Received: 30 September 2011, Revised: 12 March 2012, Accepted: 16 March 2012, Published online in Wiley Online Library: 18 April 2012

(wileyonlinelibrary.com) DOI: 10.1002/env.2145

Regional climate model assessment using statistical upscaling and downscaling techniques[†]

Veronica J. Berrocal[a*], Peter F. Craigmile[b] and Peter Guttorp[c,d]

Climate models are mathematical models that describe the temporal evolution of climate, oceans, atmosphere, ice, and land-use processes, across a spatial domain via systems of partial differential equations. Because these models cannot be solved analytically, the model output is generated numerically over grid boxes. Regional climate models (RCMs), or the dynamic downscaling of global climate models to regional scales, are often used for planning purposes, and it is important to assess carefully the uncertainty of such models. We evaluate the Swedish Meteorological and Hydrological Institute (SMHI) RCM by comparing its model output at the grid box level, with the predictions obtained from two observation-driven spatio-temporal statistical models. The "downscaling model" combines the spatially and temporally smoothed climate model output with temperature observations at synoptic stations in a spatio-temporal linear statistical model. The "upscaling model" describes the observational temperature alone at the daily scale, via a spatio-temporal model that includes a wavelet-based trend, spatially varying seasonality, along with volatility and long-range dependence terms. Both statistical models have the ability to make predictions at a seasonal scale, both at point and grid box level. In the years 1962–2007 in South Central Sweden, we show that the climate model performs well in predicting the annual and seasonal average temperature at three reserved stations, but there are interesting differences among the model output and the statistical model-based predictions at the grid box level. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Gaussian processes; point and areal prediction; regional climate models; space–time statistical modeling; wavelets

1. INTRODUCTION

The assessment of regional climate models (RCMs) using data is a nontrivial task. Climate, being the distribution of weather and other climatic factors over long periods (Rossow *et al.*, 2005; Guttorp and Xu, 2011), can not be measured directly. Rather, a variety of quantities (including weather) are measured, and usually, their long-term averages are compared with the model output. The assessment reports of the Intergovernmental Panel on Climate Change contain many such comparisons (e.g. Solomon *et al.*, 2007). It is important to note that climate models do not produce time series that are directly comparable with weather data. Rather, one needs to look at the *distribution* of variables in the model and in the observations.

A further complication is that a RCM, sometimes called a dynamic downscaling, operates on a relatively small area and needs to use boundary values for the global distribution of the atmosphere, oceans, etc. Hence, discrepancies between the distribution of variables in model output and data could be due to inadequacies in the regional model or in the process generating boundary values, typically a general circulation model (GCM). To separate out these two sources of error, one can run a regional model using observed weather, usually in the form of a reanalysis (running a current weather forecast model on historical data to obtain the best available representation of the atmosphere) (Samuelsson *et al.*, 2011). There may, of course, still be errors due to the boundary process, but the assumption usually made is that this error is smaller than the error resulting from a general circulation model.

In Guttorp and Xu (2011), the output of a RCM was compared to a data series collected in central Stockholm. Since the regional model averages over land types (water, forest, open air) and the data were collected in a park, the direct comparison between model and data, even in the distributional sense, was not appropriate. In this article, we use posterior predictions from two different Bayesian statistical models to assess the open air values from the Swedish Meteorological and Hydrological Institute (SMHI) RCM. This RCM uses boundary values from the ERA40 reanalysis (Uppala *et al.*, 2005). To investigate effects at relatively short, but climatically relevant, temporal scales, we make our assessments at quarterly time scales (winter, spring, summer, and autumn) from 1963 to 2007. The two statistical models use observed surface temperatures measured at 17 stations in South Central Sweden, taken from the SMHI synoptic data bank. The first model relates

[*] *Correspondence to: Veronica J. Berrocal, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A. E-mail: berrocal@umich.edu*

[a] *Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, U.S.A.*

[b] *Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, U.S.A.*

[c] *Department of Statistics, University of Washington, Seattle, WA 98195-4322, U.S.A.*

[d] *Norwegian Computing Center, Oslo, Norway*

[†] *This article is published in Environmetrics as a special issue on Advances in Statistical Methods for Climate Analysis, edited by Peter Guttorp, University of Washington, Norwegian Computing Center, Stephan R. Sain, National Center for Atmospheric Research, Christopher K. Wikle, University of Missouri.*

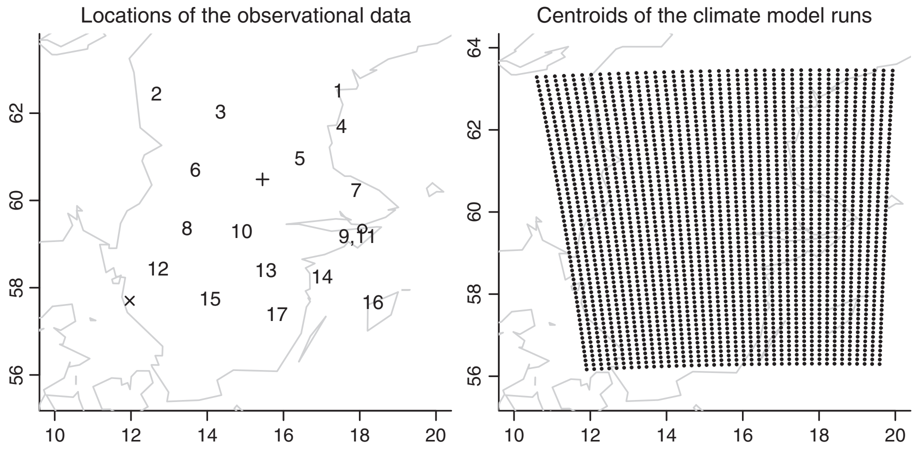
Locations of the observational data
Centroids of the climate model runs
1
2
3
4
5
6
+
7
8
10
9,11
12
13
14
×
15
17
16
62
60
58
56
64
10
12
14
16
18
20

Figure 1. Two maps of the study area of South Central Sweden. On the left hand panel, the numbers denote the location of the 17 stations, while the symbols denote the reserved stations (x, Göteborg; o, Stockholm; +, Borlänge). The centroids of the climate model runs are shown on the right hand panel

observed quarterly average temperatures to a smoothed version of the RCM output (a statistical downscaling). The second model builds an involved space–time statistical model for daily average temperatures without the use of RCM output. The predictions from the second model are then statistically upscaled to a quarterly time scale.

Our models are defined so that they both allow predictions at the point and grid box (areal) level. Hence, we can provide a direct assessment of the RCM at what we hope are more representative spatial scales. Any direct comparison of the RCM to data needs a statistical upscaling, subject to the appropriateness of the model used to upscale. Since the downscaling model is a form of data assimilation, a comparison of the upscaling and downscaling predictions can identify model biases and possibly model uncertainty. Even though it is driven by the RCM, the downscaling model (as the upscaling model) allows us to obtain point predictions at any site. Comparison of any of the predictions to data at observed stations allows us to assess the appropriateness of both the statistical and the climate models.

The paper is organized as follows. In Section 2, we introduce the Swedish temperature series and RCM runs. The statistical model for downscaling is described in Section 3, while the upscaling model is described in Section 4. Section 5 presents our results of downscaling and upscaling to temperature data from three reserved stations (at the point level) and compares the quarterly fields obtained from both approaches (at the grid box level) to the climate model output. We conclude with a discussion in Section 6.

2. DATA

2.1. Swedish temperature series

The observed data were selected from the SMHI 52 station synoptic network, by choosing stations between 12 °E and 19 °E longitude and 57 °N to 63 °N latitude, leaving 17 stations with the earliest data from 1 December 1962 through 30 November 2007. The left panel of Figure 1 shows the network. Further site-specific information is provided in a table in the online supplementary material[‡]. We used the daily mean temperature calculated using the Ekholm–Modén (Alexandersson, 2002) formula that is a function of the daily minimum and maximum temperatures as well as the temperatures at 6, 12, and 18 UTC. The data have been quality controlled but not homogenized (adjusted for local inconsistencies with neighboring series due to, for example, heat island effects; see for example, Lund *et al.*, 2007; Menne and Williams, 2009). On the daily scale, a small percent of data were missing (0.15% across all 17 stations). In our analysis, we imputed a single missing daily value using the average of the daily temperature at the previous and subsequent day and imputed a sequence of missing values using the average of the daily temperatures at the same day in the previous and subsequent year. We then calculated quarterly means (winter: DJF, spring: MAM, summer: JJA, autumn: SON) for each station and quarter that had at least 80 days of daily means available. For model comparisons, we used data from Borlänge, from the Göteborg synoptic station (located just outside the chosen region) and from the Stockholm observatory (not part of the synoptic network). The three symbols on Figure 1 indicate the location of these reserved stations. The Stockholm series is the only one used that has been homogenized and corrected for the urban heat island effect (Moberg *et al.*, 2002).

2.2. Regional climate model runs

Output of the SMHI regional model Rossby Centre Atmospheric (RCA) version 3 (Samuelsson *et al.*, 2011) was available from 1 December 1962 to 30 November 2007 at 12.5 × 12.5 km spatial resolution. For the period from December 1962 to August 2002, boundary conditions for the RCM were provided by the ERA40 (Uppala *et al.*, 2005), while from August 2002 onwards, they were provided by the ECMWF operational analysis or ERA-INTERIM (Uppala *et al.*, 2008). Each 12.5 × 12.5 km grid box was further fractioned by the RCA model into different surface types: lake and/or sea water, lake and/or sea ice, open land, and forest, the latter both with the possibility of being partly covered by snow. Surface temperatures over each surface type were derived using energy balance equations specific to the surface type.

[‡] Supporting information may be found in the online version of this article.

The overall average 2-m temperature over the grid box, as well as the average 2-m temperature over the open land portion of the grid box (including both open land and snow-covered land), over the forest portion of the grid box (including both bare soil and snow), over the lake and/or sea water, and over the lake and/or sea ice portion of the grid box were produced by the climate model as output. Since the synoptic stations are located on open land, in comparing meteorological station data with climate model output, we used the output relative to the average 2-m temperature over open land and snow. Centroids of the RCM grid boxes are displayed in Figure 1, clearly indicating that the region covered by the climate model was roughly identical to the region where the observation stations lie.

The SMHI regional model RCA3 produced predictions of 2-m temperature with a 3-h horizon. The output was provided to us aggregated at the daily scale, and we further aggregated it to the quarter-year scale (again winter: DJF, spring: MAM, summer: JJA, autumn: SON). We thus worked with RCM predictions of quarterly average temperature. For the sake of nomenclature, in the remainder of the article, we will refer to the quarterly mean temperature predictions obtained by aggregating and averaging the climate model daily predictions simply as the RCM output.

Samuelsson *et al.* (2011) suggests that the RCM predicts 2-m temperature reasonably well in Southern Sweden. This is illustrated in the left hand panels of Figure 2, which displays the mean temperatures in different quarters at the 17 synoptic stations considered in our

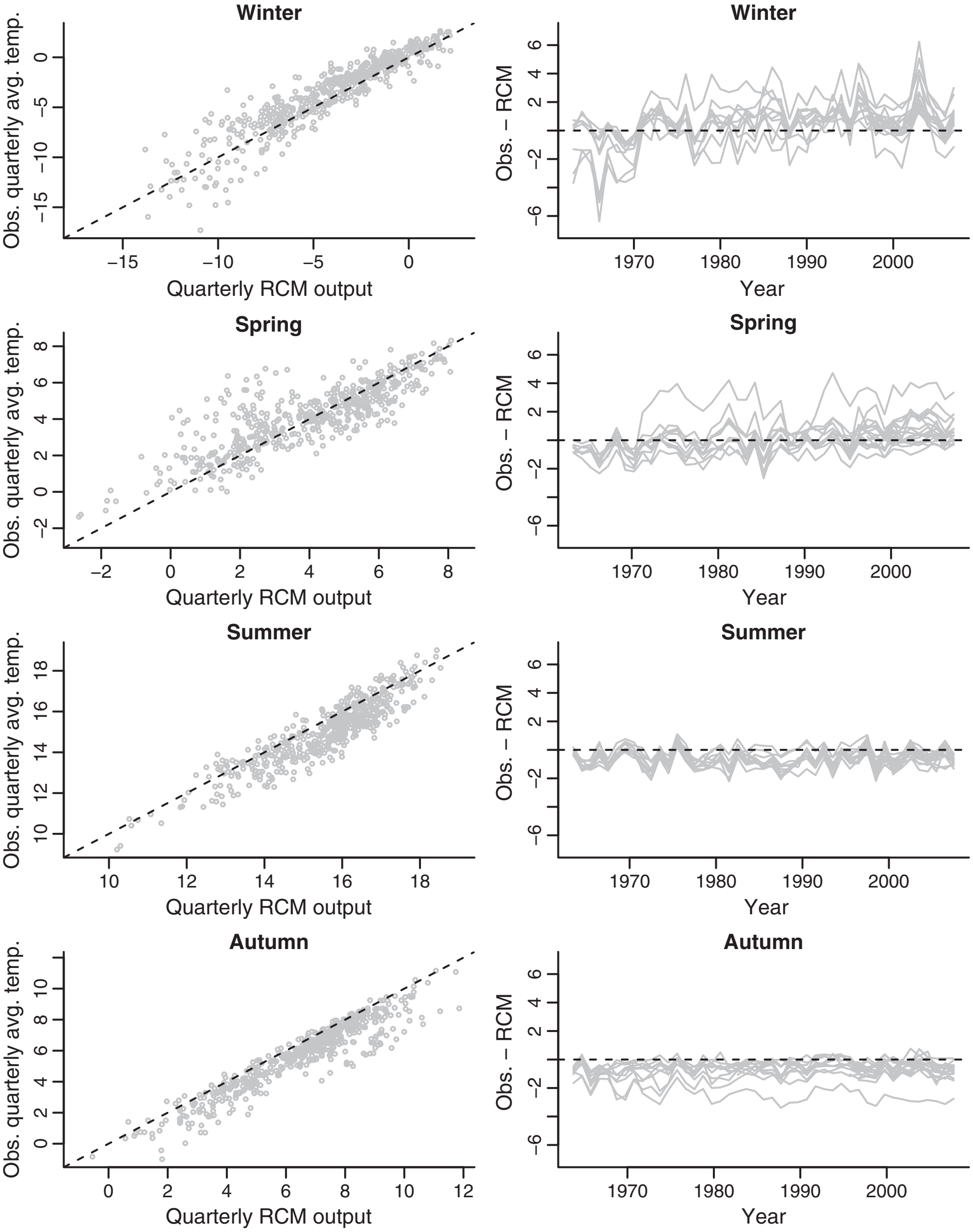

Figure 2. The left hand panels display, for different quarters, the observed mean temperatures versus the regional climate model (RCM) values at each of the 17 synoptic stations. The right hand panels show time series plots of these differences over the 17 stations, for each of the four different quarters. The temperature units are degrees Celsius

application and the RCM output at the 12.5 × 12.5 km grid boxes that contain the stations. On average, during 1962–2007, the difference between the climate model output at the grid boxes containing the stations and the quarterly mean temperature at the 17 stations was about 0.15 °C, indicating a slight warm bias in the RCM, relative to the observations.

However, the scatter plots are hiding some interesting spatial and temporal differences. To investigate such dependencies, the right hand panels of Figure 2 shows time series plots of the observed quarterly mean temperature minus the quarterly RCM output for different quarters at the 17 synoptic meteorological stations. These panels indicate substantial variation by location and over time and illustrate a tendency for the RCM to predict cooler temperatures in winter and spring and warmer temperatures in summer and autumn. There may be also temporal trends in the differences—later years in winter and spring tend to produce bigger discrepancies.

3. A STATISTICAL MODEL FOR DOWNSCALING

In this section, we present the Bayesian hierarchical model used to downscale the RCM output. Let $Y(\boldsymbol{s},t)$ be the quarterly average temperature at site $\boldsymbol{s}$ for quarter t, and let $x(B,t)$ denote the corresponding quarterly average temperature at grid box B predicted by the RCM. An earlier downscaler model proposed by Berrocal *et al.* (2010) downscaled the output $x(B,t)$ at grid box B by establishing a spatial linear model relating each $Y(\boldsymbol{s},t)$ with $x(B,t)$, with B grid box containing $\boldsymbol{s}$. To address potential misalignment between a site $\boldsymbol{s}$ and its putatively associated grid box B and to exploit potential useful information contained in neighboring grid boxes, we extended the previous downscaler model, and following Berrocal *et al.* (2011), for each site $\boldsymbol{s}$ and quarter t, we introduce a smoothed RCM output $\widetilde{x}(\boldsymbol{s},t)$. The smoothed RCM output is related to the observed quarterly average temperature $Y(\boldsymbol{s},t)$ via the following spatial linear model:

$$Y(\boldsymbol{s},t) = \widetilde{\beta}_0(\boldsymbol{s},t) + \beta_1 \widetilde{x}(\boldsymbol{s},t) + \epsilon(\boldsymbol{s},t) \tag{1}$$

where $\epsilon(\boldsymbol{s},t) \overset{iid}{\sim} N(0,\tau^2)$ and $\widetilde{\beta}_0(\boldsymbol{s},t) = \beta_{0,t} + \beta_0(\boldsymbol{s},t)$. Note that since $\widetilde{x}(\boldsymbol{s},t)$ is varying both in space and time, we keep the regression coefficient β_1 fixed in time while we allow the intercept term $\beta_{0,t}$ to vary in time. These regression coefficients $\beta_{0,t}$ and β_1 can be interpreted as calibration terms (i.e., location and scale adjustments) for the smoothed RCM output, while the spatially and temporally varying term $\beta_0(\boldsymbol{s},t)$ provides a local adjustment to $\beta_{0,t}$ at each site $\boldsymbol{s}$ for each quarter t.

The smoothed RCM output $\widetilde{x}(\boldsymbol{s},t)$ at site $\boldsymbol{s}$ and quarter t is obtained from the RCM output by taking a weighted average of the entire output $\{x(B_k,t), k=1,\ldots,g\}$ at quarter t with weights that are spatially and temporally varying. More formally, for each $\boldsymbol{s}$ and $t=1,\ldots,T$, we define $\widetilde{x}(\boldsymbol{s},t)$ as

$$\widetilde{x}(\boldsymbol{s},t) = \sum_{k=1}^{g} w_k(\boldsymbol{s},t)\, x(B_k,t) \tag{2}$$

where g is the number of RCM grid boxes. The weights $w_k(\boldsymbol{s},t)$ are, in turn, modeled as follows: given a latent spatio-temporal Gaussian process $Q(\mathbf{r},t)$ and a kernel function $K(\cdot;\lambda)$, the weights $w_k(\boldsymbol{s},t)$ are obtained as

$$w_k(\boldsymbol{s},t) = \frac{K(\boldsymbol{s}-\mathbf{r}_k;\lambda)\cdot \exp(Q(\mathbf{r}_k,t))}{\sum_{l=1}^{g} K(\boldsymbol{s}-\mathbf{r}_l;\lambda)\cdot \exp(Q(\mathbf{r}_l,t))} \tag{3}$$

where $\{\mathbf{r}_k, k=1,\ldots,g\}$ denotes the set of centroids of the RCM grid boxes $\{B_k, k=1,\ldots,g\}$. As in Berrocal *et al.* (2011), we take the kernel function $K(\cdot;\lambda)$ to be an exponential kernel with parameter λ, that is, $K(\boldsymbol{s}-\mathbf{r}_k;\lambda) = \exp(-\lambda|\boldsymbol{s}-\mathbf{r}_k|)$, where $|\boldsymbol{s}-\mathbf{r}_k|$ denotes the distance between location $\boldsymbol{s}$ and grid box centroid $\mathbf{r}_k$. The decay parameter λ of the kernel function determines how many grid boxes around site $\boldsymbol{s}$ contribute nonnegligibly to $\widetilde{x}(\boldsymbol{s},t)$. Anticipating a displacement error of no more than three grid boxes (that is, about 36 km), we chose to allow up to third-order neighbors of the grid box that contains $\boldsymbol{s}$ to contribute to $\widetilde{x}(\boldsymbol{s},t)$, and thus, we set λ equal to 0.08 km^{-1}, so that weights for grid boxes at least four boxes away were no more than 0.02.

We now specify a model for all spatio-temporal processes in (1) and (2), that is, $\beta_0(\boldsymbol{s},t)$ and $Q(\mathbf{r},t)$. In both cases, we assume that for each quarter t, $\beta_0(\boldsymbol{s},t)$ and $Q(\mathbf{r},t)$ are independent realizations of independent mean zero Gaussian processes provided with an exponential covariance function. In particular, for each quarter t, we assume that $\beta_0(\boldsymbol{s},t)$ is a Gaussian process with an exponential covariance function with marginal variance $\sigma^2_{\beta_0}$ and year-specific decay parameter $\phi_{0,t}$, whereas the Gaussian process $Q(\mathbf{r},t)$ is modeled to have marginal variance σ^2_Q and decay parameter ϕ_Q. Since the decay parameter ϕ_Q controls the smoothness of the latent process $Q(\mathbf{r},t)$ and thus the smoothness in the weights $w_k(\mathbf{r},t)$, to have a set of weights that are spatially varying, the decay parameter ϕ_Q should not be very small. In light of this and after performing several experiments to assess the sensitivity of the results on the magnitude of ϕ_Q, we set ϕ_Q equal to 0.12 km^{-1}, corresponding to a practical range of approximately 25 km, that is, two climate model grid boxes. From the definition of the weights $w_k(\boldsymbol{s},t)$ in (3), it is clear that the process $Q(\mathbf{r},t)$ is not identified: adding a constant to it would not alter the weights. We remedy this by imposing a "sum to zero" constraint that we implement during model fitting.

We considered modeling the temporal dependence in $\beta_0(\boldsymbol{s},t)$ and $Q(\mathbf{r},t)$ by assuming that they were dynamically evolving in time. However, experiments with a different dataset revealed that allowing the decay parameter $\phi_{0,t}$ to vary in time while assuming independence in time for the processes $\beta_0(\boldsymbol{s},t)$ and $Q(\mathbf{r},t)$ yielded a better predictive performance.

We carry out inference for this model within a Bayesian framework—thus, we complete the specification of our model by placing priors on all the remaining model parameters. For the additive time-varying calibration term of the smoothed RCM output, $\beta_{0,t}$, we specify a normal distribution with mean 0 and large variance corresponding to a vague prior, that is, $\beta_{0,t} \overset{iid}{\sim} N\left(0,\varsigma_0^2\right)$; for the multiplicative calibration term β_1, we specify a normal distribution with mean 1 and a large variance. For the error variance τ^2 and the marginal variances,

$\sigma^2_{\beta_0}$ and σ^2_Q, we specify inverse Gamma distributions with prior scale and shape parameters taken to yield vague prior distributions. In particular, following Berrocal *et al.* (2011), for the marginal variance σ^2_Q of the latent process $Q(\mathbf{r}, t)$, we use a prior shape of 2 and a prior scale of 1. Finally, on the decay parameter $\phi_{0,t}$ we place a discrete uniform prior on a grid of 200 values ranging from 0.005 to 0.5.

We fit the model using a Markov chain Monte Carlo (MCMC) algorithm that we ran for 30,000 iterations, discarding the first 10,000 samples for burn in. At each MCMC iteration, we updated the calibration terms ($\beta_{0,t}$ and β_1) and the variance parameters $\left(\tau^2, \sigma^2_{\beta_0}, \sigma^2_Q\right)$ using a Gibbs sampling algorithm, whereas we used a Metropolis–Hastings algorithm for the decay parameter $\phi_{0,t}$ and for the latent process $Q(\mathbf{r}, t)$. Given the large number of climate model grid boxes ($g = 2640$), to alleviate the computational burden encountered in updating the latent process $Q(\mathbf{r}, t)$, the weights $w_k(\boldsymbol{s}, t)$, and the smoothed RCM output $\tilde{x}(\boldsymbol{s}, t)$, we used the predictive process approach of Banerjee *et al.* (2008). Thus, we chose $m = 308$ knots by systematically subsampling the RCM grid box centroids, selecting knots every three rows and three columns, and we replaced the latent Gaussian process $Q(\mathbf{r}, t)$ in (2) with the predictive process $\widetilde{Q}(\mathbf{r}, t)$. More details on the computational aspects of this model can be found in the online supplementary material.

To assess the predictive performance of the downscaling model, in Section 5, we will compare observed quarterly mean temperatures at three stations not used in the fitting with point-level predictions of quarterly mean temperature. Within the downscaling model framework, point-level predictions of quarterly average temperature can be obtained from the posterior predictive distribution and can be estimated via the posterior predictive mean. In Section 5, we also compare the gridded RCM output with the downscaling model predictions of quarterly average temperature over the climate model grid boxes. Such predictions are obtained by employing numerical integration methods, taking a systematic sample of q sites within each RCM grid box. We estimate then the quarterly average temperature over a grid box via the average of the posterior predictive means of the quarterly temperature at the q sites within the grid box.

4. A STATISTICAL MODEL FOR UPSCALING

The statistical model that we upscale to a quarterly-time-scale level is a modified version of the statistical model defined in Craigmile and Guttorp (2011) (details of the changes made are given in the succeeding text). This space–time model was fitted to the daily mean temperatures introduced in Section 2.1: let $D(\boldsymbol{s}, t)$ denote the daily mean temperature at site $\boldsymbol{s}$ and year t (with $t = \text{year} + \text{day in the year}/365.25$). We assume that $D(\boldsymbol{s}, t)$ are independent set of $N(Z(\boldsymbol{s}, t), \sigma^2)$ random variables, where $Z(\boldsymbol{s}, t)$ is the latent mean temperature series at site $\boldsymbol{s}$ and year t, and $\sigma^2 > 0$ is the measurement error variance. Our model for the latent daily mean temperature is

$$Z(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) + \psi(\boldsymbol{s}, t) + \exp(\alpha(\boldsymbol{s}, t))\eta(\boldsymbol{s}, t)$$

where $\mu(\boldsymbol{s}, t)$ is a spatial-time trend term, $\psi(\boldsymbol{s}, t)$ is a seasonal term with amplitude and phase terms that spatially vary, $\alpha(\boldsymbol{s}, t)$ is a volatility term modeling the instantaneous log standard deviation that varies seasonally and spatially, and $\eta(\boldsymbol{s}, t)$ is an error process that captures short and long range dependence that can vary over space.

The space–time trend term $\mu(\boldsymbol{s}, t)$ is defined via a wavelet transform of the data, using Daubechies' "least asymmetric" wavelet filter of width eight (Daubechies, 1992). We assume that the scaling coefficients over a time scale of 512 days (smooth averages of the time series over 512 days at each spatial location; see Craigmile and Guttorp (2011) and Percival and Walden (2000) for further details) are separable space–time process with an AR(1) dependence in time and exponential dependence over space.

Let $\Delta = 1/365.25$ be the sampling interval. As per a suggestion from a referee, we modify the model for the seasonal term $\psi(\boldsymbol{s}, t)$ given in Craigmile and Guttorp (2011) to allow for yearly and half-yearly periods:

$$\psi(\boldsymbol{s}, t) = \psi_1(\boldsymbol{s})\sin(2\pi t\Delta) + \psi_2(\boldsymbol{s})\cos(2\pi t\Delta) + \psi_3(\boldsymbol{s})\sin(4\pi t\Delta) + \psi_4(\boldsymbol{s})\cos(4\pi t\Delta)$$

(This model choice can be justified through general circulation arguments.) The log standard deviation term $\alpha(\boldsymbol{s}, t)$ is also seasonal with yearly and half-yearly periods,

$$\alpha(\boldsymbol{s}, t) = \alpha_0(\boldsymbol{s}) + \alpha_1(\boldsymbol{s})\sin(2\pi t\Delta) + \alpha_2(\boldsymbol{s})\cos(2\pi t\Delta) + \alpha_3(\boldsymbol{s})\sin(4\pi t\Delta) + \alpha_4(\boldsymbol{s})\cos(4\pi t\Delta)$$

and $\eta(\boldsymbol{s}, t)$ is a stationary Gaussian mean-zero space–time process with spatially varying spectral density function at frequency f and site $\boldsymbol{s}$ given by

$$S(f; \boldsymbol{s}) = \Delta|2\sin(\pi f\Delta)|^{-2\delta(\boldsymbol{s})}\exp\left(\sum_{j=1}^{p}\theta_j(\boldsymbol{s})\cos(2\pi f\Delta j)\right)$$

Thus, $\eta(\boldsymbol{s}, t)$ is a spatially varying fractional exponential process of order p. For this analysis, we set $p = 2$, which provides a reasonable trade-off between capturing short-range time series dependence and over-complicating the model.

We complete the model by assuming that the spatial processes, $\psi_1(\cdot)$, …, $\psi_4(\cdot)$, $\alpha_0(\cdot)$, … $\alpha_4(\cdot)$, $\theta_1(\cdot)$, $\theta_2(\cdot)$, are mutually independent Gaussian processes with a constant mean and exponential covariance function defined using the Euclidean distance, a variance parameter, and a spatial range parameter. We model a transformation of the spatially varying long range dependence parameter, $\delta(\cdot)$, $\kappa(\boldsymbol{s}) = \log\left([1/2 + \delta(\boldsymbol{s})]/[1/2 - \delta(\boldsymbol{s})]\right)$, independently of the other spatial processes, assuming this transformed process to be Gaussian with a constant mean and exponential covariance function. We assume independence between the prior distributions for the hyperparameters. For the mean of $\kappa(\cdot)$, we assume a normal distribution with mean 0 and variance 2; the mean of the other spatial processes was each given a diffuse normal distribution with mean 0 and variance 100. For each spatial process, we assumed vague prior distributions for the variance parameter (inverse gamma with shape 0.01 and rate 0.01) and a gamma distribution with shape 10 and rate 1 for the range parameter. Due to documented problems in estimating the measurement error variance and the process variance, we fix the standard deviation of the measurement error at $\sigma = 0.2$ °C (Folland *et al.*, 2001) (Craigmile and Guttorp (2011) provides a robustness study in which that value of σ is altered.) For the space–time trend process, we give the mean a diffuse normal distribution with mean 0 and variance 100, the AR(1) parameter a normal distribution with mean 0.4 and variance 0.04 truncated to $(-1, 1)$, the variance parameter an inverse gamma distribution with shape 0.01 and rate 0.01, and the range parameter a gamma distribution with shape 10 and rate 1.

We fit the model using Markov chain Monte Carlo (MCMC), simplifying the inference via a wavelet transform of the data (see Craigmile and Guttorp (2011) and the supplementary material of that article for details of the algorithm and a discussion of convergence assessments—a summary of the posterior distributions is given in the online supplementary material that accompanies this article). After a burn in of 2000 samples, we based our Bayesian inference on a subsample of 5000 draws (50,000 further samples, subsampled every ten).

In Section 5, for our comparisons with the reserved stations, we sampled from the posterior distribution of the latent daily mean temperatures $Z(\boldsymbol{s}, t)$ at the locations of the three reserved stations. We averaged the daily values over the quarters to produce posterior distributions of the latent quarterly mean temperatures. For comparisons with the climate model output, we used the same numerical integration method as for the downscaling model to obtain posterior predictive means of the quarterly temperatures over each RCM grid box.

5. RESULTS

In this section, we present results for the quarterly average temperature data described in Section 2.1. In Section 5.1, we comment on the posterior distributions of parameters in the downscaling and upscaling models. Then, in Section 5.2, we look at their predictive performance at point level, comparing the predictions yielded by these models with the held-out observations at Borlänge, Stockholm, and Göteborg. Finally, in Section 5.3, we compare, at the grid box level, the quarterly average temperature fields produced by the RCM with the prediction fields produced by the statistical models for downscaling and upscaling.

5.1. Posterior summaries of the model parameters

Examining the posterior distributions of the parameters $\beta_{0,t}$, β_1, and $\beta_0(\boldsymbol{s}, t)$ in the downscaling model provides insight on the calibration of the RCM. Numerical and graphical summaries of these parameters are provided in the supplementary material. The time-invariant multiplicative calibration parameter β_1 has a posterior mean of 0.93 and standard deviation of 0.01, indicating a slight scale inequality between the quarterly average surface temperatures and the RCM output. The posterior mean values for the quarter-specific additive calibration term $\beta_{0,t}$ range from a minimum of -3.3 °C to a maximum of 2.3 °C; the average posterior standard deviation for this parameter is 0.31 °C. Confirming what was already observed in Figure 2, the posterior means of the additive calibration terms $\{\beta_{0,t}\}$ are mostly negative in the autumn quarters from 1962 to 2007, indicating a tendency of the RCM to overpredict temperature in autumn.

In the downscaling model, the spatial variability in the additive error of the RCM is accounted for by $\beta_0(\boldsymbol{s}, t)$. Inspection of the posterior means of $\beta_0(\boldsymbol{s}, t)$ over time reveals that, in general, there is no clear spatial pattern in the additive error $\beta_0(\boldsymbol{s}, t)$ during spring and summer. However, during winter and autumn, clear spatial patterns are evident. In particular, in winter, the RCM tends to underpredict temperature in the southern part of the domain and tends to overpredict it in the north, while in autumn, it tends to predict warmer temperatures than observed over the entire region. As a consequence, in winter, the posterior mean values of $\beta_0(\boldsymbol{s}, t)$ are mostly positive in the south and mostly negative in the north, while in autumn, the posterior means of $\beta_0(\boldsymbol{s}, t)$ are generally negative over the entire spatial domain.

The posterior distribution of the parameters in the upscaling model provides no information about the relationship between surface temperature and the RCM output, but it allows us to understand the components of variation in the daily surface temperatures. Craigmile and Guttorp (2011) fit the upscaling model to a longer period; even with a slightly different model for the seasonal term, the results for the shorter period used in this article are similar. Some numerical and graphical summaries of the parameters are provided in the supplementary material to this article. Inspection of the posterior summaries indicates that there are strong temporal and seasonal patterns in the daily temperatures over South Central Sweden. There is evidence of oscillatory but slightly warming time trends. Naturally, temperatures are colder in winter and warmer in summer, but the amplitudes of the yearly seasonality are higher in the north of the domain. The half-yearly seasonality is much weaker but stronger over the sea in the east of the domain. The phase of the yearly seasonality is more negative in the north, and the phase of the half-yearly seasonality is more negative in the north-east, indicating that not all locations are perfectly in phase with one another. In terms of the noise, the median of the long range dependence parameter $\delta(\boldsymbol{s})$ has a posterior mean value of 0.282, with a 95% credible interval between 0.204 and 0.342, indicating evidence of long-range dependence. Spatial maps of $\delta(\boldsymbol{s})$ show that the long range parameter tends to be higher in the south-west and south-east of the domain. The overall volatility of the noise in the temperatures is larger in the north of the domain, and there is also significant but nontrivial spatially varying seasonal volatility.

The results presented here for both models are robust to moderate changes of the prior distributions. More extreme changes (e.g., changing $\phi_{0,t}$ in the downscaling model by several orders of magnitude or doubling the variability of the range parameters of the spatially varying parameters in the upscaling model) did introduce some changes in the variability of the posterior predictions.

5.2. Reserved stations

5.2.1. Borlänge

The Borlänge station is located in an airport at location 15°30'28" E, 60°25'46" N, at an elevation of 152 m. Due to changing ownership of the airport, the station has been moved twice and has long stretches of missing data. This is clearly illustrated in the left hand panels of Figure 3, which shows the quarterly average temperature at Borlänge (black lines) along with the average quarterly temperature predicted

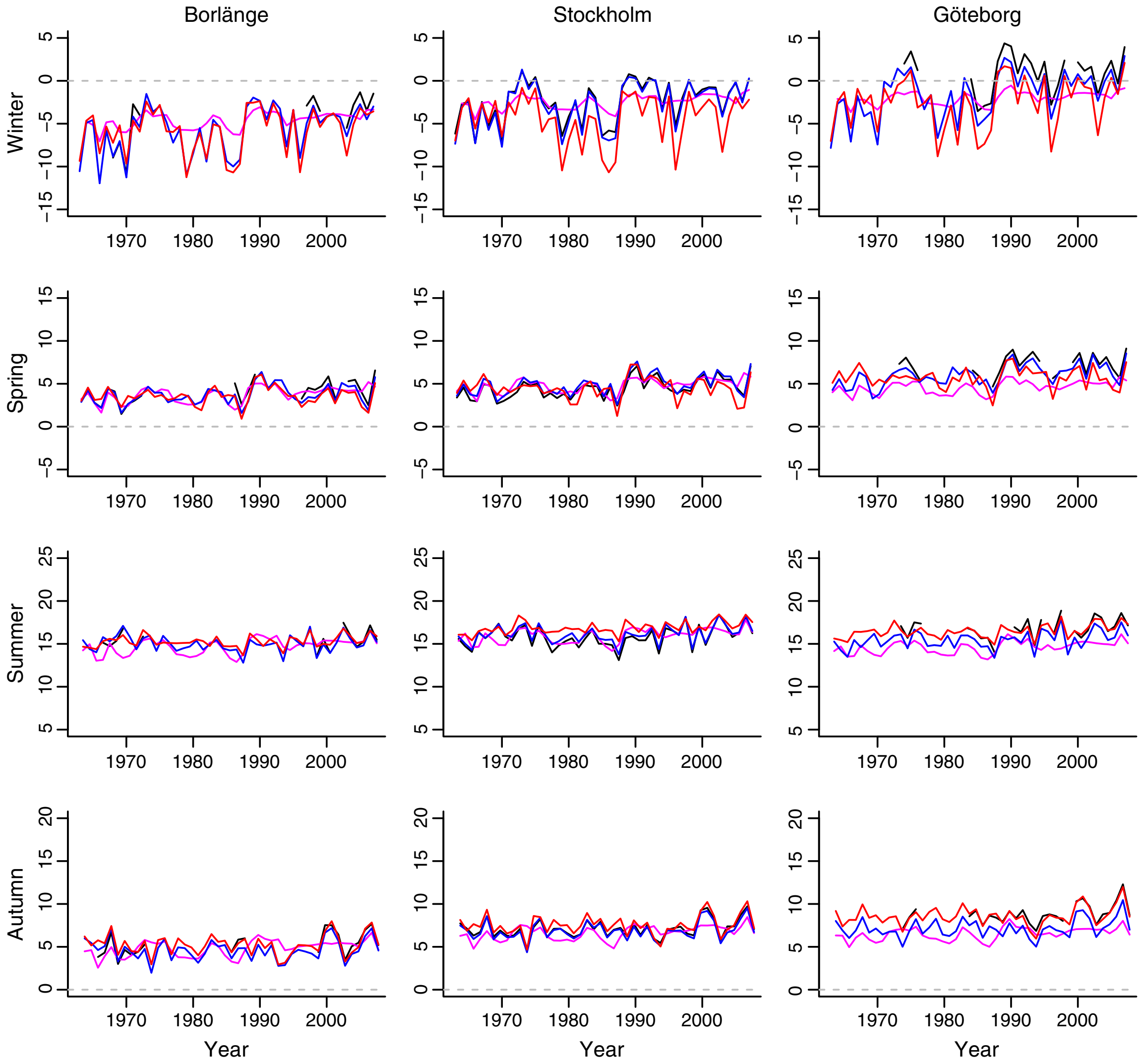

Figure 3. Time series plots comparing the quarterly average temperatures for each quarter (rows) at the three reserved locations (columns). In each panel, the observed data are black, the RCM output is red, the downscaling predictive mean is blue, and the upscaling predictive mean is magenta

by the RCM at the 12.5 × 12.5 km grid box that contains Borlänge (red lines) and the quarterly average temperature predicted by the downscaling (blue lines) and upscaling statistical models (magenta lines). The figure clearly shows that the RCM predicts temperature in Borlänge rather well: the quarter-specific root mean squared error (RMSE) for the winter, spring, summer, and autumn quarter for the RCM are, respectively, 2.0 °C, 1.1 °C, 0.6 °C, and 0.7 °C. The downscaling model is strongly driven by the RCM and, with few exceptions, yields predictions that are very similar to those provided by the RCM, although with a slightly better RMSE: its RMSEs for the winter, spring, summer and autumn quarter are, respectively, 1.2 °C, 0.7 °C, 0.5 °C, and 0.7 °C. On the other hand, the upscaling model yields predictions that are quite different from the climate model and tends to be less in agreement with the observed data particularly in the winter quarter. The quarter-specific RMSEs for the upscaling model are, respectively, 2.6 °C, 1.2 °C, 1.4 °C, and 1.3 °C, for the winter, spring, summer and autumn quarter.

Unlike the RCM, the downscaling and upscaling models provide information on the uncertainty of their predictions. We quantify this uncertainty via the 95% predictive intervals, but in the interests of space, these are not shown. The predictive intervals for the downscaling model are narrower than those of the upscaling model. This can be explained by the fact that the downscaling model exploits the information contained in the entire model output, that is, in all the 2640 grid boxes, whereas the upscaling model predicts temperature at Borlänge using only the information contained in the station data.

5.2.2. Stockholm

The Stockholm observatory has one of the longest continuous temperature series in Sweden, going back to 1756. It is located in a park in central Stockholm at 18°03’17” E and 59°20’30” N, at an altitude of 45 m. As for Borlänge, we assessed the predictive performance of the climate model and that of the upscaling and downscaling models by comparing their predictions with the observed data. The middle panels of Figure 3 compare the observed quarterly average temperature with predictions by the climate model, the downscaling model, and the upscaling model. Both the downscaling and upscaling model yield good predictions, particularly in the winter and summer quarters. In these two quarters, the climate model tends to predict temperatures in Stockholm that are either colder than observed (winter) or warmer than observed (summer), leading to worse RMSEs than the upscaling and downscaling models. For these two quarters, the RMSEs for the climate model, the upscaling model, and the downscaling model are, respectively, equal to 2.5 °C, 1.8 °C, and 0.4 °C (winter) and 1.3 °C, 1.1 °C, and 0.4 °C (summer).

During spring and autumn, the downscaling model yields predictions that are very similar to the RCM predictions even though they have a better RMSE. Indeed, the RMSEs for the downscaling model and the climate model are, respectively, 0.4 °C and 1.0 °C in spring and 0.3 °C and 0.4 °C in autumn. Over these two quarters, the upscaling model yields slightly worse predictions than the other two models, and its predictions have an RMSE of, respectively, 0.9 °C and 1.1 °C. In general, the upscaling model tends to produce predictions that are smoother than observed and fail to capture the peaks in the quarterly average temperature in Stockholm over time.

5.2.3. Göteborg

The Göteborg station is an automatic station, located in an industrial area, close to roads and railroads at 11°59’40” E and 57°42’58” N and an altitude of 2 m. The station has been moved four times. As Figure 1 shows, Göteborg is close to the boundary of the RCM spatial domain and, differently from Stockholm (which also is at the edge of the spatial domain), does not have many stations located nearby. This affects the predictions of the upscaling model as the right panel of Figure 3 demonstrates. For the downscaling model, the lack of observational information produces predictions that tend to follow the RCM predictions closely, although shifted upwards or downwards depending on the quarter and on the size of the additive calibration term $\beta_{0,t}$. In general, the downscaling and upscaling models predict quarterly average temperature in Göteborg that are colder than observed, while the RCM displays such a tendency mostly in the winter and spring quarters. A possible explanation for the overall cold bias of the predictions yielded by both the downscaling and RCM may be the fact that the Göteborg data have not been homogenized for urbanization effects (so they are higher than data from non-urbanized sites). The regional model does not take urbanization into account (E. Kjellström, personal communication), so the predicted temperatures at all quarters will be too low, although it is not obvious why the winter is so low.

5.2.4. Joint verification

In Sections 5.2.1 through 5.2.3, we have assessed the predictive performances of the RCM, the downscaling model, and the upscaling model at individual sites only. However, it is also important to verify whether these models capture the spatial dependence in 2-m temperature. To verify this, we again used the three reserved stations of Borlänge, Stockholm, and Göteborg, and for each pair of stations, we looked at the difference in their observed quarterly average temperature between 1962 and 2007. We constructed similar differences also for the predictions. Thus, for the climate model, we computed the difference between the output at the grid boxes that contain the stations, while for the upscaling and downscaling models, we simply took the difference in the predictive posterior means.

Figure 4 presents plots of the differences for the three pairs of reserved stations. These figures show a moderate level of similarity between the temperatures at the three stations: for each pair of stations, the differences were almost consistently either always positive or always negative and within certain ranges. With the exception of few years, the RCM was able to reproduce the dependence in temperature between the various stations. The downscaling model, being strongly driven by the RCM, also captured the association among the quarterly average temperatures at the three reserved stations quite well. The upscaling model reproduced the spatial association between Borlänge and Stockholm quite well; however, its performance for the two pairs involving Göteborg was impaired by the fact that the upscaling predictions in Göteborg were always much lower than the observations.

5.3. Seasonal fields

We now make comparisons of the predictions of the quarterly average temperatures from the statistical upscaling, statistical downscaling, and the RCM at the grid box levels. As the RCM output is given in terms of quarterly average temperature over 12.5 × 12.5 km grid boxes, the downscaling and upscaling areal predictions at the grid box level have been obtained using the numerical integration procedure discussed in Section 3. Given the limitations of the data and the RCM, we restrict to compare our predictions only at grid boxes over mainland Sweden; locations over Norway or the sea were discarded.

Before making comparisons between the different models, we quantify the level of uncertainty in the quarterly average predictions of the upscaling and downscaling models. Both models had similar posterior predictive standard deviations over time, ranging from 0.3 to 1.6 °C. However, the spatial gradient in the standard deviations of the predictions was different between the two models. The upscaling model generally had larger predictive uncertainty in the northwest region of the domain, where fewer stations are located, whereas the standard deviation of the downscaling predictions showed no significant spatial pattern. Again, this is not unexpected since the downscaling model combines the information contained in the RCM output with the station data.

To quantify discrepancies between the models, for each season we computed the difference between the predicted average quarterly temperature obtained using the downscaling model and the RCM and the difference between the upscaling model and the RCM. By inspecting these differences over time, interesting spatio-temporal trends can be discerned. Due to space limitations in Figure 5, we only present plots for the four quarters of an illustrative year, 2002. Dynamic plots of the differences over the quarters studied are available in the

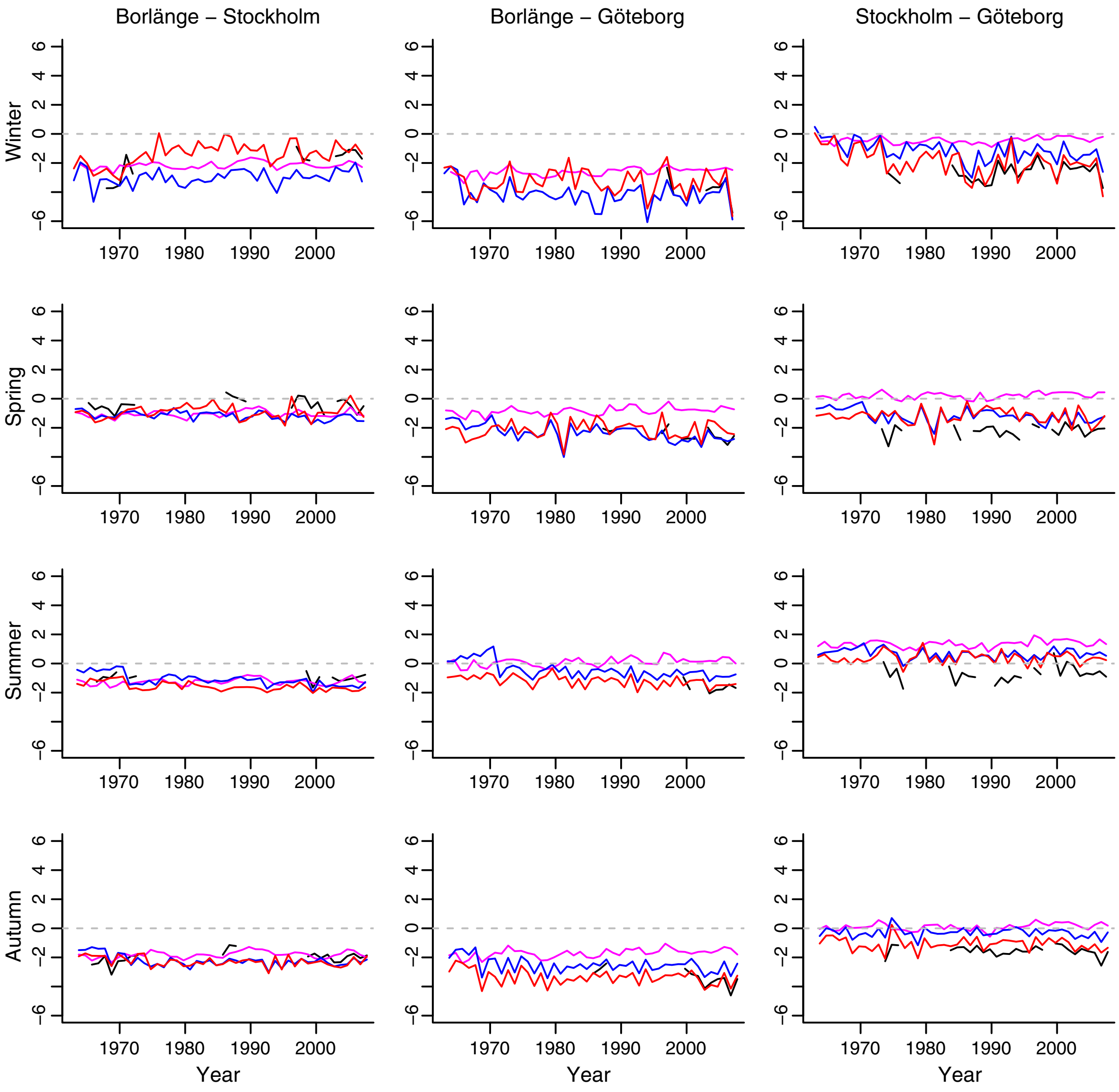

Figure 4. Differences in the observed average quarterly temperature (black line), climate model predictions (solid red line), downscaling predictive mean (blue line), and upscaling predictive mean (magenta line) by quarter (rows) at the three pairs of reserved stations (columns)

supplementary movie file named `animated_spatial_differences.mov`. We now discuss both Figure 5 as well as the spatio-temporal trends for 1963–2007.

In Figure 5, we notice that the relationship between the statistical models and the RCM predictions changes by quarter and by model. The left hand panels show that the climate predictions are colder than the downscaling predictions in winter and spring 2002, they are mixed in summer 2002, and they are hotter than the downscaling predictions in autumn 2002. There is more of a spatial pattern as we compare the upscaling predictions with the climate model output. In winter and spring 2002, the climate predictions are colder than the upscaling predictions in the north but are hotter in the south. The differences are stronger in magnitude as we move north-west or south-west. The differences between the upscaling predictions and the climate model output are more extreme and in a different direction than the differences between the downscaling predictions and the climate model output in summer 2002 and autumn 2002.

Extending our comparsions to the entire time period, we see that the downscaling predictions, as expected, tend to match closely the climate model output, presenting also similar spatial patterns. This is not the case for the upscaling predictions: their difference with the climate model output is typically more extreme, with clear seasonal and spatial patterns. The differences are strongest in the north-west corner of the domain, and they display strong spatial trends from north to south. The upscaling predictions are typically warmer than the climate model output in the north, and with the exception of some extreme quarters (e.g., winter 1998), they tend to be colder than the climate model output in the south. In these so-called extreme quarters, both the downscaling and upscaling models predict warmer temperatures than the RCM. (A reviewer pointed out that some of the later differences that are observed could have been introduced by the change from ERA40 to ERA-INTERIM boundary conditions for the RCM.)

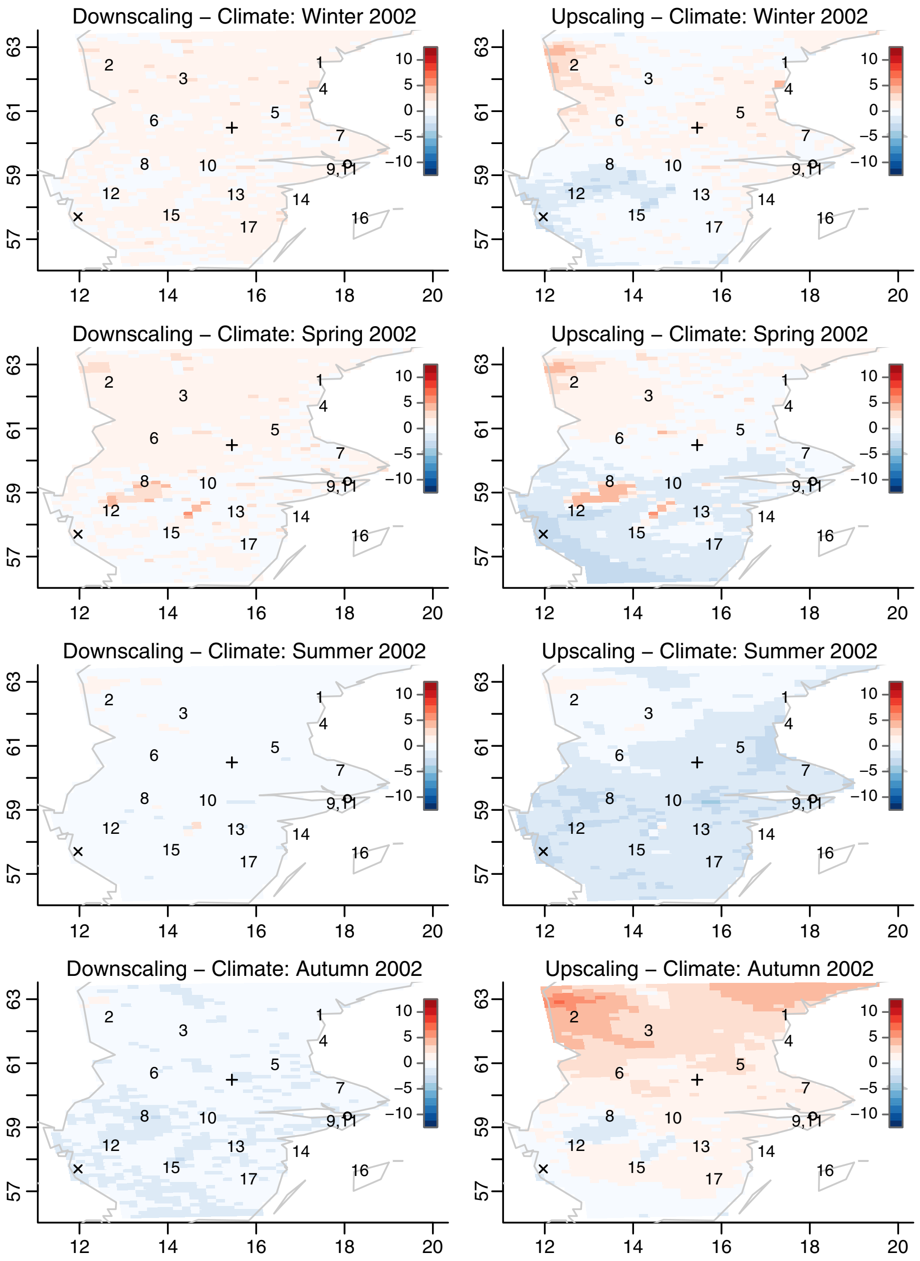
Downscaling – Climate: Winter 2002
Upscaling – Climate: Winter 2002
Downscaling – Climate: Spring 2002
Upscaling – Climate: Spring 2002
Downscaling – Climate: Summer 2002
Upscaling – Climate: Summer 2002
Downscaling – Climate: Autumn 2002
Upscaling – Climate: Autumn 2002
57
59
61
63
12
14
16
18
20
10
5
0
−5
−10
1
2
3
4
5
6
7
8
9, 11
10
12
13
14
15
16
17
×
+

Figure 5. Spatial maps for each quarter (rows) illustrating the difference between the predictive means for the downscaling model and the RCM output (left panels) and the difference between the predictive means for the upscaling model and the RCM output (left panels)

6. DISCUSSION

Because there is no gold standard, assessing the performance of a climate model is, in some respects, an impossible task. Our solution in this article was first to use statistical downscaling and upscaling models, involving observational data and possibly climate model output, as a way to produce predictions of climate on different informative temporal and spatial scales. We then assessed the climate model by comparing its output with the predictions yielded by these statistical models. As expected, the downscaling model incorporating both RCM output and observational data produced predictions that were closest to the RCM ones. However, there were indications of systematic differences between the upscaling and downscaling model predictions and the climate model output, and these differences varied seasonally. We hope that identifying systematic deviations, both spatially and temporally, encourages further research into the development of RCMs.

We assessed the predictive performance of the climate model and of the statistical downscaling and upscaling models by comparing their predictions with data from three reserved stations. Comparison with the held-out data allowed us not only to evaluate the predictive performance of each model but also to indirectly assess model assumptions. There are a number of modeling assumptions that need to be considered when statistical models are used to build space–time predictions of temperature. In our application, the northwest part of the region studied is mountainous. Taking into account altitude as a covariate in the upscaling model could presumably improve the predictions there (altitude is part of the regional model). More generally, the calculations in this article were based on spatially stationary models. In the downscaling model, we assume that the space–time intercept and the latent processes that are used to construct the space–time smoothing weights are each stationary Gaussian processes. In the upscaling model, the space–time trend, the noise component, and the spatially varying parameters are each stationary Gaussian processes. Although the downscaling model is slightly protected against assuming stationarity by incorporating smoothed climate model output, the upscaling model may not be robust to nonstationarity. Since the differences occur at higher altitudes, we could try to understand this disparity further by including altitude as a covariate in our models. Meteorological factors such as prevailing winds, or drivers of climate, may affect the covariance structure, making it nonstationary. It is possible (Reich *et al.*, 2011; Schmidt *et al.*, 2011) to include covariates in the covariance structure, but the added computational complexity would be substantial.

The reason for the apparent additional precision in the downscaling is that the analysis is conditional upon the regional model field. The latter, of course, has its own uncertainty. Kjellström *et al.* (2011) used different initial conditions (for fixed boundary conditions) to assess the uncertainty in RCA3. On annual scales, a comparison of the widths of the credibility sets for the upscaling and downscaling predictions at the reserved sites indicated a reasonably constant difference between these credible sets of about 3 °C. This can be thought of as a rough estimate of the climate model uncertainty. On the other hand, the standard deviations for average annual temperatures in Kjellström *et al.* (2011) were about 2.5 °C, which is sufficiently similar to support the intuitive argument (there is more variability on quarterly time scales).

The spatial extent of a climate model depends on the time scale used. When looking at seasonal data, the difference between a grid box and point measurements is less crucial (since the spatial field of quarterly averages is fairly smooth) than if one looked at daily or 3-h time steps (see the discussion in Sun *et al.*, 2002). Keeping this issue in mind, there would be some interest to see how the RCM and the statistical models varied on shorter-than-seasonal time scales. Such comparisons will be more computationally expensive.

Acknowledgements

We are grateful to Grigory Nikulin of SMHI who provided the RCA3 model output, separated into land use classes. The synoptic data were obtained from the SMHI web site at `http://www.smhi.se/klimatdata/meteorologi/dataserier-for-observationsstationer-1961-2008-1.7375`, and were last accessed September 19, 2011. Peter Craigmile's research is supported by National Science Foundation grant DMS-0906864, and Peter Guttorp's research is supported by National Science Foundation grant DMS-1106862. We thank two anonymous reviewers and an associate editor for comments that greatly improved this work.

REFERENCES

Alexandersson H. 2002. Temperatur och nederbörd i Sverige 1860-2001. *Technical Report ,* Sveriges meteorologiska och hydrologiska institute.

Banerjee A, Gelfand AE, Finley AO, Sang H. 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B* **70**: 825–848.

Berrocal VJ, Gelfand AE, Holland DM. 2010. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics* **15**: 176–197.

Berrocal VJ, Gelfand AE, Holland DM. 2011. Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*, DOI: 10.1111/j.1541-0420.2011.01725.x, (to appear in print). URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2011.01725.x/abstract.

Craigmile PF, Guttorp P. 2011. Space-time modeling of trends in temperature series. *Journal of Time Series Analysis* **32**: 378–395.

Daubechies I. 1992. *Ten lectures on wavelets*, Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM: Philadelphia, PA.

Folland CK, Rayner NA, Brown SJ, Smith TM, Shen SSP, Parker DE, Macadam I, Jones PD, Jones RN, Nicholls N, Sexton MH. 2001. Global temperature change and its uncertainties since 1861. *Geophysical Research Letters* **28**: 2621–2624.

Guttorp P, Xu J. 2011. Climate change, trends in extremes, and model assessment for a long temperature time series from Sweden. *Environmetrics* **22**: 456–463.

Kjellström E, Nikulin G, Hansson U, Strandberg G, Ullerstig A. 2011. 21st century changes in the European climate: uncertainties derived from an ensemble of regional climate model simulations. *Tellus A* **63**: 24–40.

Lund RB, Wang XL, Reeves J, Lu Q, Gallagher C, Feng Y. 2007. Changepoint detection in periodic and autocorrelated time series. *Journal of Climate* **20**: 5178–5190.

Menne MJ, Williams CN. 2009. Homogenization of temperature series via pairwise comparisons. *Journal of Climate* **22**: 1700–1717.

Moberg A, Bergström H, Ruiz Krigsman J, Svanered O. 2002. Daily air temperature and pressure series for Stockholm 1756-1998. *Climatic Change* **53**: 171–212.

Percival DB, Walden A. 2000. *Wavelet Methods for Time Series Analysis*. Cambridge University Press: Cambridge.

Reich BJ, Eidsvik J, Guindani M, Nail AJ, Schmidt AM. 2011. A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentrations. *Annals of Applied Statistics* **5**: 2425–2447.

Rossow WB, Tselioudis G, Polak A, Jakob C. 2005. Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures. *Geophysical Research Letters* **L21812**, DOI: 10.1029/2005GL024584.

Samuelsson P, Jones CG, Willén U, Ullerstig A, Gollvik S, Hansson U, Jansson C, Kjellström E, Nikulin G, Wyser K. 2011. The Rossby Centre regional climate model RCA3: model description and performance. *Tellus A* **63**: 4–23.

Schmidt A, Guttorp P, O'Hagan A. 2011. Considering covariates in the covariance structure of spatial processes. *Environmetrics* **22**: 487–500.

Solomon S, Qin D, Manning M, Chen Z, Marquis B, Averyt KB, Tignor M, Miller HL. 2007. *Climate change 2007: the physical science basis*. Cambridge University Press: Cambridge.

Sun L, Zidek JV, Le ND, Özkaynak H. 2002. Interpolating Vancouver's daily ambient PM_{10} field. *Environmetrics* **13**: 595–613.

Uppala S, Dee D, Kobayashi S, Berrisford P, Simmons A. 2008. Towards a climate data assimilation system: status update of ERA-Interim. *ECMWF Newsletter* **115**: 12–18.

Uppala SM, Kållberg PW, Simmons AJ, Andrae U, Da Costa Bechtold V, Fiorino M, Gibson JK, Haseler J, Hernandez A, Kelly GA, Li X, Onogi K, Saarinen S, Sokka N, Allan RP, Andersson E, Arpe K, Balmaseda MA, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm E, Hoskins BJ, Isaksen L, Janssen PAEM, Jenne R, McNally AP, Mahfouf J-F, Morcrette J-J, Rayner NA, Saunders RW, Simon P, Sterl A, Trenberth KE, Untch A, Vasiljevic D, Viterbo P, Woollen J. 2005. The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society* **131**: 2961–3012.