lower flows) and water basins. This effect diminished or disappeared for more persistent substances, such as TCS and D5 ($DT50_{water}$ = 23 and 60 days respectively), leading to more uniform concentrations, even along the main stream of the Yangtze River and the rest of the catchments. We demonstrated that, as a consequence, local modeling can substantially underestimate concentrations in fresh water (by a factor of 10 or greater, for up to 20% of the cells present in the region of study) for long-lived substances, such as TCS and D5, because it accounts for the flushing, but neglects a substantial transport from upstream regions. This is also reflected in the much larger number of upstream cells (over thousands of kilometers) substantially contributing to the concentration of long-lived substances at a given location, compared to the mostly local contributions for shorter-lived substances (Appendix A).

Main limitations – The default hydrological data set underlying the *Pangea* hydrological model, WWDRII, routes the hydrology on a 0.5°×0.5° grid. This led to mis-associations between measurement sites and rivers as defined on the grid, which required manual or semi-automatic correction. This motivated and induced a transition from the gridded WWDRII-based hydrological model to a higher resolution and catchment-based hydrological model built on HydroBASINS. This study has also identified that some vast regions of the globe that are singular (net flow close to 0 or null) for the hydrological network – and are therefore often *lost to the model* – are in fact not remote deserts, but highly populated and highly polluted regions. HydroBASINS can address this issue.

Conclusion and perspective for the future – The framework developed for this study, which couples *Pangea* with the UNILEVER ScenAT model (scenarios of emission), constitutes a milestone towards higher tier modeling approaches for ecological risk assessment. The identification of worst-case (sub)catchments in combination with source apportionment analysis for a given chemical can help prioritize further investigations. The combination of the methods presented in Chapter 5 – source apportionment based on the computation of large blocks of the matrix of fate factors – with the new hydrological model based on HydroBASINS can enable the study of regions that are known to be highly polluted (e.g. in the region of the Yin river), and whose hydrology is currently absent from the model. We foresee that the next relevant limitation will be associated with input data, which may fail to accurately define the locations and modes of entry for emissions in the hydrological system.

6.2 Local to global population exposure to emissions from point sources

This study has focused on human exposure and analyzed the evolution of the population iF, the fraction of the emission taken in by the population, as a function of the distance from source. We simulated emissions from 126 point sources (stacks) corresponding to the location of SWTPs in France and computed the radial distribution of population iFs through inhalation and ingestion. We extended this study by simulating emissions from ~10,000 point sources covering France more densely.

Summary and insights – We found that a substantial fraction of emissions may be taken in by the population farther than 100 km away from point sources (78.5% of the inhaled benzene and 54.1% of the ingested 2,3,7,8-TCDD). We demonstrated that a spatially explicit model based on the multimedia principles of a model such as USEtox corroborates the result from Lohman and Seigneur (2001) when applied to a case of steady, long-term emissions over Europe.[48] This highlights that assessments limited to a local/regional domain **may** miss a substantial fraction of the cumulative exposure, and emphasizes the importance to consider both local and long-range exposures. The sensitivity study in Section 4.6 evidences that this result is stable if we vary the period and resolution of the atmospheric wind data set. We then simulated fate, transport, and population exposure associated with emissions from ~10,000 locations covering France more densely, and built maps of global iFs corresponding to all emission locations. These maps can be seen as *potentials* for generating population intakes. We found large inhalation iFs for benzene associated with emission locations upwind from regions with large populations (mainly in Paris, Belgium, Germany, and the Netherlands), and we found that the patterns in maps of global ingestion iFs for both TCDD and B[a]P match well the distribution of large crops (maize, wheat, etc) and animal-based agricultural regions in France. We compared global iFs distributions (based on ~10,000 emission scenarios) with USEtox results for urban and rural emissions in Europe. There is a good agreement for inhalation, and a systematic discrepancy of one order of magnitude for total ingestion, where *Pangea* estimates are larger than those of USEtox. This work has also tested and demonstrated the feasibility of simulating large numbers of emission scenarios in *Pangea*.

Main limitations – The primary limitation is that this is a modeling study, so results could not be evaluated against measurements or more specialized atmospheric models apart from the initial result that Lohman and Seigneur obtained through the ISCLT dispersion model. The key reasons are that there is no large-scale emissions inventory (in fact, point sources are not the main sources of emission of the three substances tested) and that, to our knowledge, very few large-scale inventories of environmental

[48] For PCDD/F, less than 10% of emissions from tall stacks (simulations from six different geographic locations in the US) may be deposited within 100 km from the source.

levels are available.[49] The second limitation is that *Pangea* uses meteorological fields from global atmospheric models, the resolutions of which do not allow for capturing fine details of atmospheric dispersion close to ground level (which depends on the terrain, constructions, vegetation, etc.). For this reason, we built the first layer of the atmospheric grid to cover altitudes from ground level to ~100 m above ground level, and we assumed full vertical mixing. While *Pangea* could further discretize this region more, full mixing is the best assumption that could be made given the current maximum resolution of the wind data set.

Conclusion and perspective for the future – Population exposure associated with emissions from a point source results from the conjunction of two opposed trends: environmental concentrations that are decreasing with the distance from source and the size of the exposed population that is generally increasing with distance. This may lead to situations in which a substantial population intake occurs farther from sources than typical external radii of local modeling. When exposure doses due to background levels are already high, or for adverse effects with no threshold in the dose-response curve, this induce more cases of adverse effect remotely than locally. Technically, the first part of the study (radial evolution of cumulative iF) demonstrates the feasibility of systematically performing local to global assessments for point sources. The second part of the study (~10,000 emission locations) was performed before a recent update in the exposure model of *Pangea*, which allows it to more efficiently filter cells that are singular for the hydrological network. After the transition to HydroBASINS, a next step is to re-analyze this result and discuss how it is influenced by the choice of hydrological model (in a fashion similar to re-analyses following the evolution of GEOS-Chem).

6.3 Receptor perspective and source apportionment

We finally extended the previous emitter-oriented studies and the framework to support receptor-oriented analyses, i.e. source apportionment. We simulated the fate and transport of 43 substances emitted from 4,101 point sources defined by the Australian NPI for the period 2014-2015. We computed subsequent population exposure, effects (cases of non-cancer and cancer), and severity (DALY). Formaldehyde, benzene, and styrene were determined to be the three top contributors in terms of DALYs. We demonstrated the technical feasibility of multimedia, large-scale source apportionment. Exploiting NPI meta-information and source apportionment allows *Pangea* to analyze impacts per industrial sector, per industry, per region, or overall, and to trace back to major contributors (from

[49] Some national inventories of emissions are available, but they are often not exhaustive (e.g. industrial and/or self-reported). A few inventories of environmental levels could be used in future work, however, such as the US EPA National Dioxin Air Monitoring Network (NDAMN) and the Global Atmospheric Passive Sampling (GAPS) network.

emission inventories) associated with environmental levels and population exposure at any point on the globe.

Summary and insights – The emitter perspective showed that for a continental region such as Australia, with a high spatial variation in population density, intake fractions and local exposures vary substantially (0.68 ppm to 33 ppm for benzene, and 5.6 ppt to 9.5 ppm for formaldehyde) depending on the source and population densities, especially for short-lived chemicals. We found that *Pangea* and USEtox are generally within an order of magnitude when computing global intake fractions. This shows that for global intake fractions, USEtox urban and continental (rural) archetypes capture well the spatial variability that Pangea accounts for, and therefore that *Pangea* primarily enables the user to further refine the calculation of each source-specific iF.

Main limitations – The primary limitation is that NPI is limited to industrial sources[50] and there is no large-scale consistent inventory of environmental levels for the substances present in NPI. This prevents to perform and evaluation against measurements. The study was also limited to atmospheric emissions, because a substantial fraction of emissions defined by NPI are atmospheric, and because the current hydrological model is singular over most of Australia.

Conclusion and perspective for the future – The multi-scale approach enabled us to assess the spatial distribution of intake fractions both from an emitter perspective and from a receptor perspective. The emitter perspective is well suited to inform product oriented approaches such as Life Cycle Assessment, whereas the receptor perspective is well suited to allocate exposure to emission sources. This project demonstrated the feasibility of performing source apportionment based on large blocks of the matrix of fate factors. This indicates that large-scale multimedia source apportionment is achievable. The coupling with NPI and the use meta-information also showed that Pangea could manage large inventories of emissions, and trace back major contributors (per sector and/or per company) in databases.

6.4 Strengths and limitations of the Pangea framework

This thesis achieved the objective of creating a fast and spatial evaluative framework for studying substance fate and transport and subsequent population exposure.

Main characteristics and strengths – The *Pangea* framework departs from former generic and spatial multimedia fate and transport and multi-pathway exposure models with fixed parameterization and fixed geometry. It has the technical advantage of incorporating GIS abilities to enable fast, project-specific

[50] Except over specific air sheds, for which emissions from non-industrial sources (e.g. motor vehicles) were estimated, but that we did not include in the study.

multi-scale spatialization of the globe, with local resolution where relevant. Focusing on the relevance for researchers and decision-makers, its main features are as follows:

Scope	:	Spatial multi-scale, local to global, multimedia fate and transport, multi-pathway population exposure.
Flexibility	:	A GIS engine allows for building project-specific spatial parameterization (geometry and projected spatial data). Procedures for importing/processing raw data are programmed, which makes it possible to follow the growth of geo-referenced data sets as they are updated or become available.
Type of studies	:	Substances fate, transport, and population exposure, from emitter and receptor perspectives. The receptor perspective allows for large-scale multimedia source apportionment.
Typical run time	:	Less than 10 minutes for building standard projects, geometries, and projecting spatial data. Then, a few seconds per substance for steady-state solution (much more for dynamics). Large numbers of emissions scenarios can be computed in parallel. Building output files and reports can be time consuming, especially for maps.
Models and parameters	:	By default, meteorology from GEOS-Chem; hydrology from WWDRII (HydroBASINS currently evaluated); land cover from GlobCover; population from CIESIN; fate, transport, and exposure models from IMPACT/USEtox, transitioning to USEtox v2; generic parameters from USEtox; physical and chemical parameters from cascaded databases.
Main limitations	:	First-order fate and transport processes; default fate and transport models are linear; global spatial data sets are rarely high resolution.
Software	:	Common but licensed: MATLAB (with Mapping and Statistical toolboxes) and currently ArcGIS.

The main strengths of the framework are its flexibility and versatility, as well as the fact that it builds on state-of-the-art and peer-reviewed models and data sets (USEtox, GEOS-Chem, HydroBASINS). The flexibility derives from both the GIS engine and the object-oriented structure of *Pangea*. The framework is not monolithic, but constructed as a set of distinct components (core, EMs, EPMs, etc.) with normalized interfaces for integration/communication, which facilitates the selection of underlying models and data sets (even on a per-project basis). In addition, we implemented a series of tools for processing data sets and databases and for wrapping other models. This facilitates model-model comparisons and allows *Pangea* to perform them automatically/systematically, and it permits us to re-

analyze projects using updated data sets. For example, the sensitivity study and re-analysis of the radial evolution of iFs using a series of meteorological fields from GEOS-Chem (for years 2013 to 2016, and for two resolutions per year) would have otherwise been technically very complicated and time-consuming.

To highlight the strengths of the framework, it is relatively straightforward for an advanced user who wants to test an alternate atmospheric fate process model (degradation EPM in air) of her or his own to build an EPM set based on the default (USEtox) EPM set, but replacing the relevant EPM, and simulate emissions of the ~3,100 substances present in the USEtox database, which are emitted from ~10,000 locations around the globe (with grid refinement around them). It is straightforward to compare the results with results obtained using the default EPM set, and also to study the sensitivity to the meteorological fields (e.g. GEOS-Chem years 2011 to 2016, or ECMWF). Finally, it is straightforward to systematically compare the output with USEtox output.

Framework limitations – The main limitation is that *Pangea* uses and imposes first-order EPMs. This is a fixed limitation, by choice, because models that account for second-order processes belong to another category of models (computation time, data, and resource intensive, not suited for operating in global multimedia contexts). The second important limitation is a lack of data and proper methodology for modeling marine currents in the coastal zone. This prevents proper modeling of a region that is critical for human exposure, and which connects the freshwater hydrology to the oceans. This is a main limitation because we do not anticipate a change in data availability in the short to medium term. In addition, the default EPM set (USEtox) is composed of linear EPMs; this is not a limitation of the framework, strictly speaking, since other non-linear EPMs could be considered, but updating a full set of EPMs to account for non-linear processes (e.g. saturation) is a consequent work. Finally, *Pangea* is currently optimized for computing and analyzing the steady-state solution (of large numbers of emission scenarios). While it can solve the dynamics, the current solvers and tools for analyzing the dynamics are at a developmental stage.

Quality and consistency of physical and chemical input parameters – Numerous databases of physical and chemical properties are available, but none covers all substances and all properties. *Pangea* addresses this issue by cascading databases with a per-property prioritization schema. For example, EPIsuite's experimental values have a priority over extrapolated or modeled values, and degradation half-lives from ECHA have priority over degradation half-lives from USEtox. Data sources and prioritization can be fine-tuned per project or completely overridden. While this allows for obtaining values from the data sources that we think provide the highest quality, the implication is that (using the default prioritization schema) the final values for all relevant physical and chemical parameters often

come from multiple databases. By using this system for studying the distribution of degradation half-lives in particular, we could observe large variability among databases. *Pangea* hence solves the technical problem of incomplete data sources and prioritization. However, the final selection (and its quality) must be discussed for each project that is not bound to a specific database.

Quality and consistency of non-spatial input parameters – The strategy for integrating data sets in *Pangea* consists of using USEtox generic parameters when spatial parameters are not available. Generic parameters are also used as a default value when spatial parameters are not global. A typical example is precipitations that are spatially defined based on the TRMM data set, which is limited to a 60° N-S band. Projects can use USEtox generic values globally, or TRMM in this band and USEtox values outside of it, but the discrepancy at the interfaces must be carefully handled and discussed. *Pangea* currently defines about 40 generic parameters, which can be redefined in projects. The EcoHOPE project, for example, redefines the default *total suspended solids in fresh water* (TSS) value of 15 mg/L, which is the USEtox global average, as 100 mg/L, based on a literature review of TSS values in Asian rivers. Roughly half of the current generic parameters correspond to quantities that have spatial variability (e.g. TSS). The long-term objective is to spatialize them, but global spatial data sets are not necessarily available and difficult to build. On the short term, the GIS engine allows *Pangea* to define some parameters per geographical feature, e.g. TSS per main stream in Asia (when available in the literature), while using a default value elsewhere.

Quality and consistency of spatial input parameters – We have already discussed the lack of large-scale spatial inventories of emissions[51] and measured environmental levels. The other main spatial data sets are the meteorology defined by GEOS-Chem (global 2°×2.5° grid with nested 0.25°×0.3125° continental grids), the hydrology (0.5° ×0.5° grid for WWDRII and 10 levels for HydroBASINS), a high-resolution land-cover data set, a data set of population counts, and a data set of agricultural production. Meteorological fields from GEOS-Chem provide global consistent time-series for a large number of atmospheric parameters. Using a global 2°×2.5° grid with a nested 0.25°×0.3125° grid for a single continent requires processing a total of ~240 GB of raw data while offering a maximal resolution at the equator of ~34 km (between nodes). Raw data are pre-processed and the full processing is not required on a regular basis, but this evidences that it is unlikely that *Pangea*'s underlying atmospheric data set will be adequate for accounting for e.g. local terrain geometry in the short term. There is currently no global hydrological data set that accounts properly for lakes. HydroBASINS provides

[51] National inventories exist, but they are often industrial and/or self-reported, and hence not exhaustive. Further evaluation projects could nonetheless simulate global inventories of emissions of PCBs (Breivik et al., 2016, 2007), PAHs (Shen et al., 2013), PFCA (Wang et al., 2014), or PFSAs (Wang et al., 2017).

separate river and lake data sets, as well as lakes' pour points (as geo-referenced points). WWDRII was updated to account for main lakes in a multimedia framework, but the proper integration of a larger number of lakes after switching from WWDRII to HydroBASINS will be a point of special attention. Finally, there is currently no high-resolution spatial global data set of food production. *Pangea* currently distributes FAO national production inventories using the land-cover data set as a proxy. This mechanism could be improved, or better evaluated, at least by using inventories at a county level in the United States. *Pangea* is limited by the low availability of global spatial data sets. However, it implements the tools necessary for processing and importing from raw data formats, which allows us to follow the growing availability of geo-referenced data sets and their evolution.

Model evaluation – Previous spatial models based on IMPACT 2002 could only be evaluated using regional or continental averages. The influence of their geometry could not be evaluated, and they could not be tested against the evolution of their spatial parameters, as they were the outcome of single, time-consuming manual processing. *Pangea* enables more specific spatial evaluation, for example by targeting specific rivers, as detailed in Chapter 3. Results for TCS and methyl paraben indicate that we could perform a successful evaluation for the main rivers and identify river-specific issues (methyl paraben in fresh water, with Yangtze and Pearl rivers as outliers) as well as substance-specific issues (methyl paraben in sediments). The influence of the geometry can also be tested. The sensitivity study in Chapter 4 presents two situations: base geometry with a first atmospheric layer that covers altitude ranging from 0 to 127 meters (~100,000 compartments overall) and a finer geometry where this first layer is split into 10 sub-layers (~260,000 compartments overall). In both cases, a geometry refined around 126 point sources is evaluated, facilitating the development of a statistic on 126 types of local contexts, as well as the influence of the layer of emission on radial statistics. Finally, the implementation of all GIS and spatial data processing allows for following the evolution of spatial data sets and for following the subsequent evolution of results. While this would be a considerable amount of work to perform manually, *Pangea* enabled us to process ~850 GB of meteorological fields, as well as to test the influence of the resolution and the year on radial statistics (section 4.6) within a few days.

Limitations and implications for decision-making – While *Pangea* has many strengths and addresses gaps in some decision-making and research applications, it is important not to oversell it. When run with USEtox/SimpleBox EPMs, *Pangea* can reveal the predictions of these models when they are parameterized to account for project-specific spatial resolutions and parameters. In this sense, *Pangea* is already a significant technical achievement. However, as a global multi-scale framework, *Pangea* cannot be validated, or even evaluated, for all substances, all locations, all scales, etc. The question of accuracy, and absolute accuracy in particular, must therefore be addressed on a per-project basis.

Pangea thus fulfills the *technical* challenge regarding fast spatialization of a global multi-media environment, but it leaves users responsible for evaluating and validating their modeling approach. However, *Pangea* provides numerous tools for facilitating this discussion, and particularly the ability to focus on regions in which measurements are available and account for local spatial variability.

6.5 Novelty and implications of this work

The novelty of the *Pangea*-based approach presented in this thesis is that it encompasses a substantial number of studies and contexts that are highly relevant to public health and environmental sciences: simulation of fate and transport; computation of human exposure (and further effects and severity); spatial, local to global scales; simulation of large inventories of emissions; simulation of large numbers of emission scenarios; comparison with geo-referenced monitoring data; studies from emitter and receptor perspectives; and multimedia source apportionment. The novelty of this work is that it includes all of these features and topics within the same consistent framework. Several aspects deserve further discussion:

Spatial resolution versus archetypes – Previous models rely on archetypes for characterizing situations (e.g. urban or rural type of emissions, or up-wind/down-wind) because they lack the ability to exploit spatial data. *Pangea* can capture these distinct (and extreme) situations, but it can also account for all the intermediary situations. The study presented in Chapter 4 involves first a distribution of 126 emission locations that capture most of the possible situations (rural/urban, up-/down- wind/stream from large populations, or agricultural resource, or important geographic feature), and second a distribution of ~10,000 emission locations that is likely to capture them all. Interestingly, this second part and the comparison with USEtox demonstrate that when we are interested in global intake (or iF) through inhalation, USEtox results for the two archetypes (urban/rural) match well with the distributions computed by *Pangea*. For B[a]P and TCDD, they even frame the distributions, whereas for benzene, they fail to capture to upper part of the distribution. The reasons for these differences will be investigated once the full USEtox v2 EPMs are implemented in Pangea. This reveals that there are cases in which archetypes provide enough information for non-spatial models to capture the whole variability of spatial ones (when one only needs global or large-scale results). *Pangea* allows researchers to systematically study these cases, accounting for global distributions of situations (or per USEtox region for example), and identify relevant archetypes.

Spatial resolution and interface with higher resolution local/regional models – *Pangea* aims to create a bridge between generic or low-resolution models and more specialized, principally atmospheric local/regional models. The output of generic or low-resolution multimedia models could not be

evaluated or compared against the output of specialized atmospheric dispersion models (e.g. AERMOD) because there was no connection or overlap between their scales or domains. While *Pangea*'s underlying data sets and first-order EPMs prevent the achievement of high-resolution modeling, Figure 4.6 illustrates that the highest resolution meteorological field defined by GEOS-FP covers a 100 km radius (typical for local/regional dispersion models) with around 50 nodes per layer. This constitutes a good basis for comparison and for calibration of *Pangea*. The direct implication is that it becomes relevant to work on interfacing *Pangea* and specialized atmospheric dispersion models, and possibly on a wrapper for AERMOD that allows for automatic comparison (similarly to the ability of USEtox).[52]

Spatializing socio-economic factors – Inserting this work in the broader perspective of fast, total-exposure evaluation discussed in the introduction, it was particularly interesting to couple ScenAT and *Pangea* because of the ability of the former model to spatialize socio-economic factors. Although our ability to spatialize fate and transport modeling has been rather steadily increasing over the past decades, the spatialization of socio-economic factors has been encountering significant difficulties. Yet, evaluating total exposure requires evaluating exposure to products (use phase) and exposures that are specific to living and working conditions, and thus tightly bound to socio-economic factors. Shaked et al. (2011) have made an effort to include a trade model in IMPACT World, but it will likely be difficult to spatially refine such models at a finer level than countries/counties and globally in the near future. However, ScenAT demonstrates that it is possible to spatialize emissions based on sales inventories, nighttime light as a high spatial resolution proxy for GDP, and databases of GDP to product consumption. This approach is interesting and promising because it opens the path to integrating spatial socio-economic factors in *Pangea*, which is crucial to the evaluation of the total exposure.

Receptor perspective and source apportionment – Most studies of large systems have focused on an emitter perspective (starting from emissions and simulating the fate, transport, and population exposure). Source apportionment (or receptor perspective) is complex to implement, and is generally limited to single-medium contexts (e.g. only atmospheric, or only hydrological), especially when it must account for the dynamics. The techniques developed for Chapter 5 enable multimedia source apportionment at *steady-state*. While this is theoretically straightforward in first-order linear systems, using the matrix of fate factors, **FF**, it is complicated in systems as large as *Pangea*, because **FF** cannot be computed. *Pangea* can compute "large" blocks of **FF**, covering multiple media, which facilitates multimedia source apportionment. This permits systematic hotspots analysis with source apportionment, for example, and

[52] We are mentioning AERMOD because it is a set of pre/post-processors and model executables that work with configuration, input, and output files in text, image, and GIS formats, which can all be managed by *Pangea*.

consequently the performance of all types of studies that are usually conducted from an emitter perspective (e.g. radial analysis) from a (complementary) receptor analysis instead.

Feasibility of simulating large inventories of emissions – *Pangea* incorporates parsers and wrappers for a series of national emission inventories. The parser for the Australian NPI allowed for importing all available years, and the wrapper for NPI enabled *Pangea* to access the NPI as a database. This makes it easier to analyze e.g. all years for which measured environmental levels are available. More generally, the flexibility of the framework render it technically straightforward, relatively speaking, to simulate all years or versions of e.g. most national emission inventories, using geometries that are refined around major emitters and/or receptors.

Flexibility of the framework structure and sensitivity studies – Current and former multimedia models are based on a one-time fixed parameterization based on spatial data (e.g. consequent manual work from GIS specialist) and a fixed selection of fate and transport models (deeply integrated in their structure). The question of the influence of these one-time selections on the results could not be addressed. As discussed in Section 6.4, *Pangea* offers adequate flexibility for testing the influence of these choices by evaluating projects using a variety of combinations of spatial data sets and EPM sets. The next phase of the EcoHOPE project, which involves the comparison of the results based on WWDRII hydrology and HydroBASINS hydrology, will advance this ability.

Feasibility of re-analyses – More generally, the flexibility of the framework combined with the implementation of the all pre-processing for data sets and wrappers for external models could allow us to perform sensitivity studies on data sets and internal models, as well as model-model comparisons, on a systematic basis. This was demonstrated with the re-analysis of the radial study from Chapter 4 using eight additional meteorological data sets (2013-2016, with two resolutions per year), and with the comparison with USEtox results in both Chapter 3 and Chapter 4.

6.6 Future work

The EcoHOPE project is driving short-term future work, which requires finalizing the integration of HydroBASINS and the transition to the latest version of USEtox EPMs. In parallel, and after this is done, Pangea will undergo a series of updates and refactoring that will improve the framework's stability and evaluation procedures and more effectively exploit available data. In the longer term, users and projects could extend *Pangea* to become a stable reference platform for spatial fate and exposure modeling based on first-order processes.

Finalization of the integration of HydroBASINS – HydroBASINS was integrated for evaluation in *Pangea*, and this work will be finalized within the next phase of the EcoHOPE project. This implies the

exploitation of many spatially differentiated parameters that HydroBASINS/ATLAS provides, as well as working on the integration of lakes defined in a separate data set called HydroLAKES (including geo-referenced pour points). This operation is highly complex, as there is no direct way of properly combining river basins (catchments) and lakes. Lakes can span hundreds of catchments, the geographic extents of which vary with the level/resolution of HydroBASINS (which is multi-scale in *Pangea*) in a fashion defined by the refinement procedure. These catchments are not always part of the same watersheds. The strategy for combining lakes and rivers is defined by the specific needs in *Pangea* and a prioritization of the discrepancies that can be tolerated in our context. More specifically, *Pangea* needs to alter the geometry of water bodies and land cover and relevant contact surface areas in a manner that generates adequate times of residence of the water in water bodies and adequate areas for deposition/volatilization between water and air and land and air.

Finalization of the integration of USEtox v2 EPMs – The official documentation of USEtox v2 was released in early 2017, and we need to build an EPM set that strictly matches USEtox v2 (the current EPM set is transitional, combining EPMs from IMPACT 2002, from USEtox v1, and from SimpleBox). This would make it possible to peg *Pangea*'s default EPM set to the UNEP/SETAC consensus model USEtox, and could facilitate strict comparison between spatial and non-spatial approaches. This will also extend *Pangea* to support ionizable chemicals.

Finalization of the EcoHOPE project – In addition to the above-mentioned operations, the next step in the EcoHOPE project requires hotspots identification and analysis and the implementation of ionizing radiations. Source apportionment was implemented for the study presented in Chapter 5, and EcoHOPE is likely to directly benefit from this (Appendix A shows example of large-scale source apportionment in Asia).

Solidification of the framework components test procedure – Currently, *Pangea* incorporates test routines for model components in a developer package that is not fully integrated and automated. Routines for testing the atmospheric model in limit and/or trivial situations were run manually each time the atmospheric model was updated. Yet, as the framework is flexible, these routines must become more accessible to users, and *Pangea* must implement wider and more exhaustive test procedures that facilitates systematic evaluation.

Reduction to simpler systems – The first step for evaluating Pangea against a "simpler" model was to implement the systematic comparison with USEtox (13 compartments). After we finalize the implementation of USEtox v2 EPMs, it will be feasible to make further advances toward enabling *Pangea* to replicate USEtox 13 compartments system.

Extensive exploitation of recent versions of spatial data sets – Many global spatial data sets have been updated or have become accessible to *Pangea* within the last few years. While *Pangea* uses updated values, we did not increase the number of parameters that *Pangea* extracts from these data sets. For example, GEOS-FP and ECMWF provide a large number of atmospheric spatial parameters that we currently do not exploit. We will therefore establish a list of these newly available parameters and define a procedure for progressively incorporating them in *Pangea*.

Implementation of on-going test projects for global inventories of emissions and environmental levels – Chapter 3 and Chapter 5 demonstrate that *Pangea* can process and simulate large spatial inventories of emissions and compare predicted environmental concentrations with large inventories of monitored values (Chapter 3 only). *Pangea* could easily simulate all available inventories in a set of open and on-going projects, and compare predicted environmental concentrations with a database of environmental levels that we update progressively. Keeping these projects running, staying aware of the increasing availability of inventories, and performing systematic and automatic re-analyses could allow us to identify situations worth analyzing in further detail, e.g. observing a better match between predictions and measures for specific substances (or categories of substances) as inventories become more exhaustive.

6.7 Personal views about this work and the evolution of *Pangea*

From my perspective, multimedia models (aim to) implement current knowledge about chemical fate and transport into a common, unifying structure. MacLeod et al. (2010) have expressed this, describing these models as "*... a framework for organizing a wide array of information and concepts into a consistent and cohesive overall picture.*" Multimedia models are often used as evaluative models in tiered approaches during a screening phase that identifies situations that may require further analysis using more specialized methods of assessment. Risk screening requires conservatism, and would benefit from ability to define a tolerable amount of underestimates (false negatives). Like many previously developed evaluative models, *Pangea* lacks the ability to formally and precisely define such a tolerance. While implementing this *normative step* (EEA, 2007, review by G. Lammel, p.114) is a lengthy and complex task, *Pangea's* flexibility allows for the implementation of a shorter term series of small steps towards a more systematic estimate of the error. The possibility of evaluating large numbers of scenarios, processing spatial data sets at run time, and computing **K** matrices efficiently can enable an analyst to perform Monte-Carlo analyses or to define sets of parameters oriented towards conservative evaluations or best-estimates, which are more characteristic of life-cycle (impact) assessment (LCA/LCIA) studies.

Regarding model repositories, there is a plethora of models available that result from past and current modeling efforts. Models tend to have a life-cycle that is not dictated by the evolution of knowledge and science, but by the life-cycle of projects, funding, and people who can maintain them. Some models are recommended by environmental agencies but require a significant effort to deploy, so their use cannot be systematic. I strongly believe that the modeling community should capitalize on existing model availability and develop integrated assessment methods that perform model/model comparative studies automatically (or semi-automatically), as performed "manually" earlier (Armitage et al., 2007; EEA, 2007; Fenner et al., 2005; Hollander et al., 2007). With its wrappers for external models (e.g. for USEtox) and connectors, *Pangea* takes a step in this direction. Technically, *Pangea* can interact with most Excel-based models, run all executable models with batch inputs, build any configuration "text" file and process most output files of the same type, pre-process and post-process most GIS files, and use most file formats for storing large data sets (NetCDF, HDF). Keeping an evaluative objective in mind, this means that in addition to automatically running USEtox for comparison, *Pangea* could systematically run Mackay Level II/III models and run a specialized atmospheric model, such as AERMOD, with a basic parameterization (built by *Pangea* and suitable for screening).

I believe that in order to support decision-making in fast evaluative contexts, it is more reliable and resource-efficient to base assessments on the compatibility between the output of multiple models evaluated using multiple selections of data sets (source, version, year) than to focus on one model and selection of data sets in particular. From my point of view, *Pangea* could be a sound and instrumental platform for such integrated assessments.

Appendices

Appendix A – Example of source apportionment

This appendix illustrates the initial results of a source apportionment for the fresh water concentrations, for a grid cell (ID 16832) located directly downstream from Wuhan, China. It complements the EcoHOPE project presented in Chapter 3, using the tools for performing source apportionment as developed for the study presented in Chapter 5.

Figure A.1 shows the effective contributions [kg] of all cells to the mass at steady-state in cell 16832, for the actual spatial emissions of LAS defined by the project. As LAS is short-lived in fresh water ($DT50_{water}$ = 0.3 d), major contributors span over a short range upstream from cell 16832.

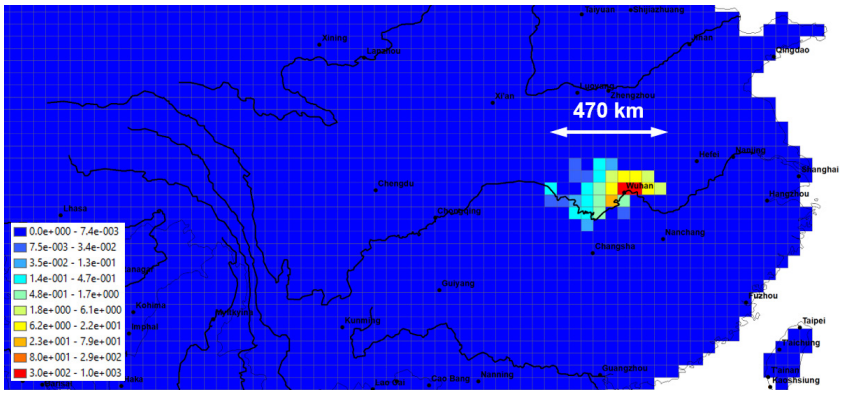


Figure A.1 – Contributions [kg] to the mass of LAS at steady-state in cell 16832 from all cells.

Figure A.2 shows the effective contributions [kg] of all cells to the mass at steady-state in cell 16832, for the actual spatial emissions of TCS defined by the project. As TCS is longer-lived in fresh water ($DT50_{water}$ = 23 d), major contributors span over a longer range upstream from cell 16832.

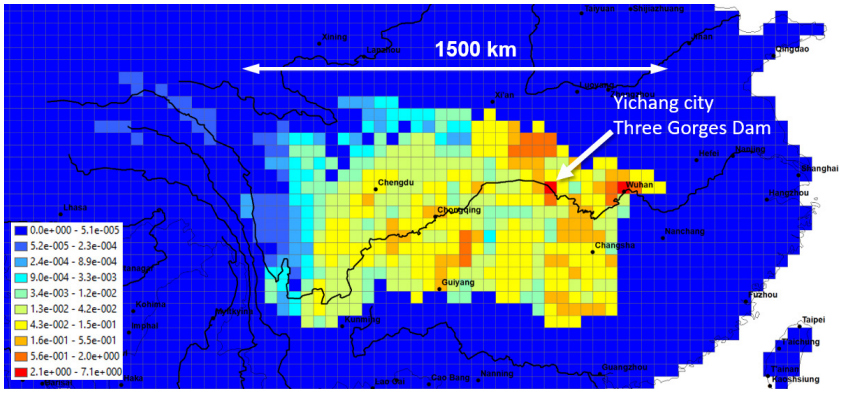


Figure A.2 – Contributions [kg] to the mass of TCS at steady-state in cell 16832 from all cells.

Appendix B – Model internals

B.1 History, main development steps, and technologies

Developments started in 2007 and the objective was to build a tool that would allow fast spatialization of generic compartmental models like IMPACT2002, on a per project basis (project-specific grids and geometry), avoiding the usual 6 months' full time effort involved in previous developments of fixed, low resolution spatial versions (e.g. IMPACT2002 Europe, IMPACT North America, IMPACT World). This required the model to incorporate a GIS engine, or to be built within a GIS. A few years earlier, ESRI had chosen the Python programming language as a scripting language for ArcGIS, and made the ArcGIS toolbox available as a Python class[53]. This made Python a relevant choice for developing *Pangea*, enabling us to use the ~900 tools from the ArcGIS toolbox. The Python language had many other advantages, one being that it was developed as a "glue language", well suited for merging (or "gluing") efficient tools and libraries together using a minimal amount of Python code, hence minimizing the time to solution (including the development phase). Python was also arising as language of choice for scientific computing, and many schools and universities offered Python classes for engineers. The major drawback was a lack of support for sparse matrices in main-stream numeric libraries[54]. MATLAB was therefore selected, initially as a temporary solution, for building *Pangea's* computation engine as an external module. ***Pangea* v1** was born: it had a Python core with projects management, data pre-processors, a GIS engine, an interface to a MATLAB core, and post-processors for the MATLAB output.

Pangea v1 was fulfilling our initial requirements: it took ~3 hours for building global geometries and projecting data, a few seconds per substance for computing steady-state solutions, and then a few minutes per substance for post-processing the output (e.g. saving to file). It had however several major drawbacks:

- Users had to understand ~1000 lines of Python code for setting up basic projects.
- Python could not call MATLAB for performing solely a few operations with sparse matrices; it had to build a complete and complex model in MATLAB[55].

[53] A class is an object-oriented concept; it can be seen as a library or as a toolbox.
[54] There were a few libraries developed as side projects with minimal support for sparse matrices, but none was stable, maintained, and covering the whole range of operations needed in *Pangea*.
[55] A large MATLAB core had therefore to be built; EPMs for example, which are involved in the construction of the sparse K matrix, had to be programmed in MATLAB.

- One of the few ways to achieve enough flexibility was to have Python build a MATLAB script based on XML templates of MATLAB blocks, and call it. This script was ~8000 lines long.
- GIS processing is rarely working in complex setups[56]. Failures in ArcGIS processing would often propagate and lead to crashes, further in the Python code, in the MATLAB code, or at the level of the interface.
- No other team collaborator could debug most projects during their initial development phase, where crashes can happen at all aforementioned levels, or could make the steps to learn Python/ArcGIS/MATLAB adequately for this purpose.
- Maintaining a system based on Python, MATLAB, XML templates, and possibly C/C++ later interfaced with Python and MATLAB for improving performance, is overly complex. It is not appropriate from a software engineering point of view beyond a very temporary solution (doomed to last longer than desirable, because there is rarely funding for model/software development/stabilization in academic projects).
- Finally, it was extremely difficult to perform all tests that arise in a context of prospective research, because implementing what could look like a "quick and easy test", would often require performing extra GIS operations (that could fail), and adding features and management procedures at all aforementioned levels.

All these reasons and the fact that MATLAB became the language of reference for scientific computing among our collaborators, lead us to developed *Pangea* v2 in parallel, with the intention to phase-out Python and ArcGIS. ***Pangea* v2** has the following main characteristics:

- Single MATLAB core, MATLAB EPMs, MATLAB EMs, etc.
- Hybrid GIS engine, cascading *Pangea* GIS[57], MATLAB Mapping, Python/ArcGIS[58], and possibly Quantum GIS later.
- Fully object-oriented design, which simplifies developments, maintenance, but also usage.

[56] GIS work is an Art, especially when dealing with large and complex data, in the sense that one generally must iteratively understand why a series of standard operations fails, and test alternative approaches. Failures can be crashes, but also inappropriate outputs, and there is often no algorithmic way to the detect the latter; it can only be detected by human eye, intelligence, and the perception that "something doesn't look right" (which cannot be programmed easily).
[57] We implemented many GIS operations directly in MATLAB, specifically for *Pangea* (e.g. zonal statistics on rectangular grids and refinement algorithms).
[58] Python/ArcGIS is an external module reduced to the minimum set of operations that really requires the full power of ArcGIS. This module is being progressively phased-out.

The hybrid GIS engine makes *Pangea* v2 able to reduce computation time for building geometries from ~3 hours to ~10 minutes for projects of average size[59]. At present, version 2 is powerful and flexible enough engine for tackling the UNILEVER project (chapter 3) and extending the initial few studies involved in the Europe solid waste treatment plant project (chapter 4) to 126 and finally 10,000 studies (to comply with reviewers' requirements).

[59] This time can still be reduced significantly, because Pangea GIS has not been parallelized yet.

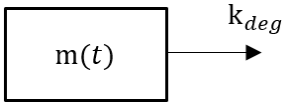
Appendix C – Linear Compartmental Systems

This section succinctly develops the theory needed to understand the construction of a mathematical compartmental system that describes the evolution of the mass/concentration of substances in the environmental, and to compute subsequent population exposure[60]. Compartmental systems have been already extensively analyzed (Jacquez and Simon, 1993, and references therein), and applied to environmental chemistry as multimedia mass-balance models (MacLeod et al., 2010). *Pangea* builds on the matrix formalism for modeling emissions to impacts presented in Rosenbaum et al. (2007), but implements it in a way that is better suited for large systems, exposed in section 2.7.

We start with a simple environmental compartment with a degradation and no source, and we progressively extend the system and the formalism until we arrive at the matrix formalism for a system with *n* compartments, relevant to *Pangea*.

C.1 Single compartment with no source, degradation rate constant

The simplest, non-trivial system is a single compartment with no source, and a single elimination process parameterized by a degradation rate constant k_{deg}:



We are interested in $\mathrm{m}(t)$ [kg], a function of the time t [s] which describes the evolution with time of the mass of e.g. a pollutant inside the compartment. The underlying assumption for the degradation, is that the quantity degraded at time t is proportional to the mass in the compartment at this time, $\mathrm{m}(t)$, with a factor of proportionality defined by k_{deg}. This defines the unit of k_{deg}:

$$\mathrm{k}_{deg}\ \left[\frac{\mathrm{kg}_{degraded}\ /\ \mathrm{s}}{\mathrm{kg}_{in\ compartment}}\right] \qquad \text{also noted} \qquad \mathrm{k}_{deg}\ \left[\frac{1}{\mathrm{s}}\right] \qquad \text{or} \qquad \mathrm{k}_{deg}\ [\mathrm{s}^{-1}]$$

This allows to write a mass-balance equation:

$$\frac{d\mathrm{m}(t)}{dt} = \underbrace{0}_{\text{input}} - \underbrace{\mathrm{k}_{deg}\ \mathrm{m}(t)}_{\text{output}} = -\mathrm{k}_{deg}\ \mathrm{m}(t) \tag{C.1}$$

We solve this equation by separation of variables and integration:

[60] We present it from a "rate coefficients point of view", but there is another formulation based on fugacity.

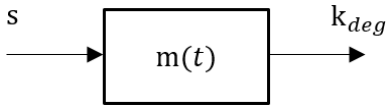
$$\frac{d\mathrm{m}(t)}{dt} = -\mathrm{k}_{deg}\,\mathrm{m}(t) \;\Rightarrow\; \frac{d\mathrm{m}(t)}{\mathrm{m}(t)} = -\mathrm{k}_{deg}\,dt \;\Rightarrow\; \int_{\mathrm{m}_0}^{\mathrm{m}_t} \frac{d\mu}{\mu} = -\mathrm{k}_{deg}\int_{t_0}^{t} d\tau$$

$$\Rightarrow\; \log(\mu)\big|_{m_0}^{m_t} = -\mathrm{k}_{deg}\,\tau\,\big|_{t_0}^{t} \;\Rightarrow\; \log(\mathrm{m}_t) - \log(\mathrm{m}_0) = \log\left(\frac{\mathrm{m}_t}{\mathrm{m}_0}\right) = -\mathrm{k}_{deg}(t - t_0) \tag{C.2}$$

$$\Rightarrow\; \frac{\mathrm{m}_t}{\mathrm{m}_0} = e^{-\mathrm{k}_{deg}(t-t_0)} \;\Rightarrow\; \mathrm{m}_t = \mathrm{m}_0\, e^{-\mathrm{k}_{deg}(t-t_0)} = \mathrm{m}(t)$$

C.2 Single compartment with a source

The case with a source term s [kg/s] transforms Eq. C.2 into an inhomogeneous equation:



$$\frac{d\mathrm{m}(t)}{dt} = \underbrace{\mathrm{s}}_{\text{input}} - \underbrace{\mathrm{k}_{deg}\,\mathrm{m}(t)}_{\text{output}} \tag{C.3}$$

We start by defining the **steady-state solution** of Eq. C.3, which is useful for expressing the solution of the dynamics later. When it exists, the steady-state is defined by the fact that all derivatives (with respect to time in this context) are null[61]. This provides a way to compute the steady-state without having to compute the limit of the dynamic solution $\mathrm{m}(t)$:

$$\frac{d\mathrm{m}_{ss}}{dt} = 0 = \mathrm{s} - \mathrm{k}_{deg}\,\mathrm{m}_{ss}$$

where m_{ss} is the mass at steady-state. Solving for m_{ss}, we get:

$$\mathrm{m}_{ss} = \frac{\mathrm{s}}{\mathrm{k}_{deg}} = \mathrm{k}_{deg}^{-1}\,\mathrm{s} \tag{C.4}$$

The steady-state solution is the constant "state" towards which the system will stabilize over long times.

For solving the dynamics, we perform a change of variable (translation) that brings us back to the previous case:

$$\tilde{\mathrm{m}}(t) = \mathrm{m}(t) - \mathrm{m}_{ss} \;\Rightarrow\; \mathrm{m}(t) = \tilde{\mathrm{m}}(t) + \mathrm{m}_{ss} \tag{C.5}$$

[61] Steady-state is not equivalent to *equilibrium*. Steady-state means "unchanging with time", whereas equilibrium implies that phases have concentrations (or pressure, temperature, etc) such that they experience no tendency for net transfer of mass. An illustration can be found in *Multimedia Environmental Models, The Fugacity Approach*, 2nd Edition, by Donald Mackay, on page 23.

Replacing in Eq. C.3, we get:

$$\frac{d\mathrm{m}(t)}{dt} = \frac{d(\tilde{\mathrm{m}}(t) + \mathrm{m}_{ss})}{dt} = \frac{d\tilde{\mathrm{m}}(t)}{dt} = -\mathrm{k}_{deg}\left(\tilde{\mathrm{m}}(t) + \mathrm{k}_{deg}^{-1}\,\mathrm{s}\right) + \mathrm{s} = -\mathrm{k}_{deg}\,\tilde{\mathrm{m}}(t) - \mathrm{s} + \mathrm{s}$$

and finally:

$$\frac{d\tilde{\mathrm{m}}(t)}{dt} = -\mathrm{k}_{deg}\,\tilde{\mathrm{m}}(t)$$

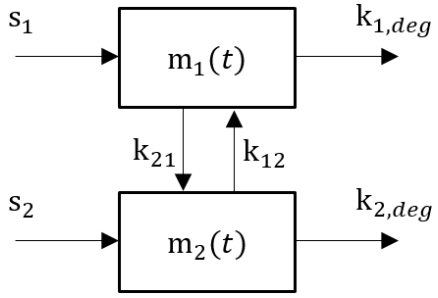
that we can solve as Eq. C.2, before performing the inverse change of variable of the solution:

$$\begin{aligned} \tilde{\mathrm{m}}(t) = \tilde{\mathrm{m}}_0\, e^{-\mathrm{k}_{deg}\,t} = \mathrm{m}(t) - \mathrm{m}_{ss} = (\mathrm{m}_0 - \mathrm{m}_{ss})\, e^{-\mathrm{k}_{deg}\,t} \\ \mathrm{m}(t) = (\mathrm{m}_0 - \mathrm{m}_{ss})e^{-\mathrm{k}_{deg}\,t} + \mathrm{m}_{ss} \quad \text{or} \quad \mathrm{m}(t) = \underbrace{(\mathrm{m}_0 - \mathrm{m}_{ss})}_{\text{offset}}\, e^{-\mathrm{k}_{deg}\,t} + \underbrace{\mathrm{k}_{deg}^{-1}\,\mathrm{s}}_{\mathrm{m}_{ss}} \end{aligned} \tag{C.6}$$

This shows that $\mathrm{m}(t)$ is the sum of two contributions: the mass at steady-state, m_{ss}, and an offset which is the difference between the initial condition, m_0, and the mass at steady-state. This offset is multiplied by an exponential function which is 1 at $t = 0$, and tends towards 0 for $t \to \infty$.

C.3 System with multiple compartments

The figure below illustrates a case with two compartments:



In this case, we have one mass-balance equation per compartment, which form a system, and these equations are coupled (the derivative of the mass in a compartment depends on the mass in the other compartment, which is the solution of the other equation):

$$\begin{cases} \dfrac{d\mathrm{m}_1(t)}{dt} = \left(-\mathrm{k}_{1,deg} - \mathrm{k}_{21}\right)\mathrm{m}_1(t) + \mathrm{k}_{12}\mathrm{m}_2(t) + \mathrm{s}_1 \\ \dfrac{d\mathrm{m}_2(t)}{dt} = \mathrm{k}_{21}m_1(t) + \left(-\mathrm{k}_{2,deg} - \mathrm{k}_{12}\right)\mathrm{m}_2(t) + \mathrm{s}_2 \end{cases} \tag{C.7}$$

Eq. C.7 is comparable to the differential equations that we had in the one compartment case, Eq. C.2, with additional terms which represent inputs from and outputs to all other compartments because of the coupling. For an n-compartment system, the expression for the *i*-th compartment is:

$$\frac{d\mathrm{m}_i(t)}{dt} = \left(-\mathrm{k}_{i,deg} - \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{k}_{ji} \right) \mathrm{m}_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{k}_{ij}\, \mathrm{m}_j(t) + \mathrm{s}_i \tag{C.8}$$

Defining:

$$\mathrm{k}_{ii} = -\mathrm{k}_{i,deg} - \sum_{\substack{j=1 \\ j\neq i}}^{n} \mathrm{k}_{ji} \tag{C.9}$$

Eq. C.8 can be simply written:

$$\frac{d\mathrm{m}_i(t)}{dt} = \sum_{j=1}^{n} \mathrm{k}_{ij}\, \mathrm{m}_j(t) + \mathrm{s}_i \tag{C.10}$$

Solving this coupled system requires more than a basic change of variable, and we can benefit switching to a matrix formulation.

C.4 Matrix formulation of system with multiple compartments

Eq. C.10 can be expressed in a matrix form:

$$\frac{d\vec{\mathrm{m}}(t)}{dt} = \mathbf{K}\, \vec{\mathrm{m}}(t) + \vec{\mathrm{s}} \tag{C.11}$$

with

$$\vec{\mathrm{s}} = \begin{pmatrix} \mathrm{s}_1 \\ \vdots \\ \mathrm{s}_n \end{pmatrix}, \text{ a constant source vector}$$

$$\vec{\mathrm{m}}(t) = \begin{pmatrix} \mathrm{m}_1(t) \\ \vdots \\ \mathrm{m}_n(t) \end{pmatrix}, \text{ a time varying mass vector}$$

$$\mathbf{K} = \left(\mathrm{k}_{ij}\right), \text{ a matrix of rate coefficients, where diagonal elements are defined by C.9}$$

Non-diagonal terms are positive or null and represent inputs into compartments. Element k_{23}, for example, is the transfer coefficient from compartment 3 to compartment 2; it appears in Eq. C.11

multiplied by $m_3(t)$ as input to compartment 2. Diagonal elements are negative[62] and represent total elimination, as a sum of degradation and outputs to all other compartments. Element k_{44}, for example, is minus the sum of the degradation in compartment 4 and transfer coefficients from compartment 4 to all other compartments. It appears in Eq. C.11 multiplied by $m_4(t)$, which defines the total output/elimination from compartment 4.

C.5 Steady-state of multi-compartment system, matrix of fate factors

We compute the steady-state by imposing a null derivative in Eq. C.11:

$$\frac{d\vec{m}(t)}{dt} = \vec{0} = \mathbf{K}\,\vec{m}(t) + \vec{s}$$

where $\vec{m}_{ss}$ is the vector of masses at steady-state. Solving for $\vec{m}_{ss}$, we get:

$$\mathbf{K}\,\vec{m}_{ss} + \vec{s} = \vec{0} \quad \Rightarrow \quad \mathbf{K}\,\vec{m}_{ss} = -\vec{s} \quad \Rightarrow \quad \mathbf{K}^{-1}\mathbf{K}\,\vec{m}_{ss} = -\mathbf{K}^{-1}\,\vec{s}$$

and finally:

$$\vec{m}_{ss} = -\mathbf{K}^{-1}\,\vec{s} \quad \text{sometimes noted} \quad \vec{m}_{ss} = \mathbf{FF}\,\vec{s} \quad \text{with} \quad \mathbf{FF} = -\mathbf{K}^{-1} \qquad \text{(C.12)}$$

We see in particular that the steady-state of the homogeneous system (characterized by $\vec{s} = \vec{0}$) is $\vec{0}$. The **FF** matrix is the matrix of *fate factors*, which "scales and distributes" components of the source vector into components of the mass at steady-state vector. Its unit is technically [$kg_{at\ steady-state}$ per $kg_{emitted}$/s], which is often simplified to [s]. There is a whole body of "theory" about interpreting elements of the **K** and the **FF** matrices, that pertains to small systems where **K** can be inverted and elements of **FF** analyzed individually. For larger systems like in *Pangea*, however, it is not possible to invert **K** and therefore not possible to obtain **FF**; yet, the linear system:

$$\mathbf{K}\,\vec{m}_{ss} = -\vec{s}$$

can be solved numerically for $\vec{m}_{ss}$ (section 2.10.3).

[62] They must even be strictly negative, as a zero on the diagonal means no elimination and would lead to a solution that diverges.

C.6 Dynamics of multi-compartment system

The dynamic solution of Eq. C.11 can be computed analytically, through a change of basis towards a basis of eigen vectors of **K**. I develop it here for reference, with minimum detail, because is currently not implemented in *Pangea* in a fashion that makes it more interesting than using a numeric solver. Computing **V** and **Λ**, the matrices of eigen vectors (written in column) and eigen values (diagonal) of **K** respectively, we can perform of change of basis on both sides of Eq. C.11:

$$\mathbf{V}^{-1}\ \frac{d}{dt}\vec{\mathrm{m}}(t) = \mathbf{V}^{-1}\ (\ \mathbf{K}\,\vec{\mathrm{m}}(t) + \ \vec{\mathrm{s}}\ )$$

where $\mathbf{V}^{-1}$ can be moved into the derivative in the left-hand side (as it is constant), and distributed into the parenthesis on the right-hand side. We insert $\mathbf{I} = \mathbf{V}\mathbf{V}^{-1}$ make changes of basis appear:

$$\frac{d}{dt}\left(\mathbf{V}^{-1}\,\vec{\mathrm{m}}(t)\right) = \mathbf{V}^{-1}\,\mathbf{K}\,\vec{\mathrm{m}}(t) + \mathbf{V}^{-1}\,\vec{\mathrm{s}} = \mathbf{V}^{-1}\,\mathbf{K}\,\underbrace{\mathbf{V}\,\mathbf{V}^{-1}}_{\mathbf{I}}\,\vec{\mathrm{m}}(t) + \mathbf{V}^{-1}\,\vec{\mathrm{s}}$$

$$\frac{d}{dt}\underbrace{(\ \mathbf{V}^{-1}\,\vec{\mathrm{m}}(t)\ )}_{\vec{\tilde{\mathrm{m}}}(t)} = \underbrace{\mathbf{V}^{-1}\,\mathbf{K}\,\mathbf{V}}_{\tilde{\mathbf{K}}}\ \underbrace{\mathbf{V}^{-1}\,\vec{\mathrm{m}}(t)}_{\vec{\tilde{\mathrm{m}}}(t)} + \underbrace{\mathbf{V}^{-1}\,\vec{\mathrm{s}}}_{\vec{\tilde{\mathrm{s}}}(t)}$$

$$\frac{d\vec{\tilde{\mathrm{m}}}(t)}{dt} = \tilde{\mathbf{K}}\,\vec{\tilde{\mathrm{m}}}(t) + \ \vec{\tilde{\mathrm{s}}} \qquad \text{(C.13)}$$

where $\tilde{\mathbf{K}}$ is another notation for **Λ** (matrix of eigen values λ_i of **K**):

$$\tilde{\mathbf{K}} = \mathbf{V}^{-1}\mathbf{K}\,\mathbf{V} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = \mathbf{\Lambda}$$

While Eq. C.13 looks identical to Eq. C.11, it is not. The change of basis transform the full **K** matrix into $\tilde{\mathbf{K}}$, which is diagonal. $\tilde{\mathbf{K}}$ hence describes a system of uncoupled equations, that can be solved independently of each other, like Eq. C.3. The solution is:

$$\vec{\tilde{\mathrm{m}}}(t) = e^{\tilde{\mathbf{K}}\,t}\left(\vec{\tilde{\mathrm{m}}}_0 + \tilde{\mathbf{K}}^{-1}\,\vec{\tilde{\mathrm{s}}}\right) - \tilde{\mathbf{K}}^{-1}\,\vec{\tilde{\mathrm{s}}} = e^{\tilde{\mathbf{K}}\,t}\left(\vec{\tilde{\mathrm{m}}}_0 - \vec{\tilde{\mathrm{m}}}_{ss}\right) + \vec{\tilde{\mathrm{m}}}_{ss}$$

which can be back-transformed by inverse change of bases:

$$\vec{\mathrm{m}}(t) = \mathbf{V}\,\vec{\tilde{\mathrm{m}}}(t) = \ \mathrm{V}\left(e^{\tilde{\mathbf{K}}\,t}\left(\vec{\tilde{\mathrm{m}}}_0 - \vec{\tilde{\mathrm{m}}}_{ss}\right) + \vec{\tilde{\mathrm{m}}}_{ss}\right)$$

$$= \ \mathbf{V}\,e^{\tilde{\mathbf{K}}\,t}\,\underbrace{\mathbf{V}^{-1}\mathbf{V}}_{\mathbf{I}}\left(\vec{\tilde{\mathrm{m}}}_0 - \vec{\tilde{\mathrm{m}}}_{ss}\right) + \mathbf{V}\,\vec{\tilde{\mathrm{m}}}_{ss} = \mathbf{V}\,e^{\tilde{\mathbf{K}}\,t}\,\mathbf{V}^{-1}(\underbrace{\mathbf{V}\,\vec{\tilde{\mathrm{m}}}_0}_{\vec{\mathrm{m}}_0} - \underbrace{\mathbf{V}\,\vec{\tilde{\mathrm{m}}}_{ss}}_{\vec{\mathrm{m}}_{ss}}) + \underbrace{\mathbf{V}\,\vec{\tilde{\mathrm{m}}}_{ss}}_{\vec{\mathrm{m}}_{ss}}$$

And thus:

$$\vec{\mathrm{m}}(t) = \mathbf{V}\,e^{\tilde{\mathbf{K}}\,t}\,\mathbf{V}^{-1}(\vec{\mathrm{m}}_0 + \mathbf{K}^{-1}\,\vec{\mathrm{s}}) - \mathbf{K}^{-1}\,\vec{\mathrm{s}} = \mathbf{V}\,e^{\tilde{\mathbf{K}}\,t}\,\mathbf{V}^{-1}(\vec{\mathrm{m}}_0 - \vec{\mathrm{m}}_{ss}) + \vec{\mathrm{m}}_{ss} \qquad \text{(C.14)}$$

While this solution is simple and elegant, it cannot be used directly, because we cannot compute eigen values and vectors of **K** matrices as large those involved in *Pangea*. Even if we could, **V** would be dense and it could not be stored in memory[63]. However, Eq. C.14 can be developed and expressed as a sum:

$$\vec{\mathrm{m}}(t) = \mathbf{V}e^{\tilde{\mathbf{K}}\,t}\,\vec{\mathrm{c}} + \vec{\mathrm{m}}_{ss} = \sum_{j=1}^{n} \mathrm{c}_j e^{\lambda_j\,t}\,\vec{\mathrm{v}}_j + \vec{\mathrm{m}}_{ss} \tag{C.15}$$

with $\vec{\mathrm{c}} = \mathbf{V}^{-1}(\vec{\mathrm{m}}_0 - \vec{\mathrm{m}}_{ss}) = \begin{pmatrix} \mathrm{c}_1 \\ \vdots \\ \mathrm{c}_n \end{pmatrix}$.

Pangea currently implements a numeric ODE solver by part for solving the dynamics, and the focus of my thesis was not on dynamic solutions. An extension of *Pangea* is, that will investigate the possibility to use and iterative eigenvalue algorithm for evaluating a small set of relevant *eigen* pairs, that would allow us to approximate the sum in Eq. C.15. However, while begin slow, using a numeric ODE solver has the advantage that it allows us to address problems with non-constant coefficients $\vec{\mathrm{s}}(t)$ and/or $\mathbf{K}(t)$:

$$\frac{d\vec{\mathrm{m}}(t)}{dt} = \mathbf{K}(t)\,\vec{\mathrm{m}}(t) + \vec{\mathrm{s}}(t) \tag{C.16}$$

[63] The minimal size of matrices in Pangea lies around 70,000×70,000. Storing a single dense matrix of this size as double precision (8 bytes) floating point numbers would take ~40GB of RAM.

Appendix D – Supporting information

D.1 A Global Framework to Model Spatial Ecosystems Exposure to Home and Personal Care Chemicals in Asia

D.1.1 Calculation of emission inventories

D.1.1.1 Market consumption

Triclosan – To derive the total market tonnage for the selected countries the Global New Product Database (GNPD) (Mintel, 2016) and Euromonitor sales data (Euromonitor, 2016) was used. Market penetration (e.g. TCS is 11% of all liquid soap on the market) was first calculated from the GNPD (Mintel, 2016). This was done by selecting all product categories containing TCS. Categories were either included or excluded based on their applicability for down the drain environmental risk assessment (e.g. Bar soap is included, Diapers & Feminine Hygiene Wipes excluded). For each of the TCS categories included, the total number of variants globally were compared to the number of variants using TCS to determine the fraction of TCS entering the global market (i.e. market penetration). This approach assumes that the fraction of TCS entering the market will be the same in all selected countries.

The inclusion level of TCS was based on a typical level of 0.3% in HPC products (EC, 2007; Zhu et al., 2016) for Europe. It is acknowledged inclusion levels will vary geographically, however, it is assumed inclusion levels will be equivalent for each product type across Asia. Zhu et al. (2016) used the same value to calculate TCS usage in China across all categories. Euromonitor 2013 sales data for each category and country were downloaded. By multiplying the "Product sold" (Euromonitor), "Inclusion level" and "market penetration" fraction (Mintel), the total tonnage for each category can be calculated (Eq. D.1.1):

$$T = S \times M \times I \tag{D.1.1}$$

Where:

S is the Euromonitor sale volume

M is the GNPD market penetration of ingredients in products

I is the ingredient inclusion level in product

T is the total tonnage

By adding the total tonnage for each category together, a final tonnage estimate can be made for each country.

Methyl paraben – Refer to the GNPD/Euromonitor method used for TCS. Only Personal care products were used in this method as home care inputs are negligible. Inclusion levels were derived from Cosmetics Ingredients Review (Andersen, 2008). Where a range was provided the mean of the two values was used. It is acknowledged inclusion levels will vary geographically, however, it is assumed inclusion levels will be equivalent for each product type.

Decamethylcyclopentasiloxane (D5) – Refer to the GNPD/Euromonitor method used for TCS. Only Personal care products were used in this method as home care inputs are negligible. Median inclusion levels from Personal Care Products Council (PCPC, 2017) were used.

Linear alkylbenzene sulfonate (LAS) – To derive the total market tonnage, it was assumed that LAS is present in all detergent formats (i.e. bars, powders, liquids, hand wash, automatic wash, standard and concentrated). Category sales data from Euromonitor were extracted for each product type. An average inclusion level of 13% (HERA, 2013) was applied to detergent products and applied to the Euromonitor sales data to determine a LAS tonnage for each country. The tonnages calculated are in the same order of magnitude as per capita literature values (Whelan et al., 2012) when extrapolated using population data.

Table D.1.1 - Inclusion levels of chemicals in different categories of products.

Chemicals	Main Categories	Sub-categories	Inclusion level
TCS	Oral Hygiene Soap & Bath Products Dishwashing Products Fabric Care Skincare Hair Products Diapers & Feminine Hygiene Deodorants Hard Surface Care Toilet Care Fragrances Shaving & Depilatories		0.30%(1)
LAS	Fabric care		13%(2)
Methyl Paraben	Deodorants		0.15%(3)
	Hair products	Conditioner	0.25%(3)

		Hair Styling	0.22% (3)
		Hair Treatments	0.22% (3)
		Shampoo	0.25% (3)
	Shaving & Depilatories	Depilatory Products	0.25% (3)
		Shaving Preparations	0.21% (3)
	Skincare	Body Care	0.275%(3)
		Face/Neck Care	0.32% (3)
		Hand/Nail Care	0.275% (3)
		Lip Care	0.32% (3)
		Sun - After Sun	0.275 %(3)
		Sun - Sun/Sunbed Exposure	0.275% (3)
	Soap & Bath Products	Bar Soap	0.2%(3)
		Liquid Soap	0.2%(3)
		Shower Products	0.2% (3)
		Bath Additives	0.4% (3)
D5	Deodorants		40.5% (4)
	Hair products	Conditioner	44.5% (4)
		Hair Styling	0.3% (4)
		Hair Treatments	46%(4)
		Shampoo	0.02%(4)
	Skincare	Body Care	44%(4)
		Face/Neck Care	5.9%(4)
		Hand/Nail Care	44% (4)
		Sun - After Sun	12.5%(4)
		Sun - Sun/Sunbed Exposure	25% (4)
	Soap & Bath Products	Liquid Soap	7.00%(4)
		Shower Products	7.00%(4)

Notes: (1) value of inclusion level reported by EC (2007)
(2) value of inclusion level reported by HERA (2013)
(3) mean value of inclusion levels reported by Andersen (2008)
(4) mean value of inclusion levels reported by PCPC (2016)

D.1.1.2 Discharge pathways scenarios

ScenAT and SimpleTreat were used to calculate the total concentration entering the water, soil and air compartment for input into *Pangea*. Figure S1 is a conceptual model of the discharge pathway scenarios from consumer use to the environment. This was done for each administrative unit in ScenAT, which was then resampled for input into *Pangea's* WWDRII 0.5°×0.5° native grid.

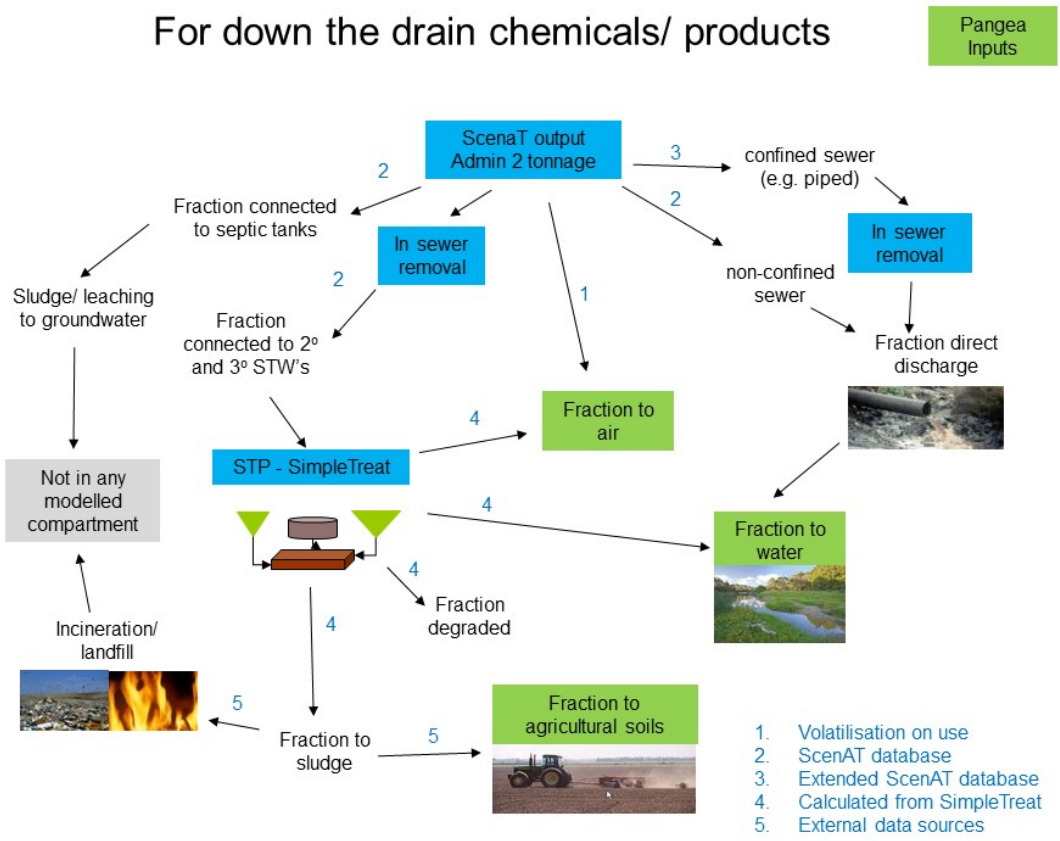


Figure D.1.1 - Discharge pathways scenarios for input into *Pangea*.

For the aquatic compartment, the $PEC_{effluent}$ needs to be calculated. The method used to calculate the $PEC_{effluent}$ corresponds to the local scale surface water PEC following the European Technical Guidance Document (EC, 2007) (Eq. 2). Dilution is not included as *Pangea* accounts for this parameter using WWDRII global gridded hydrological model.

$$\mathrm{PEC_{effluent}}\ (mg\ L^{-1}) = (1 - F)\frac{T \cdot 10^{9}}{W \cdot 365 \cdot P} \qquad \text{(D.1.2)}$$

Where

F is the fraction removed in STP

T is the annual emissions of chemical ingredient (tonnes, 1000kg)

W is the per capita water use (L/day/person)

P is the population of the county

See tables D.1.6 to D.1.8 with details of input parameters (i.e. population, water use and STP connectivity) for equation 1.

The air compartment input is derived from volatilization on consumer use (i.e. D.1.5) and volatilization during sewage treatment (i.e. SimpleTreat) (Table D.1.2).

Table D.1.2 - Primary removal and discharge fractions for the selected HPC chemical.

	Fraction to air on use	**In sewer removal fraction**	**SimpleTreat fraction to FW**	**SimpleTreat fraction to sludge**	**SimpleTreat fraction to air**
TCS	0	0	0.054	0.5	0
Methyl Paraben	0	0	0.1245	0.00092	0.003
LAS	0	0.5[(2)]	0.011	0.206	0
D5	0.9[(1)]	0	0.05	0.73	0.22

[(1)] MacKay et al., 2015
[(2)] HERA et al., 2013

The soil compartment input into *Pangea* is derived from SimpleTreat (Table D.1.2) and sludge application to soil practices (Table D.1.3). There are many sludge disposal mechanisms such as incineration, to landfill, to agricultural land or dumped directly into the ocean. This type of data is difficult to obtain as many countries in Asia do not have specific statistics on these mechanisms. We therefore took an approach of providing a qualitative value (Low: 0.1, Medium: 0.4, High: 0.7) for each country and where actual values are available these are used. Where no value is available for a country read across from a neighboring country is used (Table D.1.3).

Table D.1.3 – Fraction of sludge to agricultural soil.

Countries	Fraction to ag soil	Qualitative	Reference
Bangladesh	0.7	High	Read across from India
Cambodia	0.1	Low	Read across from Vietnam
China	0.45*	-	iwaterwiki (2016)
India	0.7	High	China Water Risk (2016)
Indonesia	0	Low	Read across from Malaysia
Japan	0.1	Low	China Water Risk (2016)
Laos	0.1	Low	Thailand Read across
Malyasia	0	Low	ISWA (2016)
Pakistan	0.7	High	Read across from India
Phillipines	0.1	Low	Read across from Vietnam
Sri Lanka	0.7	High	Read across from India
Thailand	0	Low	Sreesai et al. (2013)
Vietnam	0.1	Low	Spinosa et al. (2007)
South Korea	0.4	Medium	China Water Risk (2016)

*Actual value

D.1.2 Input parameters

D.1.2.1 Physico-chemical properties

Physico-chemical properties are used as input into SimpleTreat, to determine fate in sewage treatment plants, and *Pangea* to determine fate in the environment.

Table D.1.4 - Main fate and transport chemical properties and parameters for the four substances under study.

		LAS	**TCS**	**Methyl Paraben**	**D5**
		68411-30-3	3380-34-5	99-76-3	541-02-6
Kaw	[-]	1.5×10^{-6} (2)	2.1×10^{-7} (1)	1.5×10^{-7} (2)	1.4×10^{3} (1)
Kow	[-]	2.5×10^{1} (1)	7.9×10^{4} (1)	4.6×10^{1} (1)	1.1×10^{8} (1)
Koc	[L kg^{-1}]	2.5×10^{3} (3)	2.3×10^{4} (2)	8.6×10^{1} (2)	1.5×10^{5} (1)
$DT50_{air}$	[d]	3.3×10^{-1} (1)	6.6×10^{-1} (2)	1.5×10^{1} (4)	4.5×10^{-1} (1)
$DT50_{water}$	[d]	2.9×10^{-1} (5)	2.3×10^{1} (6)	8.3×10^{-1} (7)	6.0×10^{1} (8)
$DT50_{sediments}$	[d]	7.0×10^{1} (3)	2.3×10^{2} (9)	8.3×10^{0} (9)	9.7×10^{1} (10)
$DT50_{soil}$	[d]	7.0×10^{0} (3)	2.3×10^{1} (11)	8.3×10^{-1} (11)	9.7×10^{0} (8)

(1) ECHA database
(2) EPIsuite modeled
(3) HERA 2013
(4) Toxnet
(5) Whelan et al. 2007
(6) Singer et al. 2002
(7) Harman et al. 2015
(8) UK Environment Agency 2009
(9) $DT50_{sediments} = DT50_{water} \times 10$ according to TGD in EC, 2003
(10) $DT50_{sediments} = DT50_{soil} \times 10$ assuming 10% of sediments are aerobic
(11) $DT50_{soil} = DT50_{water}$ assuming rate is representative of aerobic medium

Spatial datasets of administrative boundaries for each country were compiled and organized for all countries, as summarized in Table D.1.5 below. The spatial unit of analysis covers three levels: Admin-0 (country), Admin-1 (state) and Admin-2 (county/district) for each country.

Table D.1.5 – Spatial administrative boundary data by country.

Unilever Market Cluster	Country	Date Acquired	Source Data Reference
North Asia	China	September 2013	China Data Centre, University of Michigan
North Asia	Japan	February 2013	GADM
North Asia	South Korea	February 2013	GADM
South Asia	Bangladesh	January 2011	GADM
South Asia	India	September 2010	Geographic Enterprises, LLC
South Asia	Pakistan	July 2011	DEMOBASE, United States Census Bureau
South Asia	Sri Lanka	January 2011	GADM
South East Asia	Cambodia	January 2011	GADM
South East Asia	Indonesia	January 2011	GADM
South East Asia	Laos	January 2011	GADM
South East Asia	Malaysia	January 2011	GADM
South East Asia	Philippines	January 2011	GADM
South East Asia	Thailand	January 2011	GADM
South East Asia	Vietnam	January 2011	GADM

Table D.1.6 – Population data obtained for each country.

Country	Extent Available	Census Data Year Released	Date Data Acquired	Source Data Reference
China	Admin-2	2000	September 2010	National Bureau of Statistics, China
India	Admin-2	2011	May 2011	Office of The Registrar General and Census Commissioner, India
Bangladesh	Admin-2	2001	July 2011	Bangladesh Bureau of Statistics
Cambodia	Admin-1	2008	July 2011	National Institute of Statistics, Cambodia
Indonesia	Admin-2	2008	July 2011	Statistics Indonesia
Laos	Admin-1	2005	July 2011	Laos Statistics Bureau
Malaysia	Admin-2	2000	July 2011	Department of Statistics, Malaysia
Pakistan	Admin-2	1998	July 2011	DEMOBASE, US Census Bureau
Philippines	Admin-2	2007	July 2011	National Statistics Office, Philippines
Sri Lanka	Admin-2	2001	July 2011	Department of Census and Statistics, Sri Lanka
Thailand	Admin-2	2000	July 2011	National Statistical Office , Thailand
Vietnam	Admin-2	2009	July 2011	General Statistics Office of Vietnam
Japan	Admin-2	2010	February 2011	Statistics Bureau, Japan
South Korea	Admin-2	2010	February 2011	Statistics Korea, Republic of South Korea

Table D.1.7 – Domestic water use data obtained by country.

Country	Extent Available	Year of Data	Date Data Acquired	Source Data Reference
China	Admin-1	2008	May 2011	China Statistical Yearbook (2009)
India	Admin-0	2005	June 2011	Stockholm Environment Institute (2011)
Bangladesh	Admin-0	1996	August 2011	Peter Gleick, Basic Water Requirements for Human Activities (1996)
Cambodia	Admin-0	1996	August 2011	Peter Gleick, Basic Water Requirements for Human Activities (1996)
Indonesia	Admin-0	1996	August 2011	Peter Gleick, Basic Water Requirements for Human Activities (1996)
Laos	Admin-0	2000	August 2011	World Resources (2000-2001)
Malaysia	Admin-0	2000	August 2011	World Resources (2000-2001)
Pakistan	Admin-0	2000	August 2011	World Resources (2000-2001)
Philippines	Admin-0	2000	August 2011	World Resources (2000-2001)
Sri Lanka	Admin-0	2000	August 2011	Peter Gleick, The World's Water (2000)
Thailand	Admin-0	2007	August 2011	Hoekstra and Chapagain (2008)
Vietnam	Admin-0	2000	August 2011	Peter Gleick, The World's Water (2000)
Japan	Admin-0	2008	April 2013	Japan Waste Water Association (2009)
South Korea	Admin-0	2005	April 2013	Ministry of Environment, South Korea (2009)

Table D.1.8 – Sewage Treatment Plant connectivity data obtained by country.

Country	Extent Available	Year of Data	Date Data Acquired	Source Data Reference (With STP Treatment)
China	Admin-1	2008	May 2011	China Statistical Yearbook (2009)
India	Admin-1	2005	June 2011	Central Pollution Control Board, India (2008)
Japan	Admin-1	2003	August 2011	UN Stats (2009)
South Korea	Admin-1	2010	August 2011	OECD Stats (2011)

Indonesia also has information on Septic tank connectivity (World Bank, 2008), which was accounted for in the discharge scenario pathway for input into *Pangea*.

D.1.3 *Pangea*: refinement potential and grids

The refinement potential is based on the following components: region of interest (extended Asia), population density, proximity to main stream (necessary because we need refined cells for comparing with monitoring data, which are located along main streams).

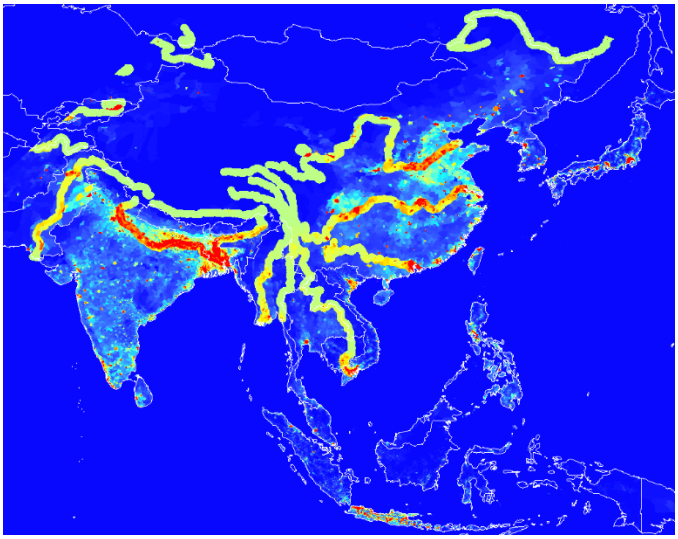


Figure D.1.2 – Refinement potential (RP).

The iterative refinement procedure outputs the following *Results* grid, which is also the first layer of the atmospheric grid. Upper layers are obtained by stacking intermediary refinement steps, in reversed order.

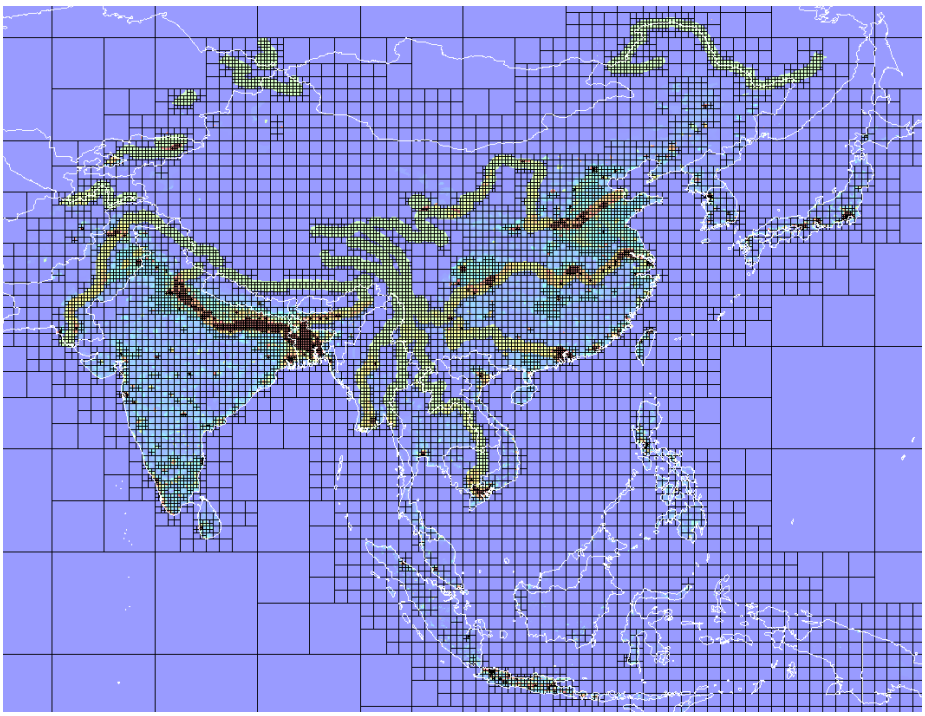


Figure D.1.3 – First layer of the atmospheric grid.

The terrestrial grid is obtained by clustering the WWDRII 0.5°x0.5° native grid. Clusters are finally disaggregated over the region of interest, to enforce maximum/native resolution.

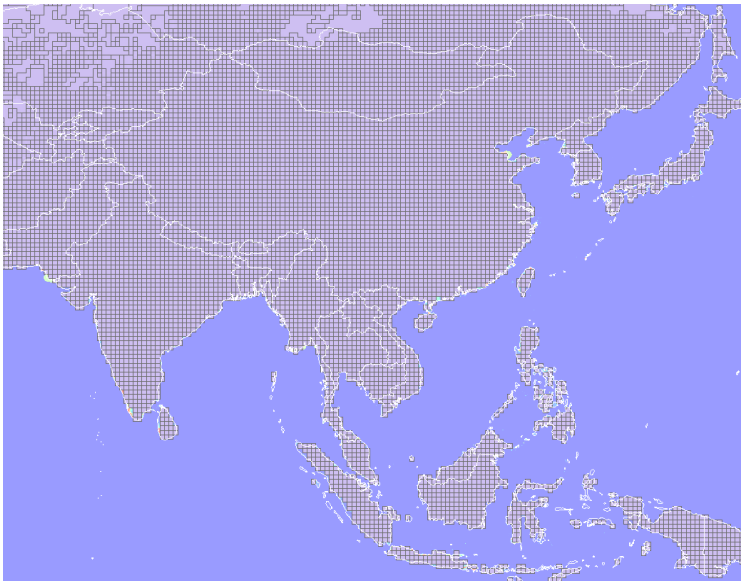


Figure D.1.4 – Terrestrial grid of clusters, un-clustered over the region of interest to enforce maximum/native resolution (of the hydrological model).

D.1.4 Monitoring: data collection

Table D.1.9 – Summary of monitoring data used for model verification for freshwater.

Freshwater	Country	No of stations	Regime	Author
Triclosan	India (Cauvery, Tamirapani River)	24	Feb-09	Ramaswamy et al., 2011
	China (Shima, Danshui River)	21	Jul-12, Dec-12	Chen et al., 2014
	China (Yangtze River)	23	Jul-13, Dec-13	Liu et al 2015
	China (Yellow River)	15	May-08, Nov-08	Wang et al 2012
	China (Sha, Dongjiang River)	15	Dec-12, Mar-13, Jun-13	Zhang et al 2015
	China (Pearl River)	13	Mar-08, May-08	Yu et al., 2011
	China (Xiaoqing River)	13	Mar-11, Aug-11	Wang et al 2014
	South Korea (Jung-rang Creek)	3	May-13	Ryu et al 2014
	Japan (Kyoto, Saitama, Tokushima)	16	Sept-10 to Feb-11	Kimura et al., 2014
Methyl Paraben	China (Shima, Danshui River)	21	Jul-12, Dec-12	Chen et al., 2014
	Japan (Kyoto, Saitama, Tokushima)	16	Sept-10 to Feb-11	Kimura et al., 2014
	China (Yangtze River)	23	Jul-13, Dec-13	Liu et al 2015
	China (Pearl River)	13	Mar-08, May-08	Yu et al., 2011
	China (Sha, Dongjiang River)	15	Dec-12, Mar-13, Jun-13	Zhang et al 2015
LAS	China (Tinajin)	9	Feb-09	Mu and Wen, 2013
	Philippines (Laguna de bay)	6	Dec-99, Mar-00	Eichhorn et al., 2001

Table D.1.10 – Summary of monitoring data used for model verification for sediment.

Sediment	Country	No of stations	Regime	Author
Triclosan	India (Cauvery, Tamirapani River)	24	Feb-09	Ramaswamy et al., 2011
	China (Shima, Danshui River)	21	Jul-12, Dec-12	Chen et al., 2014
	China (Yangtze River)	23	Jul-13, Dec-13	Liu et al 2015
	China (Yellow River)	15	May-08, Nov-08	Wang et al 2012
	China (Sha, Dongjiang River)	15	Dec-12, Mar-13, Jun-13	Zhang et al 2015
	China (Xiaoqing River)	13	Mar-11, Aug-11	Wang et al 2014
Methyl Paraben	China (Shima, Danshui River)	21	Jul-12, Dec-12	Chen et al., 2014
	China (Yangtze River)	23	Jul-13, Dec-13	Liu et al 2015
	China (Sha, Dongjiang River)	15	Dec-12, Mar-13, Jun-13	Zhang et al 2015
LAS	China (Tinajin)	9	Feb-09	Mu and Wen, 2013

D.1.5 Results

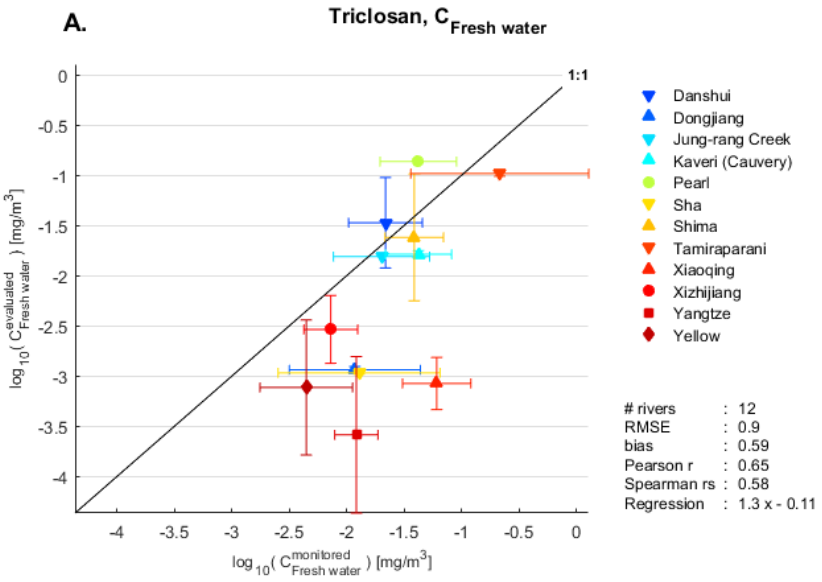


Figure D.1.5 – Comparison of local/evaluated concentrations versus monitored concentrations: local evaluations severely underestimate low-end concentrations in large rivers, when modeled PECs (accounting for upstream history and inputs from other media) match well monitored concentrations (Figure 4.A in main text).

D.2 Multi-scale spatial modeling of human exposure from local sources to global intake

D.2.1 Media and environmental process models (EPMs)

Figure D.2.1 shows media involved in *Pangea*, and substances fate and transport processes taken into account. We distinguish processes (or phenomena) happening within each compartment of each medium, such as degradation, from transport phenomena that describe transfers between compartments representing similar or different media, such as advection and diffusion.

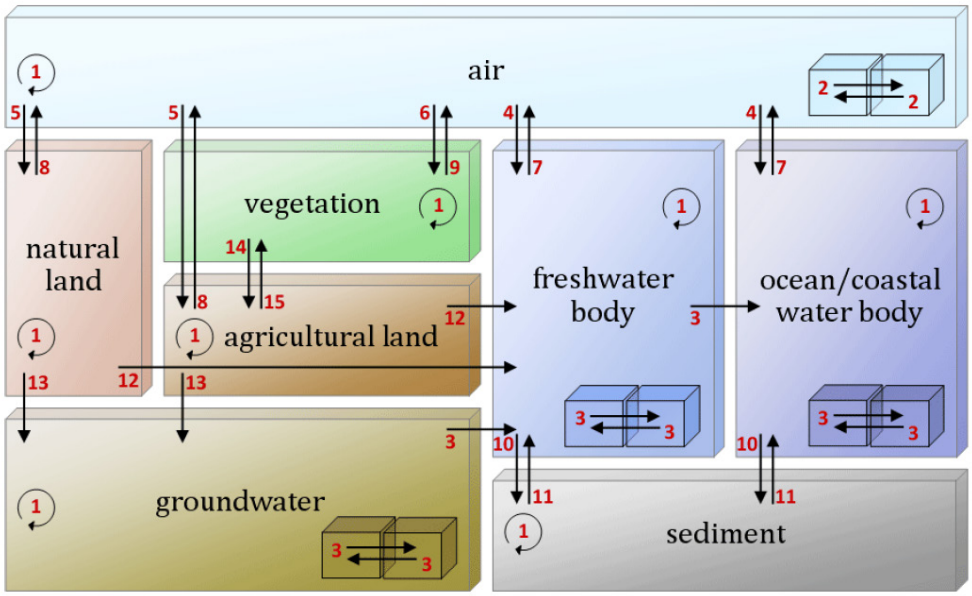


Figure D.2.1 – Set of media and related environmental processes present in *Pangea* (vegetation is not yet implemented).

Models that describe phenomena involved in substances fate and transport are defined in *Pangea* as environemental processes models (EPMs); they are listed in Table D.2.1. *Pangea* accounts as well for the exposure as a removal pathway, in the computation of the fate. While generally considered negligible, exposure-inducted removal is significant for pollutants which bioaccumulate in fish, and in regions/cells which contain only a small amount of surface fresh water with respect to the size of the population. Not accounting for fish production as a removal pathway from the environment can lead to situations where fishes bioaccumulate more pollutant than the quantity available in the whole body of water in which they are immersed. Exposing populations to large fresh water concentrations (through drinking) in regions with almost no surface fresh water fails at acounting for the fact that in these regions most of the fresh water used as drinking water is ground water.

Table D.2.1 – Set of environmental processes models (EPMs) used in this study. Mechanisms: degradation (deg.), advection (adv.), diffusion (diff.), deposition (dep.).

EPM	Mechanism
Elimination:	
air	deg.
nat. and agr. lands	deg.
water	deg.
sediments	deg.
Transfer:	
air → air	adv.
air → water	dry/wet dep. + diff.
air → natural land	dry/wet dep. + diff.
air → agricultural land	dry/wet dep. + diff.
water → water	adv.
water → air	diff.
water → sediments	dep.
sediments → water	resuspension
natural land → air	diff.
natural land → water	runoff
agricultural land → air	diff.
agricultural land → water	runoff

Table D.2.2 – Main non-spatial, chemical dependent, EPMs parameters, used as a basis for computing elimination and transfer coefficients.

Parameter	Description
$t_{1/2,air}$	Degradation half-life in air.
$t_{1/2,water}$	Degradation half-life in bulk water.
$t_{1/2,sediments}$	Degradation half-life in bulk sediments.
$t_{1/2,soils}$	Degradation half-life in bulk soil.
$K_{air/water}$	Air/water partition coefficient.
$K_{octanol/water}$	N-octanol/water partition coefficient.
$K_{oc/water}$	Soil organic carbon/water partition coefficient.

Table D.2.3 – Non-spatial, chemical independent, EPMs parameters, used as a basis for computing elimination and transfer coefficients.

Parameter	Description
$C_{air,particles,suspended}$	Total average suspended particle concentration in air.
$fr\_C_{air,aerosol,OM}$	Organic fraction of tropospheric aerosol.
$a_{particles/gas-octanol/air}$	Coefficient linear relation Kparticle/gas -Koctanol/air
$p_{air,rain,year}$	Average annual precipitation rate in air.
$p_{air,rain,wet}$	Precipitation rate in air during the wet period, i. e. intensity of a rain event.
$v_{air,dep,dry,particles}$	Deposition velocity of particles in air during wet period.
$a_{air,dep,dry,gas-u10}$	Correlation factor between deposition velocity of gas phase in air during the dry period and mean wind velocity.
u10	Mean wind velocity at 10m above (terrestrial/aquatic) surface level.
$t_{air,rain,int}$	Interval period between two consecutive rain events in air.
$r\_W_{air,wet,particles}$	Yearly average washout ratio of aerosols (particles) in air.
$fr\_V_{air,rain,water}$	Fraction of volumetric water in air at air/water droplet equilibrium.
$C_{water,pcom}$	Concentration of colloidal organic matter per unit volume of bulk water.
$C_{water,ps}$	Concentration of suspended solids per unit volume of bulk water.
$\Delta l_{air,boundary}$	Diffusion path length (thickness) of air boundary layer.
$\Delta l_{water,boundary}$	Diffusion path length (thickness) of water boundary layer.
$D_{gas,H2O}$	Diffusion coefficient of water vapor in gas.
MW_{H2O}	Molecular weight of water.
$D_{water,O2}$	Diffusion coefficient of gaseous oxygen in water (at +25°C).
MW_{O2}	Molecular weight of oxygen.
$fr\_M_{water,coc}$	Fraction of colloidal organic carbon in water suspended solids.
$h_{sediment}$	Mean depth of sediment body.
$fr\_M_{sediment,resusp}$	Fraction of deposited substance mass that undergoes resuspension.
$\rho_{sediment,ps}$	Density of sediment solid phase (particles).
$v_{water/sediment,sed,ps}$	Mean sedimentation velocity of water particles to sediment.
$fr\_V_{sediment,ps}$	Fraction of volumetric solids (particles) in sediment.
$OC_{sediment}$	Organic carbon content in sediment.
$\Delta l_{soil,boundary}$	Effective diffusion coefficient in bulk soil.
$\rho_{soil,matrix}$	Density of dry soil matrix (i. e. not the density of dry bulk soil, which is the density of dry soil matrix multiplied by the fraction of soil solids).
$fr\_V_{soiltop,ps}$	Fraction of volumetric particles in bulk soil.
$fr\_V_{soiltop,paq}$	Fraction of volumetric water in bulk soil.
$fr\_V_{soiltop,pg}$	Fraction of volumetric gas in bulk soil.
ρ_{water}	Density of water.
$OC_{soiltop}$	Organic carbon content in soil.
$v_{water/sediment,sed}$	Sedimentation velocity from water to sediments.

Table D.2.4 – Main spatial EPMs parameters, used as a basis for computing elimination and transfer coefficients.

Category	Description
Grids geometry	Layers parameters (e.g. altitude), cells dimensions, shapes, coordinates.
Grids structure topology	Connections between cells from same and different grids/layers.
Connections geometry	Contact surfaces areas.
Sources	Locations, intensities.
Hydrology	Discharges, depths, volumes, topology of water network, runoff.
Wind	Horizontal and vertical velocities.
Land covers	Categories of land cover as % of cells surfaces areas.
Population	Count per cell.
Food production	Food production per country.

D.2.2 Data sets and pollutants characteristics

This section provides a description of three major data sets as used in the current *Pangea* setup: GEOS-Chem, GlobCover, and WWDRII. This selection is motivated by the fact that these data sets were processed before being integrated in *Pangea*, whereas the other data sets, listed in Table 4.1 (main text), are used directly as they are distributed by their developers.

D.2.2.1 GEOS-Chem

Website: http://acmg.seas.harvard.edu/geos

The wind field dataset is based on a 2005 'reference' run from GEOS-Chem which provides wind velocities for GEOS-Chem-specific atmospheric layers in the u and v directions. This dataset is provided with a six-hourly resolution and a 2°×2.5° spatial resolution. For each node of the grid, velocities (signed) in the u and v directions are split into speeds (positive) in the u_+, u_-, v_+, v_- directions. A yearly statistic is computed for each direction. This provides a dataset of wind speeds that reflect bidirectional yearly average transfers. Vertical wind speeds are determined by continuity, starting from the lowest level and setting vertical flows to zero at ground level. This data set is then interpolated on a study-specific basis to a resolution suitable for the geometry of the atmospheric grid.

D.2.2.2 GlobCover

Website: http://ionia1.esrin.esa.int/

The composition of the terrestrial grid is determined based on the GlobCover v2.2 data set. It provides a 10 arc second resolution raster whose pixels define categories of land covers among the 23 categories listed below (including category "no data").

These categories are aggregated into three *Pangea*-specific categories: freshwater, natural land, and agricultural land, for defining terrestrial media, and four *Pangea*-specific categories: freshwater, natural land, crop land, and grass land, for distributing food production data (available on a per country basis) onto the terrestrial grid.

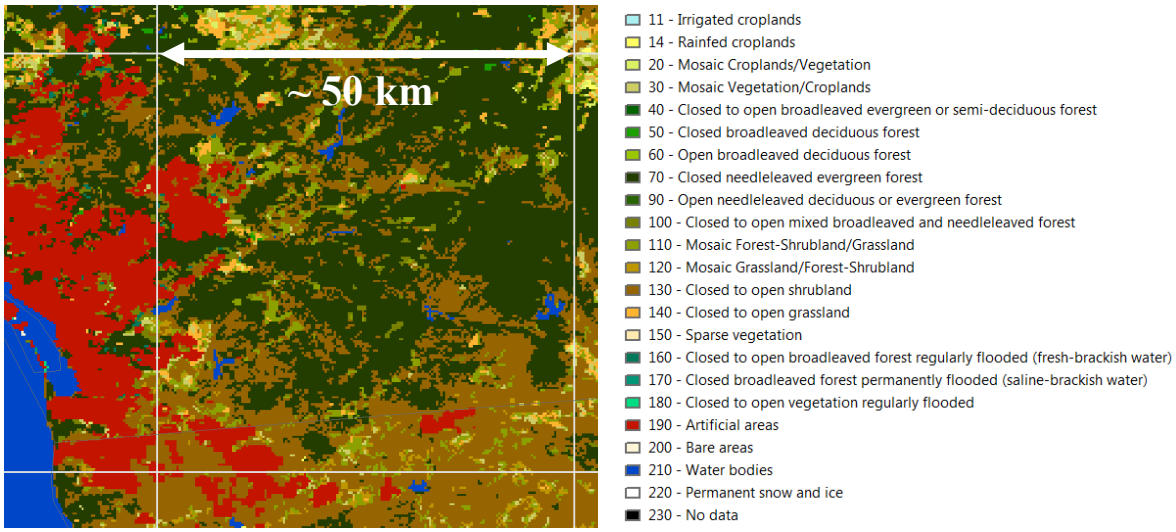


Figure D.2.2 – GlobCover raster (here in the region of San Diego, USA) and categories. The grid represents a native cell from the WWDRII dataset described in the next section.

D.2.2.3 WWDRII

Website: http://wwdrii.sr.unh.edu

The freshwater network model in *Pangea* is based on the gridded $0.5° \times 0.5°$ global water network defined by the World Water Development Report II (WWDRII) as adapted by Helmes et al. (2012, section 2.2.1, advection). The native grid of this model is made of approximately 65,000 cells. WWDRII provides gridded datasets for a variety of hydrological parameters (e.g. discharges, runoff, and precipitation) as well as topological information (the direction of the discharge for each cell).

D.2.3 Pollutants characteristics

Table D.2.5 provides values for the most relevant non-spatial chemical properties and parameters used in assessing the fate and transport of the six pollutants under study. *Pangea* uses a hierarchy of data sources, favoring experimental values when available. Numbers in parentheses link to references for these values. The air-water partition coefficient, K_{aw}, is computed from the Henry constant; the reference provided in Table D.2.5 is for the latter.

Table D.2.5 – Major fate and transport-related chemical properties and parameters, for the six pollutants under study. For PM 2.5, removal is determined by particulate dry and wet deposition. Numbers in parentheses are references to literature.

		PCB118	**2,3,7,8-TCDD**	**B[a]P**	**Benzene**	**1,2-Dichloro-benzene**	**PM2.5**
		31508-00-6	1746-01-6	50-32-8	71-43-2	95-50-1	
K_{aw}	[-]	5.9×10^{-3} (1)	2.1×10^{-3} (2)	3.1×10^{-5} (6)	2.3×10^{-1} (2)	8.0×10^{-2} (2)	-
K_{ow}	[-]	4.9×10^{6} (1)	6.3×10^{6} (2)	1.3×10^{6} (2)	1.3×10^{2} (2)	2.7×10^{3} (2)	-
K_{oc}	[L kg^{-1}]	1.0×10^{5} (2)	6.8×10^{4} (2)	1.8×10^{5} (6)	7.1×10^{1} (2)	9.5×10^{2} (2)	-
$DT50_{air}$	[d]	3.2×10^{1} (3)	1.4×10^{1} (3)	2.1×10^{-1} (6)	8.7×10^{0} (3)	2.5×10^{1} (3)	∞
$DT50_{water}$	[d]	1.3×10^{3} (4)	1.7×10^{2} (5)	1.3×10^{0} (6)	3.8×10^{1} (3)	3.8×10^{1} (3)	∞
$DT50_{sediments}$	[d]	1.9×10^{3} (4)	3.8×10^{4} (5)	5.4×10^{2} (3)	3.4×10^{2} (3)	3.4×10^{2} (3)	∞
$DT50_{soil}$	[d]	1.9×10^{3} (4)	3.8×10^{4} (5)	1.2×10^{2} (3)	7.5×10^{1} (3)	7.5×10^{1} (3)	∞

Appendix E – Illustrations

E.1 Difficulty to detect modeling failures when there is no crash

Figure E.1 is a map of fate factors (FF) associated with a given receptor location. For sake of simplicity, imagine that it is the corresponding map of environmental concentrations (it would look alike, and it is simpler to reason in concentration). It is the output resulting from a failed GIS operation that did not crash.

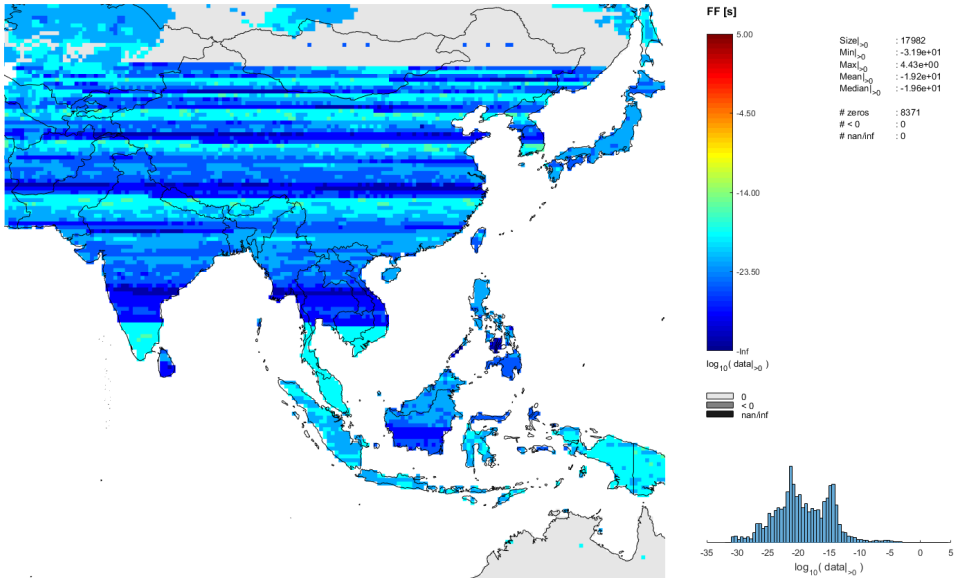


Figure E.1 – Map of fate factors resulting from computations based on a failed GIS processing that did not crash.

Visually, it is obviously wrong; we know that no distribution of concentrations (or fate factors) associated with a natural spread of pollutants can have this pattern of horizontal bands, a gap in the values cut at a given latitude except for some patches, etc. Yet, how to program the detection of such cases? While this is a trivial task for the human eye and intelligence, there is no simple and generic algorithmic way to detect these situations when ranges are correct but the spatial distribution is wrong.

Bibliography

Anderson, P.N., Hites, R.A., 1996. OH radical reactions: The major removal pathway for polychlorinated biphenyls from the atmosphere. Environ. Sci. Technol. 30, 1756–1763. doi:10.1021/es950765k

Armitage, J.M., Cousins, I.T., Hauck, M., Harbers, J. V, Huijbregts, M.A.J., 2007. Empirical evaluation of spatial and non-spatial European-scale multimedia fate models: Results and implications for chemical risk assessment. J. Environ. Monit. 9, 572–581. doi:10.1039/b700680b

Aronson, D., Weeks, J., Meylan, B., Guiney, D.P., Howard, H.P., 2012. Environmental release, environmental concentrations, and ecological risk of N,N-diethyl-m-toluamide (DEET). Integr. Environ. Assess. Manag. 8, 135–166. doi:10.1002/ieam.271

Attias, L., Boccardi, P., Boeije, G., Brooke, D., De Bruijn, J., Comber, M., Dolan, B., Fischer, S., Heinemeyer, G., Koch, V., Lijzen, J., Müller, B., Murray-Smith, R., Rikken, M., Tadeo, J., Vermeire, T., 2005. European union system for the evaluation of substances: The second version. Chemosphere 59, 473–485. doi:10.1016/j.chemosphere.2005.01.062

Bachmann, T.M., 2006. Hazardous Substances and Human HealthExposure, Impact and External Cost Assessment at the European Scale, Hazardous Substances and Human HealthExposure, Impact and External Cost Assessment at the European Scale. Elsevier. doi:http://dx.doi.org/10.1016/S0927-5215(06)80020-2

Bennett, D.H., Margni, M.D., McKone, T.E., Jolliet, O., 2002a. Intake fraction for multimedia pollutants: A tool for life cycle analysis and comparative risk assessment. Risk Anal. 22, 905–918. doi:10.1111/1539-6924.00260

Bennett, D.H., McKone, T.E., Evans, J.S., Nazaroff, W.W., Margni, M.D., Jolliet, O., Smith, K.R., 2002b. Defining intake fraction. Environ. Sci. Technol. 36, 206A–211A.

Bennett, D.H., McKone, T.E., Matthies, M., Kastenberg, W.E., 1998. General formulation of characteristic travel distance for semivolatile organic chemicals in a multimedia environment. Environ. Sci. Technol. 32, 4023–4030. doi:10.1021/es980328g

Bey, I., Jacob, D.J., Yantosca, R.M., Logan, J.A., Field, B.D., Fiore, A.M., Li, Q., Liu, H.Y., Mickley, L.J., Schultz, M.G., 2001. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. J. Geophys. Res. Atmos. 106, 23073–23095.

Beyer, A., Mackay, D., Matthies, M., Wania, F., Webster, E., 2000. Assessing long-range transport potential of persistent organic pollutants. Environ. Sci. Technol. 34, 699–703. doi:10.1021/es990207w

Breivik, K., Armitage, J.M., Wania, F., Sweetman, A.J., Jones, K.C., 2016. Tracking the Global Distribution of Persistent Organic Pollutants Accounting for E-Waste Exports to Developing Regions. Environ. Sci. Technol. 50, 798–805. doi:10.1021/acs.est.5b04226

Breivik, K., Sweetman, A., Pacyna, J.M., Jones, K.C., 2007. Towards a global historical emission inventory for selected PCB congeners - A mass balance approach. 3. An update. Sci. Total Environ. 377, 296–307. doi:10.1016/j.scitotenv.2007.02.026

Brooke, D.N., Crookes, M.J., Gray, D., Robertson, S., 2009. Environmental Risk Assessment Report: Decamethylcyclopentasiloxane. doi:978-1-84911-029-7

Cambridge Environmental Research Consultants, 2016. ADMS5 [WWW Document]. URL http://www.cerc.co.uk/environmental-software/ADMS-model.html

Charles, R., 2004. Modelling Pesticides Residues. Ecole Polytechnique Fédérale de Lausanne (EPFL). doi:3123(2004)

Chen, Z.F., Ying, G.G., Liu, Y.S., Zhang, Q.Q., Zhao, J.L., Liu, S.S., Chen, J., Peng, F.J., Lai, H.J., Pan, C.G., 2014. Triclosan as a surrogate for household biocides: An investigation into biocides in aquatic environments of a highly urbanized region. Water Res. 58, 269–279. doi:10.1016/j.watres.2014.03.072

Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Wilson, R.B., Lee, R.F., Peters, W.D., Brode, R.W., 2005. AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. J. Appl. Meteorol. 44, 682–693. doi:10.1175/JAM2227.1

CMAS, n.d. Community Multiscale Air Quality Model [WWW Document]. URL https://www.cmascenter.org/cmaq/

Cousins, I.T., Mackay, D., 2001. Strategies for including vegetation compartments in multimedia models. Chemosphere 44, 643–654. doi:10.1016/S0045-6535(00)00514-2

Crettaz, P., Pennington, D., Rhomberg, L., Brand, K., Jolliet, O., 2002. Assessing human health response in life cycle assessment using ED10s and DALYs: Part 1 - Cancer effects. Risk Anal. 22, 931–946. doi:10.1111/1539-6924.00262

Davis, T., Rajamanickam, S., Sid-Lakhdar, W., 2016. A survey of direct methods for sparse linear systems.

EEA, 2007. Feasibility study: modelling environmental concentrations of chemicals from emission data. Eur. Environ. Agency. doi:ISSN 1725-2237

Eichhorn, P., Flavier, M.E., Paje, M.L., Knepper, T.P., 2001. Occurrence and fate of linear and branched alkylbenzenesulfonates and their metabolites in surface waters in the Philippines. Sci. Total Environ. 269, 75–85. doi:10.1016/S0048-9697(00)00825-1

ESA, GlobCover, 2009. GlobCover.

Euromonitor, 2016. No Title [WWW Document]. Euromonitor Int. Database. URL http://www.euromonitor.com (accessed 4.14.16).

Evans, J.S., Wolff, S.K., Phonboon, K., Levy, J.I., Smith, K.R., 2002. Exposure efficiency: An idea whose time has come? Chemosphere 49, 1075–1091. doi:10.1016/S0045-6535(02)00242-4

Fantke, P., Friedrich, R., Jolliet, O., 2012. Health impact and damage cost assessment of pesticides in Europe. Environ. Int. 49, 9–17. doi:10.1016/j.envint.2012.08.001

Fantke, P., Wannaz, C., Shaked, S., 2010. Pangea - Implementation of environmental process models (EPM).

Fenner, K., Scheringer, M., Macleod, M., Matthies, M., McKone, T., Stroebe, M., Beyer, A., Bonnell, M., Le Gall, A.C., Klasmeier, J., Mackay, D., Van De Meent, D., Pennington, D., Scharenberg, B., Suzuki, N., Wania, F., 2005. Comparing estimates o persistence and long-range transport potential among multimedia models. Environ. Sci. Technol. 39, 1932–1942. doi:10.1021/es048917b

Fernández, P., Grimalt, J.O., 2003. On the global distribution of persistent organic pollutants. Chimia (Aarau). 57, 514–521.

Franco, A., Price, O.R., Marshall, S., Jolliet, O., Van den Brink, P.J., Rico, A., Focks, A., De Laender, F., Ashauer, R., 2016. Toward refined environmental scenarios for ecological risk assessment of down-the-drain chemicals in freshwater environments. Integr. Environ. Assess. Manag. doi:10.1002/ieam.1801

Franco, A., Struijs, J., Gouin, T., Price, O.R., 2013. Evolution of the sewage treatment plant model SimpleTreat: applicability domain and data requirements. Integr. Environ. Assess. Manag. 9, 560–568.

Gouin, T., Harner, T., 2003. Modelling the environmental fate of the polybrominated diphenyl ethers. Environ. Int. 29, 717–724. doi:10.1016/S0160-4120(03)00116-8

Gouin, T., Thomas, G.O., Cousins, I., Barber, J., Mackay, D., Jones, K.C., 2002. Air-surface exchange of polybrominated diphenyl ethers and polychlorinated biphenyls. Environ. Sci. Technol. 36, 1426–1434. doi:10.1021/es011105k

Hageman, K.J., Bogdal, C., Scheringer, M., 2015. Long-Range and Regional Atmospheric Transport of POPs and Implications for Global Cycling. Compr. Anal. Chem. doi:10.1016/B978-0-444-63299-9.00011-9

Hansen, K.M., Christensen, J.H., Brandt, J., Frohn, L.M., Geels, C., 2004. Modelling atmospheric transport of α-hexachlorocyclohexane in the Northern Hemisphere with a 3-D dynamical model: DEHM-POP. Atmos. Chem. Phys. 4, 1125–1137.

Hansen, K.M., Prevedouros, K., Sweetman, A.J., Jones, K.C., Christensen, J.H., 2006. A process-oriented inter-comparison of a box model and an atmospheric chemistry transport model: Insights into model structure using α-HCH as the modelled substance. Atmos. Environ. 40, 2089–2104. doi:10.1016/j.atmosenv.2005.11.050

Hauschild, M.Z., Huijbregts, M., Jolliet, O., Macleod, M., Margni, M., Van De Meent, D., Rosenbaum, R.K., McKone, T.E., 2008. Building a model based on scientific consensus for life cycle impact assessment of chemicals: The search for harmony and parsimony. Environ. Sci. Technol. 42, 7032–7037. doi:10.1021/es703145t

Helmes, R.J.K., Huijbregts, M.A.J., Henderson, A.D., Jolliet, O., 2012. Spatially explicit fate factors of phosphorous emissions to freshwater at the global scale. Int. J. Life Cycle Assess. 17, 646–654. doi:10.1007/s11367-012-0382-2

HERA, 2013. No Title, Human and environmental risk assessment on ingredients of European household cleaning products. Linear alkylbenzene sulphonate.

Hertwich, E.G., 2001. Intermittent rainfall in dynamic multimedia fate modeling. Environ. Sci. Technol. 35, 936–940. doi:10.1021/es000041v

Hodges, J.E., Vamshi, R., Holmes, C., Rowson, M., Miah, T., Price, O.R., 2014. Combining high-resolution gross domestic product data with home and personal care product market research data to generate a subnational emission inventory for Asia. Integr. Environ. Assess. Manag. 10, 237–246. doi:10.1002/ieam.1476

Hodges, J.E.N., Holmes, C.M., Vamshi, R., Mao, D., Price, O.R., 2012. Estimating chemical emissions from home and personal care products in China. Environ. Pollut. 165, 199–207. doi:10.1016/j.envpol.2011.11.009

Hollander, A., Sauter, F., den Hollander, H., Huijbregts, M., Ragas, A., van de Meent, D., 2007. Spatial variance in multimedia mass balance models: Comparison of LOTOS-EUROS and SimpleBox for PCB-153. Chemosphere 68, 1318–1326. doi:10.1016/j.chemosphere.2007.01.035

Hollander, A., Scheringer, M., Shatalov, V., Mantseva, E., Sweetman, A., Roemer, M., Baart, A., Suzuki, N., Wegmann, F., Van De Meent, D., 2008. Estimating overall persistence and long-range transport potential of persistent organic pollutants: A comparison of seven multimedia mass balance models and atmospheric transport models. J. Environ. Monit. 10, 1139–1147. doi:10.1039/b803760d

Hollander, A., Schoorl, M., van de Meent, D., 2016. SimpleBox 4.0: Improving the model while keeping it simple. Chemosphere 148, 99–107. doi:10.1016/j.chemosphere.2016.01.006

Horstmann, M., McLachlan, M.S., 1998. Atmospheric deposition of semivolatile organic compounds to two forest canopies. Atmos. Environ. 32, 1799–1809. doi:10.1016/S1352-2310(97)00477-9

Humbert, S., Manneh, R., Shaked, S., Wannaz, C., Horvath, A., Deschênes, L., Jolliet, O., Margni, M., 2009. Assessing regional intake fractions in North America. Sci. Total Environ. 407, 4812–4820.

ISWA, 2012. Waste-to-Energy, State-of-the-Art-Report.

Jacquez, J.A., Simon, C.P., 1993. Qualitative theory of compartmental systems. SIAM Rev. 35, 43–79.

Jolliet, O., Hauschild, M., 2005. Modeling the influence of intermittent rain events on long-term fate and transport of organic air pollutants. Environ. Sci. Technol. 39, 4513–4522. doi:10.1021/es049913+

Jolliet, O., Shaked, S., Friot, D., Humbert, S., Schwarzer, S., Margni, M., 2008. Multicontinental long range intake fraction of POPS:importance of food exposure and food exports. Organohalogen Compd. 70, 1939–1941.

Kehrein, N., Berlekamp, J., Klasmeier, J., 2015. Modeling the fate of down-the-drain chemicals in whole watersheds: New version of the GREAT-ER software. Environ. Model. Softw. 64, 1–8. doi:10.1016/j.envsoft.2014.10.018

Keller, V., 2006. Risk assessment of “down-the-drain” chemicals: Search for a suitable model. Sci. Total

Environ. 360, 305–318. doi:10.1016/j.scitotenv.2005.08.042

Keller, V.D.J., Williams, R.J., Lofthouse, C., Johnson, A.C., 2014. Worldwide estimation of river concentrations of any chemical originating from sewage-treatment plants using dilution factors. Environ. Toxicol. Chem. 33, 447–452. doi:10.1002/etc.2441

Kimura, K., Kameda, Y., Yamamoto, H., Nakada, N., Tamura, I., Miyazaki, M., Masunaga, S., 2014. Occurrence of preservatives and antimicrobials in Japanese rivers. Chemosphere 107, 393–399. doi:10.1016/j.chemosphere.2014.01.008

Klepper, O., Den Hollander, H.A., 1999. A comparison of spatially explicit and box models for the fate of chemicals in water, air and soil in Europe. Ecol. Modell. 116, 183–202. doi:10.1016/S0304-3800(98)00161-6

Koziol, A.S., Pudykiewicz, J.A., 2001. Global-scale environmental transport of persistent organic pollutants. Chemosphere 45, 1181–1200. doi:10.1016/S0045-6535(01)00004-2

Kuramochi, H., Takigami, H., Scheringer, M., Sakai, S.-I., 2014. Estimation of physicochemical properties of 52 non-PBDE brominated flame retardants and evaluation of their overall persistence and long-range transport potential. Sci. Total Environ. 491–492, 108–117. doi:10.1016/j.scitotenv.2014.04.004

Lammel, G., Klöpffer, W., Semeena, V.S., Schmidt, E., Leip, A., 2007. Multicompartmental fate of persistent substances: Comparison of predictions from multi-media box models and a multicompartment chemistry-atmospheric transport model. Environ. Sci. Pollut. Res. 14, 153–165. doi:10.1065/espr2006.11.363

Lehner, B., Grill, G., 2013. Global river hydrography and network routing: Baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27, 2171–2186. doi:10.1002/hyp.9740

Likens, G.E., 2010. River ecosystem ecology: a global perspective: a derivative of Encyclopedia of inland waters.

Lindim, C., van Gils, J., Cousins, I.T., 2016. A large-scale model for simulating the fate & transport of organic contaminants in river basins. Chemosphere 144, 803–810. doi:10.1016/j.chemosphere.2015.09.051

Liu, W.R., Zhao, J.L., Liu, Y.S., Chen, Z.F., Yang, Y.Y., Zhang, Q.Q., Ying, G.G., 2015. Biocides in the Yangtze River of China: Spatiotemporal distribution, mass load and risk assessment. Environ.

Pollut. 200, 53–63. doi:10.1016/j.envpol.2015.02.013

Lohman, K., Seigneur, C., 2001. Atmospheric fate and transport of dioxins: Local impacts. Chemosphere 45, 161–171. doi:10.1016/S0045-6535(00)00559-2

Mackay, D., Cowan-Ellsberry, C.E., Powell, D.E., Woodburn, K.B., Xu, S., Kozerski, G.E., Kim, J., 2015. Decamethylcyclopentasiloxane (D5) environmental sources, fate, transport, and routes of exposure. Environ. Toxicol. Chem. 34, 2689–2702. doi:10.1002/etc.2941

Mackay, D., Webster, E., 2006. Environmental persistence of chemicals. Environ. Sci. Pollut. Res. 13, 43–49. doi:10.1065/espr2006.01.008

Mackay, D., Webster, E., Cousins, I., Cahill, T., Foster, K., Gouin, T., 2001. An Introduction to Multmedia Models. Petersborough, Canada.

MacLeod, M., Bennett, D.H., Perem, M., Maddalena, R.L., McKone, T.E., Mackay, D., 2004. Dependence of intake fraction on release location in a multimedia framework: A case study of four contaminants in North America. J. Ind. Ecol. Journal of, 89–102. doi:10.1162/1088198042442360

MacLeod, M., McKone, T.E., 2004. Multimedia persistence as an indicator of potential for population-level intake of environmental contaminants. Environ. Toxicol. Chem. 23, 2465–2472. doi:10.1897/03-390

MacLeod, M., Riley, W.J., McKone, T.E., 2005. Assessing the influence of climate variability on atmospheric concentrations of polychlorinated biphenyls using a global-scale mass balance model (BETR-Global). Environ. Sci. Technol. 39, 6749–6756. doi:10.1021/es048426r

MacLeod, M., Scheringer, M., McKone, T.E., Hungerbuhler, K., 2010. The state of multimedia mass-balance modeling in environmental science and decision-making. Environ. Sci. Technol. 44, 8360–8364. doi:10.1021/es100968w

MacLeod, M., Woodfine, D.G., Mackay, D., McKone, T., Bennett, D., Maddalena, R., 2001. BETR North America: a regionally segmented multimedia contaminant fate model for North America. Environ. Sci. Pollut. Res. Int. 8, 156–63. doi:10.1007/BF02987379

Margni, M., 2003. Source to intake modeling in life cycle impact assessment. Ecole Polytechnique Fédérale de Lausanne (EPFL).

Margni, M., Pennington, D.W., Amman, C., Jolliet, O., 2004. Evaluating multimedia/multipathway model intake fraction estimates using POP emission and monitoring data. Environ. Pollut. 128,

263–277. doi:10.1016/j.envpol.2003.08.036

Matthijs, E., Holt, M.S., Kiewiet, A., Rijs, G.B.J., 1999. Environmental monitoring for linear alkylbenzene sulfonate, alcohol ethoxylate, alcohol ethoxy sulfate, alcohol sulfate, and soap. Environ. Toxicol. Chem. 18, 2634–2644. doi:10.1002/etc.5620181133

McKone, T.E., Enoch, K.G., 2002. CalTox(R), A multimedia total exposure model spreadsheet user's guide.

McKone, T.E., MacLeod, M., 2003. Tracking multiple pathways of human exposure to persistent multimedia pollutants: Regional, continental, and global-scale models. Annu. Rev. Environ. Resour. doi:10.1146/annurev.energy.28.050302.105623

Meyer, T., Lei, Y.D., Muradi, I., Wania, F., 2009a. Organic contaminant release from melting snow. 1. Influence of chemical partitioning. Environ. Sci. Technol. 43, 657–662. doi:10.1021/es8020217

Meyer, T., Lei, Y.D., Muradi, I., Wania, F., 2009b. Organic contaminant release from melting snow. 2. Influence of snow pack and melt characteristics. Environ. Sci. Technol. 43, 663–668. doi:10.1021/es8020233

Meyer, T., Wania, F., 2011. Modeling the elution of organic chemicals from a melting homogeneous snow pack. Water Res. 45, 3627–3637. doi:10.1016/j.watres.2011.04.011

Meyer, T., Wania, F., 2008. Organic contaminant amplification during snowmelt. Water Res. 42, 1847–1865. doi:10.1016/j.watres.2007.12.016

Mintel, 2016. No Title [WWW Document]. Mintel Glob. New Prod. Database.

Mu, L., Wen, J., 2013. Spatial distribution characteristics of polycyclic musks as a chemical marker in river water and sediment compared with other typical pollutants. Water Sci. Technol. 67, 1173–1180. doi:10.2166/wst.2013.642

Müller-Herold, U., 1996. A simple general limiting law for the overall decay of organic compounds with global pollution potential. Environ. Sci. Technol. 30, 586–591. doi:10.1021/es9503443

Pennington, D.W., Margni, M., Ammann, C., Jolliet, O., 2005. Multimedia fate and human intake modeling: Spatial versus nonspatial insights for chemical emissions in Western Europe. Environ. Sci. Technol. 39, 1119–1128. doi:10.1021/es034598x

Perry, S.G., Cimorelli, A.J., Paine, R.J., Brode, R.W., Weil, J.C., Venkatram, A., Wilson, R.B., Lee, R.F., Peters, W.D., 2005. AERMOD: A Dispersion model for industrial source applications. Part

II: Model performance against 17 field study databases. J. Appl. Meteorol. 44, 694–708. doi:10.1175/JAM2228.1

Pistocchi, A., 2014. GIS based chemical fate modeling principles and applications.

Prevedouros, K., Jones, K.C., Sweetman, A.J., 2004a. European-scale modeling of concentrations and distribution of polybrominated diphenyl ethers in the pentabromodiphenyl ether product. Environ. Sci. Technol. 38, 5993–6001. doi:10.1021/es049206g

Prevedouros, K., MacLeod, M., Jones, K.C., Sweetman, A.J., 2004b. Modelling the fate of persistent organic pollutants in Europe: Parameterisation of a gridded distribution model. Environ. Pollut. 128, 251–261. doi:10.1016/j.envpol.2003.08.041

Ramaswamy, B.R., Shanmugam, G., Velu, G., Rengarajan, B., Larsson, D.G.J., 2011. GC-MS analysis and ecotoxicological risk assessment of triclosan, carbamazepine and parabens in Indian rivers. J. Hazard. Mater. 186, 1586–1593. doi:10.1016/j.jhazmat.2010.12.037

Rochat, D., Margni, M., Jolliet, O., 2006. Continent-specific intake fractions and characterization factors for toxic emissions: Does it make a difference? Int. J. Life Cycle Assess. 11, 55–63. doi:10.1065/lca2006.01.238

Roeckner, E., Bäum, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagermann, S., Kirchner, I., Kornblueh, L., Manzini, E., Rhodin, A., Schlese, U., Schulzweida, U., Tompkins, A., 2003. The Atmospheric General Circulation Model ECHAM5, Part 1.

Rosenbaum, R.K., Bachmann, T.M., Gold, L.S., Huijbregts, M.A.J., Jolliet, O., Juraske, R., Koehler, A., Larsen, H.F., MacLeod, M., Margni, M., McKone, T.E., Payet, J., Schuhmacher, M., Van De Meent, D., Hauschild, M.Z., 2008. USEtox - The UNEP-SETAC toxicity model: Recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment. Int. J. Life Cycle Assess. 13, 532–546. doi:10.1007/s11367-008-0038-4

Rosenbaum, R.K., Margni, M., Jolliet, O., 2007. A flexible matrix algebra framework for the multimedia multipathway modeling of emission to impacts. Environ. Int. 33, 624–634. doi:10.1016/j.envint.2007.01.004

Ryu, J., Oh, J., Snyder, S.A., Yoon, Y., 2014. Determination of micropollutants in combined sewer overflows and their removal in a wastewater treatment plant (Seoul, South Korea). Environ. Monit. Assess. 186, 3239–3251. doi:10.1007/s10661-013-3613-5

Scheringer, M., 1996. Persistence and spatial range as endpoints of an exposure-based assessment of

organic chemicals. Environ. Sci. Technol. 30, 1652–1659. doi:10.1021/es9506418

Scheringer, M., Held, H., Stroebe, M., 2002. Chemrange 2.1—A Multimedia Transport Model for Calculating Persistence and Spatial Range of Organic Chemicals.

Scheringer, M., Jones, K.C., Matthies, M., Simonich, S., Van De Meent, D., 2009. Multimedia partitioning, overall persistence, and long-range transport potential in the context of pops and pbt chemical assessments. Integr. Environ. Assess. Manag. 5, 557–576. doi:10.1897/IEAM_2009-007.1

Scheringer, M., Strempel, S., Hukari, S., Ng, C.A., Blepp, M., Hungerbuhler, K., 2012. How many persistent organic pollutants should we expect? Atmos. Pollut. Res. 3, 383–391. doi:10.5094/APR.2012.044

Scheringer, M., Wania, F., 2003. Multimedia models of global transport and fate of persistent organic pollutants, in: Springer Press (Ed.), The Handbook of Environmental Chemistry. pp. 237–269.

Scheringer, M., Wegmann, F., Fenner, K., Hungerbühler, K., 2000. Investigation of the cold condensation of persistent organic pollutants with a global multimedia fate model. Environ. Sci. Technol. 34, 1842–1850. doi:10.1021/es991085a

Sehili, A.M., Lammel, G., 2007. Global fate and distribution of polycyclic aromatic hydrocarbons emitted from Europe and Russia. Atmos. Environ. 41, 8301–8315. doi:10.1016/j.atmosenv.2007.06.050

Semeena, V.S., Feichter, J., Lammel, G., 2006. Impact of the regional climate and substance properties on the fate and atmospheric long-range transport of persistent organic pollutants - Examples of DDT and γ-HCH. Atmos. Chem. Phys. 6, 1231–1248.

Shaked, S., 2011. Multi-Continental Multimedia Model of Pollutant Intake and Application to Impacts of Global Emissions and Globally Traded Goods. The University of Michigan, Ann Arbor.

Shatalov, V., Ilyin, I., Gusev, A., Travnikov, O., 2016. Heavy metals and persistent organic pollutants: Model assessment of pollution and research activities.

Shen, H., Huang, Y., Wang, R., Zhu, D., Li, W., Shen, G., Wang, B., Zhang, Y., Chen, Y., Lu, Y., Chen, H., Li, T., Sun, K., Li, B., Liu, W., Liu, J., Tao, S., 2013. Global atmospheric emissions of polycyclic aromatic hydrocarbons from 1960 to 2008 and future predictions. Environ. Sci. Technol. 47, 6415–6424. doi:10.1021/es400857z

Silverman, K.C., Tell, J.G., Sargent, E. V, Qiu, Z., 2007. Comparison of the Industrial Source Complex and AERMOD dispersion models: Case study for human health risk assessment. J. Air Waste Manag. Assoc. 57, 1439–1446. doi:10.3155/1047-3289.57.12.1439

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., Roeckner, E., 2013. Atmospheric component of the MPI-M earth system model: ECHAM6. J. Adv. Model. Earth Syst. 5, 146–172. doi:10.1002/jame.20015

Struijs, J., 2014. No Title, SimpleTreat 4.0: a model to predict fate and emission of chemicals in wastewater treatment plants. doi:601353005/2014

Suzuki, N., Murasawa, K., Sakurai, T., Nansai, K., Matsuhashi, K., Moriguchi, Y., Tanabe, K., Nakasugi, O., Morita, M., 2004. Geo-referenced multimedia environmental fate model (G-CIEMS): Model formulation and comparison to the generic model and monitoring approaches. Environ. Sci. Technol. 38, 5682–5693. doi:10.1021/es049261p

Vermeire, T.G., Jager, D.T., Bussian, B., Devillers, J., Den Haan, K., Hansen, B., Lundberg, I., Niessen, H., Robertson, S., Tyle, H., Van Der Zandt, P.T.J., 1997. European Union System for the Evaluation of Substances (EUSES). Principles and structure. Chemosphere 34, 1823–1836. doi:10.1016/S0045-6535(97)00017-9

Vörösmarty, C.J., 2002. Global water assessment and potential contributions from Earth Systems Science. Aquat. Sci. 64, 328–351.

Wambaugh, J.F., Setzer, R.W., Reif, D.M., Gangwal, S., Mitchell-Blackwood, J., Arnot, J.A., Joliet, O., Frame, A., Rabinowitz, J., Knudsen, T.B., Judson, R.S., Egeghy, P., Vallero, D., Cohen Hubal, E.A., 2013. High-throughput models for exposure-based chemical prioritization in the ExpoCast project. Environ. Sci. Technol. 47, 8479–8488. doi:10.1021/es400482g

Wang, L., Ying, G.G., Chen, F., Zhang, L.J., Zhao, J.L., Lai, H.J., Chen, Z.F., Tao, R., 2012. Monitoring of selected estrogenic compounds and estrogenic activity in surface water and sediment of the Yellow River in China using combined chemical and biological tools. Environ. Pollut. 165, 241–249. doi:10.1016/j.envpol.2011.10.005

Wang, X., Homer, M., Dyer, S.D., White-Hull, C., Du, C., 2005. A river water quality model integrated with a web-based geographic information system. J. Environ. Manage. 75, 219–228. doi:10.1016/j.jenvman.2004.11.025

Wang, X.K., Jiang, X.J., Wang, Y.N., Sun, J., Wang, C., Shen, T.T., 2014. Occurrence, distribution, and multi-phase partitioning of triclocarban and triclosan in an urban river receiving wastewater treatment plants effluent in China. Environ. Sci. Pollut. Res. 21, 7065–7074. doi:10.1007/s11356-014-2617-1

Wang, Z., Boucher, J.M., Scheringer, M., Cousins, I.T., Hungerbühler, K., 2017. Toward a Comprehensive Global Emission Inventory of C4–C10 Perfluoroalkanesulfonic Acids (PFSAs) and Related Precursors: Focus on the Life Cycle of C8-Based Products and Ongoing Industrial Transition. Environ. Sci. Technol. 51, 4482–4493. doi:10.1021/acs.est.6b06191

Wang, Z., Cousins, I.T., Scheringer, M., Buck, R.C., Hungerbühler, K., 2014. Global emission inventories for C4-C14 perfluoroalkyl carboxylic acid (PFCA) homologues from 1951 to 2030, Part I: Production and emissions from quantifiable sources. Environ. Int. 70, 62–75. doi:10.1016/j.envint.2014.04.013

Wania, F., Breivik, K., Persson, N.J., McLachlan, M.S., 2006. CoZMo-POP 2 - A fugacity-based dynamic multi-compartmental mass balance model of the fate of persistent organic pollutants. Environ. Model. Softw. 21, 868–884. doi:10.1016/j.envsoft.2005.04.003

Wania, F., Mackay, D., 2000. The global distribution model. A non-steady state multi-compartmental mass balance model of the fate of persistent organic pollutants in the global environment.

Wania, F., Mackay, D., 1999. Global chemical fate of α-hexachlorocyclohexane. 2. Use of a global distribution model for mass balancing, source apportionment, and trend prediction. Environ. Toxicol. Chem. 18, 1400–1407. doi:10.1897/1551-5028(1999)018<1400:GCFOHU>2.3.CO;2

Wania, F., Mackay, D., 1996. Tracking the distribution of persistent organic pollutants. Environ. Sci. Technol. 30, 390A–397A.

Wania, F., Mackay, D., Li, Y.-F., Bidleman, T.F., Strand, A., 1999. Global chemical fate of α-hexachlorocyclohexane. 1. Evaluation of a global distribution model. Environ. Toxicol. Chem. 18, 1390–1399. doi:10.1002/etc.5620180707

Wania, F., McLachlan, M.S., 2001. Estimating the influence of forests on the overall fate of semivolatile organic compounds using a multimedia fate model. Environ. Sci. Technol. 35, 582–590. doi:10.1021/es0011919

Warren, C.S., Mackay, D., Webster, E., Arnot, J.A., 2009. A cautionary note on implications of the well-mixed compartment assumption as applied to mass balance models of chemical fate in flowing

systems. Environ. Toxicol. Chem. 28, 1858–1865. doi:10.1897/08-569.1

Webster, E., Mackay, D., Di Guardo, A., Kane, D., Woodfine, D., 2004. Regional differences in chemical fate model outcome. Chemosphere 55, 1361–1376. doi:10.1016/j.chemosphere.2003.10.061

Webster, E., Mackay, D., Wania, F., 1998. Evaluating environmental persistence. Environ. Toxicol. Chem. 17, 2148–2158. doi:10.1897/1551-5028(1998)017<2148:EEP>2.3.CO;2

Wei, Y., Nishimori, M., Kobara, Y., Akiyama, T., 2008. Development of global scale multimedia contaminant fate model: Incorporating paddy field compartment. Sci. Total Environ. 406, 219–226. doi:10.1016/j.scitotenv.2008.07.053

Whelan, M.J., Hodges, J.E.N., Williams, R.J., Keller, V.D.J., Price, O.R., Li, M., 2012. Estimating surface water concentrations of “down-the-drain” chemicals in China using a global model. Environ. Pollut. 165, 233–240. doi:10.1016/j.envpol.2011.10.035

Woodfine, D.G., MacLeod, M., Mackay, D., Brimacombe, J.R., 2001. Development of continental scale multimedia contaminant fate models: Integrating GIS. Environ. Sci. Pollut. Res. 8, 164–172.

Yu, Y., Huang, Q., Wang, Z., Zhang, K., Tang, C., Cui, J., Feng, J., Peng, X., 2011. Occurrence and behavior of pharmaceuticals, steroid hormones, and endocrine-disrupting personal care products in wastewater and the recipient river water of the Pearl River Delta, South China. J. Environ. Monit. 13, 871–878. doi:10.1039/c0em00602e

Zarfl, C., Scheringer, M., Matthies, M., 2011. Screening criteria for long-range transport potential of organic substances in water. Environ. Sci. Technol. 45, 10075–10081. doi:10.1021/es2012534

Zhang, N.S., Liu, Y.S., Van den Brink, P.J., Price, O.R., Ying, G.G., 2015. Ecological risks of home and personal care products in the riverine environment of a rural region in South China without domestic wastewater treatment facilities. Ecotoxicol. Environ. Saf. 122, 417–425. doi:10.1016/j.ecoenv.2015.09.004

Zhu, Y., Price, O.R., Kilgallon, J., Rendal, C., Tao, S., Jones, K.C., Sweetman, A.J., 2016. A Multimedia Fate Model to Support Chemical Management in China: A Case Study for Selected Trace Organics. Environ. Sci. Technol. 50, 7001–7009. doi:10.1021/acs.est.5b05769

Zou, B., 2010. How should environmental exposure risk be assessed? A comparison of four methods for exposure assessment of air pollutions. Environ. Monit. Assess. 166, 159–167. doi:10.1007/s10661-009-0992-8