

# Geophysical Research Letters<sup>®</sup>

## RESEARCH LETTER

10.1029/2022GL100667

### Key Points:

- Identify five clusters in the Great Lakes region with similar runoff potential
- Generalize hybrid models developed at field scales to a continental-scale region
- Predict daily runoff risk on a 1 km-by-1 km grid over the entire Great Lakes region

### Supporting Information:

Supporting Information may be found in the online version of this article.

### Correspondence to:

Y. Hu,  
[yaohu@udel.edu](mailto:yaohu@udel.edu)

### Citation:

Ford, C. M., Hu, Y., Ghosh, C., Fry, L. M., Malakpour-Estalaki, S., Mason, L., et al. (2022). Generalization of runoff risk prediction at field scales to a continental-scale region using cluster analysis and hybrid modeling. *Geophysical Research Letters*, 49, e2022GL100667. <https://doi.org/10.1029/2022GL100667>

Received 4 AUG 2022  
 Accepted 23 AUG 2022

### Author Contributions:

**Conceptualization:** Yao Hu, Lauren M. Fry

**Data curation:** Chanse M. Ford, Yao Hu, Chirantan Ghosh, Lacey Mason, Lindsay Fitzpatrick, Amir Mazrooei, Dustin C. Goering

**Formal analysis:** Chanse M. Ford, Yao Hu

**Funding acquisition:** Yao Hu, Dustin C. Goering

**Investigation:** Chanse M. Ford, Yao Hu

**Methodology:** Chanse M. Ford, Yao Hu, Chirantan Ghosh, Lauren M. Fry

**Project Administration:** Yao Hu, Lauren M. Fry, Dustin C. Goering







**Resources:** Chanse M. Ford, Yao Hu, Lauren M. Fry, Lacey Mason

**Software:** Chanse M. Ford, Yao Hu, Chirantan Ghosh, Siamak Malakpour-Estalaki, Lacey Mason

**Supervision:** Yao Hu, Lauren M. Fry, Lacey Mason

© 2022. American Geophysical Union.  
 All Rights Reserved.

## Generalization of Runoff Risk Prediction at Field Scales to a Continental-Scale Region Using Cluster Analysis and Hybrid Modeling

Chanse M. Ford<sup>1</sup>, Yao Hu<sup>2,3</sup> , Chirantan Ghosh<sup>4</sup> , Lauren M. Fry<sup>5</sup> , Siamak Malakpour-Estalaki<sup>2</sup> , Lacey Mason<sup>5</sup> , Lindsay Fitzpatrick<sup>6</sup>, Amir Mazrooei<sup>7</sup> , and Dustin C. Goering<sup>8</sup>

<sup>1</sup>Department of Earth and Environmental Sciences, Michigan State University, East Lansing, MI, USA, <sup>2</sup>Department of Geography and Spatial Sciences, University of Delaware, Newark, DE, USA, <sup>3</sup>Department of Civil and Environmental Engineering, University of Delaware, Newark, DE, USA, <sup>4</sup>Department of Computer and Information Sciences, University of Delaware, Newark, DE, USA, <sup>5</sup>Great Lakes Environmental Research Lab (GLERL), National Oceanic and Atmospheric Administration, Ann Arbor, MI, USA, <sup>6</sup>Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan, Ann Arbor, MI, USA, <sup>7</sup>Research Application Laboratory, National Center for Atmospheric Research (NCAR), Boulder, CO, USA, <sup>8</sup>North Central River Forecast Center, National Weather Service, National Oceanic and Atmospheric Administration, Chanhassen, MN, USA

**Abstract** As surface water resources in the U.S. continue to be pressured by excess nutrients carried by agricultural runoff, the need to assess runoff risk at the field scale continues to grow in importance. Most landscape hydrologic models developed at regional scales have limited applicability at finer spatial scales. Hybrid models can be used to address the scale mismatch between model simulation and applicability, but could be limited by their ability to generalize over a large domain with heterogeneous hydrologic characteristics. To assist the generalization, we develop a regionalization approach based on the principal component analysis and *K*-means clustering to identify the clusters with similar runoff potential over the Great Lakes region. For each cluster, hybrid models are developed by combining National Oceanic and Atmospheric Administration's National Water Model and a data-driven model, eXtreme gradient boosting with field-scale measurements, enabling prediction of daily runoff risk level at the field scale over the entire region.

**Plain Language Summary** Nutrient loading is an important factor determining water quality in the Great Lakes. Transport of nutrients to surface water is often correlated with runoff, causing detrimental effects to aquatic ecosystems, such as harmful algal blooms. Runoff risk forecasts constituting an early warning system can be used to improve timing of nutrient application, leading to dual benefits of reducing nutrient transport to surface water and leaving more nutrients in the field for crop growth. However, measurements of the edge-of-field runoff are conducted at the field scale and sparse over the Great Lakes region, posing a great challenge to developing such a warning system over the continental scale. To address the challenge, we developed a generalization approach that allows predictive models developed using the runoff measurements at the field scale to be generalized to large regions with similar hydrogeologic characteristics. We can then predict the daily runoff risk level over the entire Great Lakes domain at 1 km-by-1 km resolution, which shows promise to be the backbone of the early warning system on the forecast of daily risk level for the Contiguous U.S.

## 1. Introduction

The Laurentian Great Lakes are one of the largest liquid surface freshwater reservoirs on the planet, containing more than 22,000 km<sup>3</sup> of freshwater (Grannemann et al., 2000). Major threats to this critical water resource associated with nutrient runoff are the formation of large algal blooms, nuisance benthic algae, and large hypoxic areas, which can have negative impacts on aquatic ecosystems and human health (Brooks et al., 2016; Carmichael et al., 2001; Scavia et al., 2019). Investigations into these algal blooms have largely found their formation driven by excess nutrient inputs coming from non-point source pollution from agricultural fields (Michalak et al., 2013; Paerl & Otten, 2013; Stackpoole et al., 2019). Much of this non-point source pollution occurs after snowmelt and precipitation events generate surface runoff from those fields into nearby surface waters which drain to the lakes (Hamlin et al., 2020).

**Validation:** Chanse M. Ford, Yao Hu, Chirantan Ghosh, Siamak Malakpour-Estalaki

**Visualization:** Chanse M. Ford, Yao Hu, Chirantan Ghosh, Siamak Malakpour-Estalaki

**Writing – original draft:** Chanse M. Ford, Yao Hu

**Writing – review & editing:** Chanse M. Ford, Yao Hu, Lauren M. Fry, Siamak Malakpour-Estalaki, Lacey Mason, Dustin C. Goering

One way to reduce the nutrient inputs to these surface waters is to improve the timing of nutrient application to fields such that risk of nutrient runoff is minimized (Gildow et al., 2016; Hopkins & Hansen, 2019). This requires runoff risk prediction at high spatial and temporal resolution. However, despite the advances in high resolution regional and continental scale landscape hydrologic modeling in recent decades, application of these large domain models at a local level is limited where topographic, geologic and climate heterogeneities are more evident. A hybrid modeling approach (Reichstein et al., 2019) that combines physics-based models and statistical models can be used to address the scale mismatch between model simulation and usability with low computational cost and better prediction accuracy (Hu et al., 2021).

In an effort to advance runoff risk predictions relevant to nutrient management at the farm scale over a large domain, this study adopts the hybrid models (Hu et al., 2021) developed at the field-scale where edge-of-field (EOF) runoff measurements are available. However, due to the nature of statistical models which are data and domain specific, it often requires retraining the models when they are applied to a different domain, similar to the recalibration of physics-based models, which can be constrained by the resources (i.e., computational capacity and data availability) and the know-how. To address these challenges, the objectives of this study are twofold: define clusters at the watershed level with similar runoff potential and then generalize the field-scale statistical models to the clusters to predict the risk level of daily runoff over a large domain, demonstrating the potential for management-relevant predictions for agricultural stakeholders in the domain.

## 2. Methods

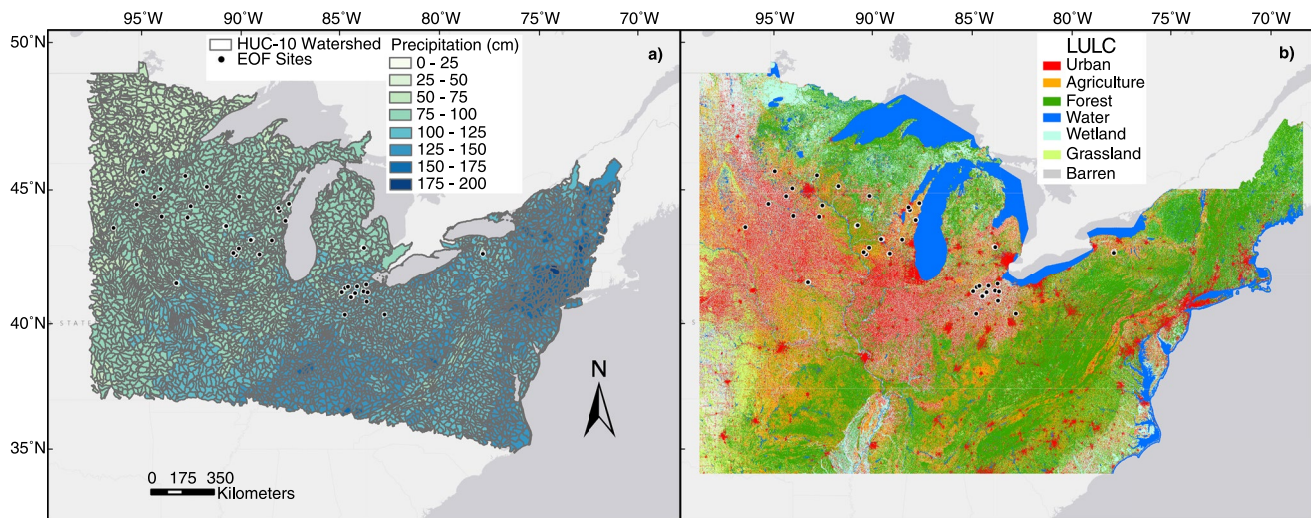
In this study, we generalize statistical models developed to predict the risk level for daily runoff at the field locations, where runoff measurements exist, to the watersheds over the Great Lake region. To do so, we developed a regionalization technique based on unsupervised machine learning methods to group watersheds by similar runoff potential over the Great Lakes domain. The regionalization is necessary to deploy statistical models to larger areas where no observation of runoff is available. This regionalization approach uses variables from both observations and outputs of physics-based models, which can directly or indirectly influence runoff generation. These variables are spatially aggregated across Hydrologic Unit Code 10-digit (HUC-10) watersheds and temporally aggregated to annual values, then cluster analysis techniques are applied to the variables to form the regionalization. Once the clusters are defined, we then train and validate the statistical models for individual clusters to predict the risk level of daily runoff using the approach described in Hu et al. (2021).

### 2.1. Study Area

The Great Lakes domain encompasses a variety of geologies and land covers, with regional precipitation differences. The entire region has a humid climate, with an average annual precipitation amount that has been increasing through time to the current 82.6 cm (Ford et al., 2021; NCEI, 2021). Much of that precipitation occurs as rain, especially in the southern basins, but areas in the Great Lakes and New England can receive significant snowfall amounts in the winter. The geology of the region ranges from the clay-rich soils in the Midwest to the heavily glaciated, sandy soils of the Great Lakes to thin soils underlain by impermeable Precambrian bedrocks in the northern basins (Soller, 2001; Soller et al., 2009). Much of the Appalachian Mountains are also captured in this study area, where the western basins are heavily agricultural lands interspersed with deciduous forests, while the southern and eastern basins tend to be more heavily urbanized. The northern basins of the study area tend to have less agricultural and urban land cover and instead contain large swaths of coniferous forests and wetlands (USGS, 2011, Figure 1b).

### 2.2. Data Preparation

A combination of observed and simulated data sets considered having influence on the runoff generation were used as inputs for regionalization. The observed data come from 79 EOF sites located in different states in the Great Lakes region (Figure 1). These are instrumented sites collecting sub-daily runoff data from the proximal farm field from 2002 to 2018, which was then aggregated to daily values (Hu, 2022). The simulated data included the gridded outputs from a retrospective run of the National Oceanic and Atmospheric Administration's (NOAA) National Water Model (NWM; NOAA, 2016) configuration of WRF-Hydro for the same period. The NWM outputs have different spatial resolution (e.g., ponded surface water on 250 × 250 m grid and soil moisture on



**Figure 1.** Study area for the regionalization which consists of 4,552 Hydrologic Unit Code 10-digit (HUC-10) watersheds and 79 edge of field sites: (a) Mean annual precipitation for each HUC-10 watershed for 3 years in the study period. (b) Land use-land cover data from NLCD (USGS, 2011) across the study domain.

1 × 1 km grid). Several other gridded products were also used as the simulated data, including NOAA's Snow Data Assimilation System (SNODAS), the Oregon State University Climate Group's PRISM model (PRISM Climate Group, 2018), and NLDAS-2 forcing data (Xia et al., 2012). The full list of variables can be found in Table S1 in Supporting Information S1.

Once the data sets were quality controlled and checked, we spatially aggregated the gridded data to HUC-10 watershed polygons (Seaber et al., 1987, Figure 1a). To do so, for each variable, we selected three wet years of data (i.e., 2006, 2010, and 2011) as rainfall is considered important for runoff generation, and then aggregated their daily values into an annual value and rasterized those annual gridded values, from which we calculated their arithmetic mean over the HUC-10 watersheds, that is, annual values of each variable for each watershed. Next, we averaged the 3-year annual values to obtain one annual value for each variable in each HUC-10. Finally, we normalized the annual values for regionalization.

### 2.3. Generalization of Statistical Models

#### 2.3.1. Regionalization

The regionalization approach aims to identify HUC-10 watersheds with similar runoff potential. The HUC-10 was chosen to balance spatial heterogeneity of variables considered for cluster analysis with computational costs, leading to a reasonable number of clusters. The approach comprises two parts: principal component analysis (PCA) and *K*-means analysis. PCA can reduce the dimensionality of the covariates that characterize the runoff potential and identifies the primary components that can describe no less than 85% of data set variability (Artoni et al., 2018; Wang et al., 2017). Those components are used to form the coordinates to transform the data into a new coordinate system. As such, the first coordinate explains the greatest variance of the data set. The reprojected data of HUC10 watersheds was then clustered using *K*-means to minimize the within sum of squares based on the algorithm of Hartigan and Wong (1979). The final number of clusters were determined to balance the minimization of within-cluster variance and the maximization of clusters containing EOF sites. These clusters formed the HUC-10 watersheds of the regionalization. The analyses were conducted in the programming language “R” (Table S2 in Supporting Information S1).

#### 2.3.2. Runoff Risk Prediction

Large magnitudes of EOF runoff often leads to a high level of runoff risk. However, rather than the use of runoff magnitude alone to define risk level, we designed a matrix to define the risk level of daily EOF runoff. The matrix comprises two components: the likelihood of the occurrence of EOF runoff (i.e., occurrence probability, OP) and the magnitude of the occurring EOF runoff (i.e., level of severity, LS) for a given day. We adopted the matrix for

two reasons: (a) It is expected to account for the cases in which high runoff risk can arise from moderate runoff with high likelihood to occur or extreme runoff with moderate likelihood to occur. For the latter case, once the runoff occurs, it can cause severe damage. (b) The matrix acts as a smoothing filter. It intends to maximize the area with the same level of runoff risk and smooth the transition of risk levels over a large domain.

We defined four intervals for LS based on historical EOF measurements ( $M_{\text{EOF}}$ ) and predicted magnitude of EOF runoffs ( $P_{\text{EOF}}$ ; Table S3 in Supporting Information S1). Built from the previous work on hybrid modeling using directed information for causal inference (Hu et al., 2021), we first selected the outputs from the NWM that exhibit statistically causal influence on the runoff generation for each cluster (Tables S5 and S6 in Supporting Information S1). Then, we trained the eXtreme Gradient Boosting (XGBoost) models (Chen & Guestrin, 2016) using the causal outputs to predict the LS and OP of daily EOF runoff at the sites where EOF measurements are taken (Figure 1a). To evaluate the performance of the XGBoost models for ungauged locations, we randomly split EOF sites by 70%/30% within the cluster under a range of split scenarios: EOF measurements from 70% of the EOF sites were used for training and the remaining 30% were for validation (Development of XGBoost Models in Supporting Information S1). Next, we generalized the XGBoost models over each cluster to predict daily LS and OP at 1 km-by-1 km resolution. Finally, we estimated the risk level of daily EOF runoff (i.e., No, Low, Medium, and High) over the entire domain using the risk level matrix defined based on LS and OP (Figure 4a).

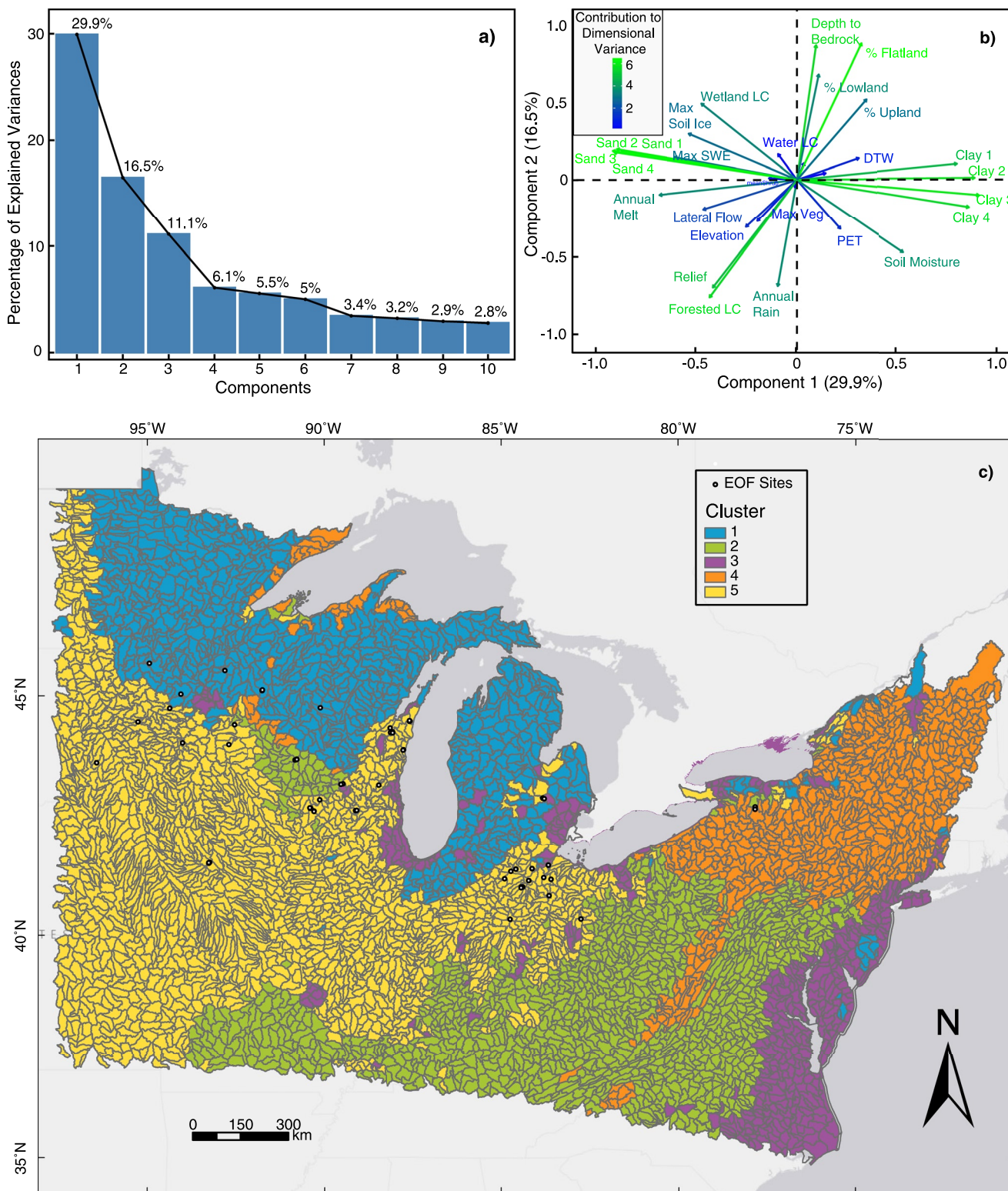
### 3. Results and Discussion

#### 3.1. Regionalization

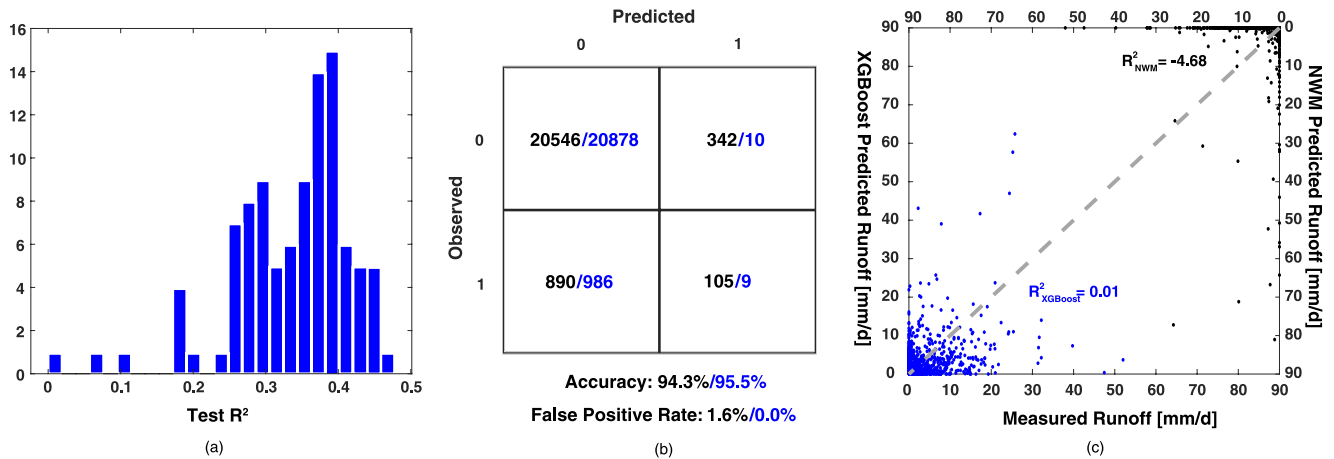
Through PCA, we obtained 30 different components, out of which the first 10 components were selected to reproject the data set for the  $K$ -means analysis, since they accounted for 86.5% (>85.0%) of the variance of the covariates (Figure 2a). Most of the variance was captured by the first two components, that is, 29.9% and 16.5% for the first and second components. For both components, geophysical and topographic variables (i.e., soil content, depth to bedrock, and relief and percent upland) were ranked high in terms of their contributions to the first two components (Figure 2b). Additionally, winter climate variables (e.g., annual melt amount and snow-water equivalent) in the first component and land use-land cover (LULC) variables (e.g., percent forested and percent wetland areas) in the second component appeared to be influential.

We obtained five different clusters through the  $K$ -means analysis (Figure 2c). The number of clusters was chosen to minimize the within-cluster variance while maximizing the number of clusters containing EOF sites for training statistical models (Table S4 in Supporting Information S1). The five clusters produced in this study are largely characterized by their LULC, soil content and topographic differences (Figure 2b). Cluster1, which contains a large portion of the Great Lakes basin, is fairly wet with a significant portion of precipitation coming from snow. It has a relatively flat, low elevation topography with large amounts of wetlands and deep sandy soils. Cluster2 is primarily located in the southern Appalachians and the southern Midwest. The areas are heavily forested with moderate relief and elevation. The region is very wet with high precipitation, the vast majority of which falls as rain on the thin, silty soils. Cluster3 is the most spatially heterogeneous, likely due to its high urban LULC. The majority of the east coast is contained in this cluster with more urbanized watersheds such as Chicago, Cleveland and St. Louis, U.S. These areas are very flat and low in elevation, with thick sandy soils. Precipitation in Cluster3 is similar to Cluster2, with high precipitation totals dominated by rain. Cluster4 is mostly located in the northern Appalachians and New England, with some areas in the western Great Lakes. This region is heavily forested, with thin sandy soils over areas of high relief and elevation. The climate in these watersheds is very wet, with the highest snowfall and ice variables across the study area. Finally, Cluster5 has the least amount of defining characteristics, likely due to the heavily agricultural Midwestern areas contained in the region. It is flat, low elevation areas dominated by thick, silty soils receiving a moderate amount of rain each year.

The landscape and hydrologic differences observed in the five clusters indicate different runoff potential, and correlate with other studies finding the physical characteristics of basins can predict the runoff ratio of the basin (Kult et al., 2014). Clusters 1, 3, and 4 are likely to produce higher runoff amounts than the remaining two clusters. Cluster1 is likely to generate runoff despite its deep sandy soils because of the amount of moisture, both liquid and solid, in those soils. The shallow water table depths and high precipitation amounts in combination could produce significant runoff, especially in the spring during the snowmelt. Cluster3's highly urbanized areas are the main driver of runoff generation for the region. Cluster4 is probably the most likely cluster to produce high



**Figure 2.** (a) The variance of the covariates described by each component in the principal component analysis (PCA). (b) The influence of covariates on the first two components in the PCA. The *x*-axis is the contributions of the vectorized covariates to the first component and the *y*-axis is the contributions to the second component. Larger absolute values indicate more influence on that component by that variable, with negative values inversely influencing the component. The color of the vectors represents the strength of the absolute contribution to the variance of the whole data set by that variable. (c) Five clusters generated by the *K*-means analysis.



**Figure 3.** Performance evaluation of the eXtreme Gradient Boosting (XGBoost) models for ungauged locations in Cluster5: (a) Histogram of the test  $R^2$  values on the magnitude prediction of daily edge-of-field (EOF) runoff for 30% of the EOF sites in Cluster5 as the result of 100 random 70%/30% splits by EOF sites ( $R^2_{\min} = 0.01$ ;  $R^2_{\text{med}} = 0.35$ ;  $R^2_{\max} = 0.48$ ). (b) Confusion matrices of the occurrence predictions of daily runoff events by the National Water Model (NWM) (black) and the XGBoost model (blue) for the split scenario with  $R^2 = 0.01$ : 0/1: no/yes for a runoff event. (c) Scatter plots of the comparisons between the observed runoff events and the predictions by the NWM (upper right, top x-axis and right y-axis) and the XGBoost model (lower left, bottom x-axis and left y-axis) for the split scenario with  $R^2 = 0.01$ .

amounts of runoff. The thin soils, high relief and large amounts of precipitation and soil ice that characterize the cluster will enhance whatever runoff is generated.

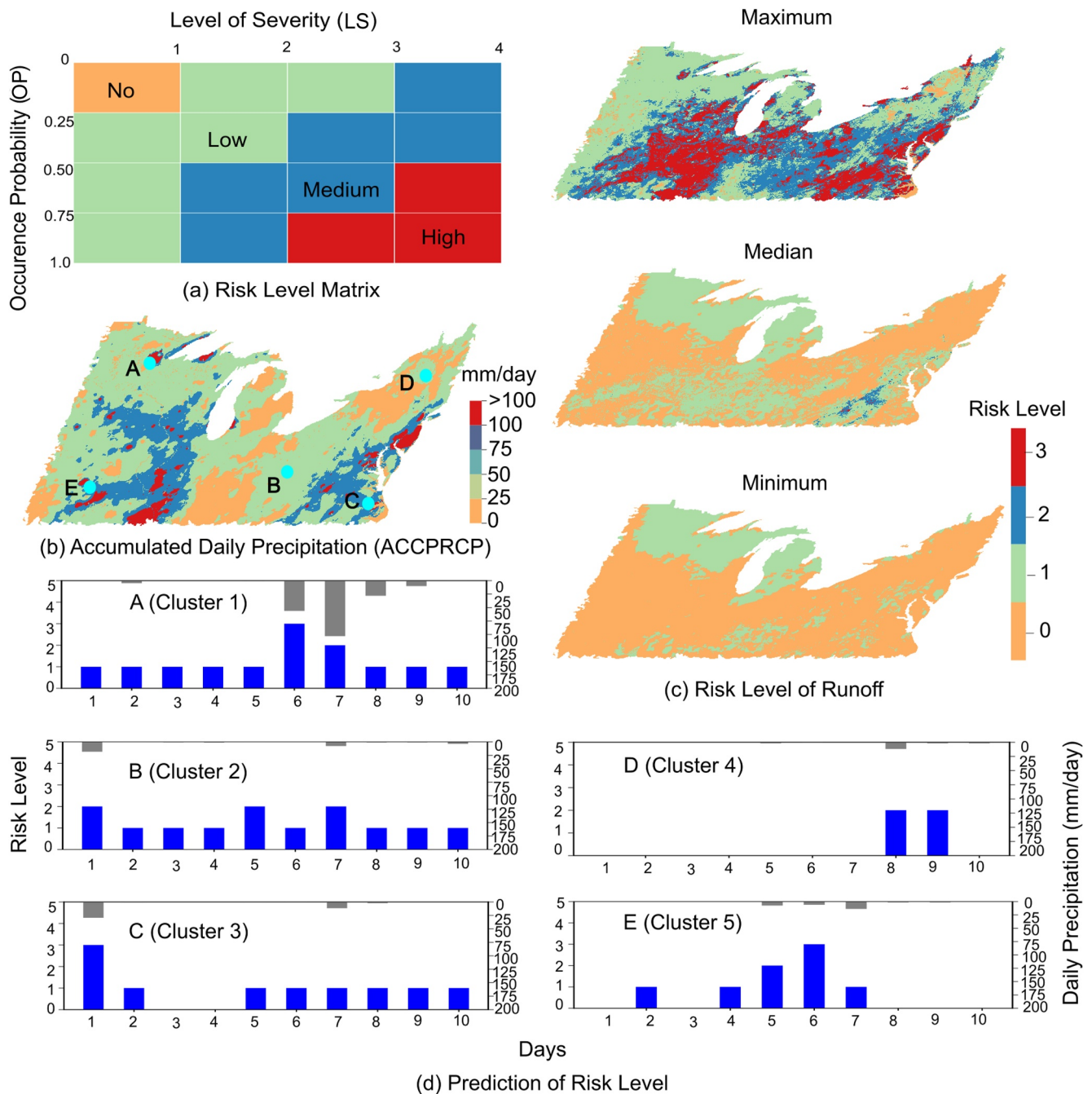
The remaining two clusters, while likely to still produce runoff from high precipitation events, will probably not have as much runoff generated as the other clusters. Cluster2 has moderate relief, but not as high as Cluster4. The cluster's heavily forested areas do receive a significant amount of precipitation each year, and the silty soils are thin, so the cluster will likely produce more runoff than the final cluster. Cluster5's flat topography, deep soils and relatively low precipitation amounts for the study area make it the least likely region for runoff to form despite its high clay content in the soil. However, Cluster5's heavy agriculture LULC makes this cluster particularly important for runoff-induced excess nutrient loads in surface waters. These nutrient loads may be exacerbated by the typical introduction of tile drainage to these agricultural fields with high clay contents and poor drainage. These two clusters are still located in the humid climate of the eastern US, so there will certainly be plenty of runoff generating events throughout the year.

### 3.2. Generalization of Statistical Models

#### 3.2.1. Prediction of Daily Runoff

For each cluster, XGBoost models were developed to predict the OP and magnitude of EOF runoff. When OP is greater than 0.5, the XGBoost predicts the occurrence of EOF runoff (1), and 0 otherwise. For the NWM, it predicts the occurrence of EOF runoff (1) only when the predicted magnitude is positive. For Cluster5, we performed 100 random 70%/30% splits by EOF sites. Given the split scenario that led to the worst performance by the XGBoost model in prediction of daily EOF runoff (i.e.,  $R^2 = 0.01$ ; Figure 3a), they still outperformed the NWM in prediction of occurrence and magnitudes of daily EOF runoff (Figures 3b and 3c). The results were consistent with the rest of the split scenarios as the predictions of XGBoost models improved, measured by  $R^2$  (Figures S1 and S2 in Supporting Information S1). For the other four clusters, similar results are found in Figures S3–S6 in Supporting Information S1.

Overall, the XGBoost models outperform the NWM in prediction of occurrence and magnitude of daily EOF runoff events for ungauged locations across clusters. This is because the XGBoost models are developed to address the insufficient representation of runoff generation process by the NWM (Hu et al., 2021). Meanwhile, the performance of the XGBoost relies on the quality of the NWM to represent the physical processes (e.g., tile drainage) critical for runoff generation, as well as the ability of the XGBoost models to use the causal outputs from the NWM. As the NWM further incorporates and improves the representation of these critical physical



**Figure 4.** (a) Risk level matrix: risk levels of daily edge-of-field runoff (No, Low, Medium, and High) determined by the occurrence probability and the level of severity for a given day. (b) Accumulated daily precipitation (mm) over 10 days from 20 to 29 December 2009, over the Great Lakes region. (c) Predicted risk level of daily runoff (minimum, medium, and maximum risk level across all 10 days) during the same period in the study area. (d) Daily precipitation (gray) and predicted risk level of daily runoff (blue) at five selected sites (one site for each cluster) during the study period.

processes, we can expect better prediction of the occurrence and magnitude of daily EOF runoff events by the NWM, further leading to better predictions by the XGBoost models.

### 3.2.2. Prediction of Runoff Risk Level

Based on the predictions of occurrence probability, OP and level of severity, LS of the daily EOF runoff by the XGBoost models (Table S6 in Supporting Information S1) for each cluster, we derived the risk level of daily runoff from 20 to 29 December 2009, over the Great Lakes region (Figures 4a and 4c). This period was chosen as

a test case to demonstrate the utility of the forecast during winter conditions when applied nutrients are vulnerable to transport due to snowmelt or rain-on-snow events following the application. The daily risk level varied spatially over the entire domain (Figure 4d) as precipitation accumulated over the 10-days period (Figure 4b). The high level of runoff risk mostly occurred in part of Clusters 2, 3, and 5 (Figure 4c) where there was a large amount of accumulated precipitation (Figure 4b). Additionally, the runoff risk level increased between day 6 and 8 when the precipitation mostly occurred (Figure 4d). This agrees with our expectation that precipitation is the major driving factor to runoff generation, and heavy precipitation often leads to high levels of runoff risk.

We also noticed that high levels of runoff risk did occur when there was moderate precipitation. This is consistent with snowmelt being another important factor that causes runoff during the winter season, which explains the non-zero risk level even without any precipitation (e.g., Site A, B, and C in Figure 4d). Additionally, geophysical, topographic variables can also play critical roles in runoff generation and determination of runoff risk level, such as LULC, soil content and condition and relief. Compared with Cluster2 which has predominant forest cover with moderate relief, Cluster3 contains highly urbanized, impervious areas which are more likely to generate runoff (Figure 4c). In fact, we did observe high levels of runoff risk in part of Cluster3 where only moderate precipitation occurred (Figure 4b). Similar to Cluster3, Clusters 1 and 4 generated medium and high levels of runoff risk even given low and moderate precipitation (Figure 4c). This finding agrees with our cluster analysis that Clusters 1 and 4 are likely to generate runoff given its soil condition and topography. Although the simulation length limits our ability to draw quantitative conclusions, it demonstrates the need for better measurements and predictions of the factors that can affect runoff generation for a given environment, as such factors are critical to the improvement of runoff risk prediction and forecast.

### 3.3. Limitation and Outlook

The accuracy of this study relies on several assumptions about the region and study period. While LULC is influential in runoff generation, changes to LULC are assumed to be negligible over the study period. It is also assumed that the 3 years of data used in the analyses are representative of those variables between the year 2002 and 2018. This assumption is necessary based on processing power and time constraints for analyzing such large data sets. In addition, many of the variables in the cluster analysis are derived from model outputs, which may introduce additional uncertainties. Overall uncertainties on the definition of clusters can propagate and further affect the generalization of statistical models. We thus plan to conduct sensitivity analysis with the cluster definition to understand the impact of uncertainty propagation on the predictions of risk level. Finally, the runoff risk level is derived based on the risk level matrix that comprises two components: how likely the runoff can occur and the level of severity (Figure 4a), which are subjective. Further advice will be solicited from the experienced stakeholders to better define the level of severity and risk level matrix.

## 4. Conclusions

In this paper, we proposed a new approach by combining cluster analysis and hybrid models (i.e., NWM and XGBoost models) to predict the risk level of daily runoff at a high spatial resolution to guide nutrient application over a large domain. By fusing the coarser, gridded NWM simulations with the field-scale measurements of EOF runoff, this approach generalizes the XGBoost models trained at field scales to larger regions with similar runoff potential. Many physics-based numerical models (e.g., regional landscape hydrologic models) are often developed and calibrated at a coarse spatial scale over a large domain. This approach would make a good candidate to enable such models to serve as useful tools for decision making that often occurs at a fine scale, especially when recalibration of these models is computationally prohibitive.

### Data Availability Statement

The data that support the findings of this study are openly available at the following URL/DOI: <https://doi.org/10.4211/hs.9460830270ec4d8b9d9c4260cca2114d>.



### Acknowledgments

This work was supported by the Great Lakes Restoration Initiative through the U.S. Environmental Protection Agency and National Oceanic and Atmospheric Administration. An award is granted to Cooperative Institute for Great Lakes Research (CIGLR) through the NOAA Cooperative Agreement with the University of Michigan (NA17OAR4320152). The NOAA GLERL contribution is No. 2010. We want to thank Laura Read, Arezoo RafieeiNasab and Aubrey Dugger for assisting with code development and questions related to the WRF-Hydro modeling system. We also thank the following agencies for providing us with daily EOF measurements, including USGS, USDA-ARS, Discovery Farms Minnesota, and Discovery Farms Wisconsin, and anonymous reviewers whose comments helped improve and clarify this manuscript.

### References

- Artoni, F., Delorme, A., & Makeig, S. (2018). Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition. *NeuroImage*, *175*, 176–187. <https://doi.org/10.1016/j.neuroimage.2018.03.016>
- Brooks, B. W., Lazorchak, J. M., Howard, M. D., Johnson, M. V. V., Morton, S. L., Perkins, D. A., et al. (2016). Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environmental Toxicology and Chemistry*, *35*(1), 6–13. <https://doi.org/10.1002/etc.3220>
- Carmichael, W. W., Azevedo, S. M., An, J. S., Molica, R. J., Jochimsen, E. M., Lau, S., et al. (2001). Human fatalities from cyanobacteria: Chemical and biological evidence for cyanotoxins. *Environmental Health Perspectives*, *109*(7), 663–668. <https://doi.org/10.1289/ehp.01109663>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).
- Ford, C. M., Kendall, A. D., & Hyndman, D. W. (2021). Snowpacks decrease and streamflows shift across the eastern US as winters warm. *Science of the Total Environment*, *793*, 148483. <https://doi.org/10.1016/j.scitotenv.2021.148483>
- Gildow, M., Aloysius, N., Gebremariam, S., & Martin, J. (2016). Fertilizer placement and application timing as strategies to reduce phosphorus loading to Lake Erie. *Journal of Great Lakes Research*, *42*(6), 1281–1288. <https://doi.org/10.1016/j.jglr.2016.07.002>
- Grannemann, N. G., Hunt, R. J., Nicholas, J. R., Reilly, T. E., & Winter, T. C. (2000). The importance of ground water in the Great Lakes Region. In *Water-resources investigations report 2000-4008*. US Geological Survey. <https://doi.org/10.1006/gcen.1999.7348>
- Hamlin, Q. F., Kendall, A. D., Martin, S. L., Whitenack, H. D., Roush, J. A., Hannah, B. A., & Hyndman, D. W. (2020). Quantifying landscape nutrient inputs with spatially explicit nutrient source estimate maps. *Journal of Geophysical Research: Biogeosciences*, *125*(2), e2019JG005134. <https://doi.org/10.1029/2019jg005134>
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. <https://doi.org/10.2307/2346830>
- Hopkins, B. G., & Hansen, N. C. (2019). Phosphorus management in high-yield systems. *Journal of Environmental Quality*, *48*(5), 1265–1280. <https://doi.org/10.2134/jeq2019.03.0130>
- Hu, Y. (2022). Edge of field runoff for the great lakes region [Dataset]. HydroShare. <https://doi.org/10.4211/hs.9460830270ec4d8b9d9c4260cca2114d>
- Hu, Y., Fitzpatrick, L., Fry, L. M., Mason, L., Read, L. K., & Goering, D. C. (2021). Edge-of-field runoff prediction by a hybrid modeling approach using causal inference. *Environmental Research Communications*, *3*(7), 075003. <https://doi.org/10.1088/2515-7620/ac0d0a>
- Kult, J. M., Fry, L. M., Gronewold, A. D., & Choi, W. (2014). Regionalization of hydrologic response in the Great Lakes basin: Considerations of temporal scales of analysis. *Journal of Hydrology*, *519*, 2224–2237. <https://doi.org/10.1016/j.jhydrol.2014.09.083>
- Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., et al. (2013). Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6448–6452. <https://doi.org/10.1073/pnas.1216006110>
- NOAA. (2016). NOAA launches America's first national water forecast model. Retrieved from <http://www.noaa.gov/media-release/noaa-launches-america-s-first-national-water-forecast-model>
- NOAA National Centers for Environmental Information (NCEI). (2021). Climate at a glance: Regional time series. Retrieved From <https://www.ncdc.noaa.gov/cag/>
- Paerl, H. W., & Otten, T. G. (2013). Harmful cyanobacterial blooms: Causes, consequences, and controls. *Microbial Ecology*, *65*(4), 995–1010. <https://doi.org/10.1007/s00248-012-0159-y>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Scavia, D., Bocaniov, S. A., Dagnew, A., Hu, Y., Kerkez, B., Long, C. M., et al. (2019). Detroit River phosphorus loads: Anatomy of a binational watershed. *Journal of Great Lakes Research*, *45*(6), 1150–1161. <https://doi.org/10.1016/j.jglr.2019.09.008>
- Seaber, P. R., Kapinos, F. P., & Knapp, G. L. (1987). *Hydrologic unit maps*. US Government Printing Office. (Vol. 2294, p. 1987). Retrieved from <https://water.usgs.gov/GIS/huc.html>
- Soller, D. R. (2001). *Map showing the thickness and character of quaternary sediments in the glaciated United States East of the Rocky Mountains*. U.S. Geological Survey. <https://doi.org/10.3133/i1970E>
- Soller, D. R., Reheis, M. C., Garrity, C. P., & Van Sistine, D. R. (2009). *Map database for surficial materials in the coterminous United States*. U.S. Geological Survey. (p. 425). Retrieved from <https://pubs.usgs.gov/ds/425/>
- Stackpoole, S. M., Stets, E. G., & Sprague, L. A. (2019). Variable impacts of contemporary versus legacy agricultural phosphorus on US river water quality. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(41), 20562–20567. <https://doi.org/10.1073/pnas.1903226116>
- United States Geological Survey (USGS). (2011). NLCD 2011 land cover coterminous United States. Retrieved from <https://mrlc.gov/data>
- Wang, Y., Liu, Y., Khan, F., & Imtiaz, S. (2017). Semiparametric PCA and Bayesian network based process fault diagnosis technique. *Canadian Journal of Chemical Engineering*, *95*(9), 1800–1816. <https://doi.org/10.1002/cjce.22829>
- Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*(D3). <https://doi.org/10.1029/2011jd016048>
- Prism climate group. (2018). Oregon State University. Retrieved from <http://prism.oregonstate.edu>

### References From the Supporting Information

- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, *37*(3), 424–438. <https://doi.org/10.2307/1912791>
- Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1–3), 271–280. [https://doi.org/10.1016/s0378-4754\(00\)00270-6](https://doi.org/10.1016/s0378-4754(00)00270-6)