2016

Environmental informatics

Noel Cressie
University of Wollongong

Sandy Burden
University of Wollongong

Clint Shumack
University of Wollongong

Andrew Zammit-Mangion
University of Wollongong

Bohai Zhang
University of Wollongong

Environmental informatics

Abstract

Environmental Informatics uses a panoply of tools from the Statistics, Mathematics, Computing, and Visualisation disciplines. It uses these tools to reveal, quantify, and validate scientific hypotheses in the environmental sciences, with the quantification of uncertainty central to its approach. There is now a strong recognition that scientific models need to incorporate stochastic components throughout: While it has always been recognised that data have a component of measurement error, attention is now being given to the quantification of model error, and it is becoming accepted by environmental scientists that probability models for the latter allows for a coherent way to make scientific inference. In Environmental Informatics, uncertainty may be assigned not only to datasets of measurements, but also to computer-generated climate-model output. Methodological advances, in the form of hierarchical statistical models and the accompanying computational developments, have expanded the scope of statistical analyses into very large spatial domains. This has led to studies of the dynamical evolution of entire spatial fields of geophysical variables, where results are given in terms of predictive distributions. Environmental Informatics is not only involved in characterising the environment, it can also be used to make decisions about mitigation and adaptation strategies. The steps taken by environmental scientists, from data to information, from information to knowledge, and from knowledge to decisions, are all taken in the presence of uncertainty. Environmental Informatics encompasses all these aspects.

NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

09-16

Environmental Informatics

Noel Cressie, Sandy Burden, Clint Shumack,

Andrew Zammit-Mangion, and Bohai Zhang

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Wiley StatsRef: Environmental Informatics

Noel Cressie[1*], Sandy Burden[1], Clint Shumack[1], Andrew Zammit-Mangion[1], and Bohai Zhang[1]

[1]Centre for Environmental Informatics, National Institute for Applied Statistics Research Australia, University of Wollongong, NSW, Australia

August 30, 2016

Abstract

Environmental Informatics uses a panoply of tools from the Statistics, Mathematics, Computing, and Visualisation disciplines. It uses these tools to reveal, quantify, and validate scientific hypotheses in the environmental sciences, with the quantification of uncertainty central to its approach. There is now a strong recognition that scientific models need to incorporate stochastic components throughout: While it has always been recognised that data have a component of measurement error, attention is now being given to the quantification of model error, and it is becoming accepted by environmental scientists that probability models for the latter allows for a coherent way to make scientific inference. In Environmental Informatics, uncertainty may be assigned not only to datasets of measurements, but also to computer-generated climate-model output. Methodological advances, in the form of hierarchical statistical models and the accompanying computational developments, have expanded the scope of statistical analyses into very large spatial domains. This has led to studies of the dynamical evolution of entire spatial fields of geophysical variables, where results are given in terms of predictive distributions. Environmental Informatics is not only involved in characterising the environment, it can also be used to make decisions about mitigation and adaptation strategies. The steps taken by environmental scientists, from data to information, from information to knowledge, and from knowledge to decisions, are all taken in the presence of uncertainty. Environmental Informatics encompasses all these aspects.

Keywords— Spatial statistics, decision-making, hierarchical statistical modelling, sea surface temperature, Big Data, remote sensing, Bayesian statistics

*ncressie@uow.edu.au

1 Introduction

Environmental Informatics uses tools from the Statistics, Mathematics, Computing, and Visualisation disciplines to reveal, quantify, and validate scientific hypotheses in the environmental sciences[1]. There are a number of names for the incorporation of statistical thinking, modelling, and analysis in the environmental sciences: Environmental Statistics has taken on a meaning that involves statistical analysis for local or regional environmental characterisations and remediations, although not exclusively so[2]. Environmetrics has become known for encompassing statistical analysis of the environment at all scales from local to global; The International Environmetrics Society (TIES) is a section of the International Statistical Institute with its own journal, *Environmetrics*. Terms like Uncertainty Analysis and Uncertainty Quantification have become synonyms for Statistics in many of the sciences, including the environmental sciences.

While it has always been recognised that data have measurement error associated with them, attention is now being given to the quantification of model error, and it is becoming accepted that probability models for the latter allow for a coherent way to make scientific inference. That uncertainty may even be assigned to computer-generated climate-model output[3].

In the last decade, computational advances and methodological advances in the form of hierarchical statistical models have expanded the scope of statistical analyses into large spatial domains, particularly the dynamical evolution of entire spatial fields of geophysical variables (e.g., see Ch.7 of the 2011 book by Cressie and Wikle[4]). An early example is a spatio-temporal hierarchial statistical model for forecasting sea surface temperatures (SSTs) in the tropical Pacific Ocean[5], which is used in this article to illustrate the steps in Environmental Informatics (EI).

EI is data driven, but the quantification of uncertainty is central to its approach. It looks where it can for technology, particularly in the fertile discipline known as Machine Learning, while linking the algorithms back to an implied statistical model. The 2009 book by Hastie, Tibshirani, and Friedman[6] shows the value of such a strategy. Bioinformatics broke free of its biostatistical bonds, and Environmental Informatics is doing the same from environmental statistics.

EI is not only involved in characterising the environment, it can also be used to make decisions about mitigation and adaptation strategies, critically in the presence of uncertainty. The organisation of this article is along the lines of a conceptual pyramid with data at its base and steps from data to information, from information to knowledge, and from knowledge to decisions[4]. As one climbs the pyramid, the uncertainty is crystallised but never eliminated. At the top of the pyramid, where decisions are made, the use of gain functions allows competing decisions (that includes the decision of doing nothing at all) to be compared in a rational, quantitative manner.

2 From Data to Information

The pyramid referred to in Section 1 has "data" at its base. The El Niño phenomenon, of warmer waters pooling off the coast of Peru, started with anecdotal data from local fishermen who noticed that it spanned the Christmas period during those years when it occurred ("El Niño" means the Christ Child). Now, through ship tracks, ocean buoys, and more recently remote sensing, the climatological community has data that give them an oceanic view of decadal oscillations of warmer and cooler water pooling in the Eastern Tropical Pacific Ocean[7]. El Niño plays an important role in the world's weather, but particularly so for weather patterns in Australia and the Americas.

Automation, micro-devices, and space-based sensors have, in recent decades revolutionised data-collection methods for the environmental sciences. Different instruments and collection mechanisms generate different types of data, which can usually be classified into (i) point-referenced data, (ii) areal data, (iii) point-pattern data, and (iv) trajectory data. The first three are discussed at length in Cressie (1993)[8], and the fourth is becoming more common with the use of micro-devices that can be attached to animals to study their migration patterns or placed on buoys to study the ocean circulation system.

EI is poised to take advantage of the data revolution; it has innovative statistical methodology at its core that draws from environmental datasets whose sizes are often so large that they require a database distributed across multiple computers in a cluster, to achieve scalability and load-balancing. For example, climate models draw on databases that include meteorology, geography, elevation, surface type, atmospheric particles (clouds, aerosols), and so forth. These Big Data require special tools. The Structured Query Language (SQL[9]) is a special-purpose programming language designed for defining, manipulating, and controlling tabled data (see Database Systems). The Hadoop distributed file system (HDFS) can store petabytes of data across multiple machines. HBase is a distributed database that runs on top of HDFS, providing SQL-like capabilities for querying very large tables of data. Algorithms to carry out analyses are now being re-designed to "go to the data" by querying databases located in remote data centres.

"Information" comes in the form of subsetting and aggregation, amongst other operations on the data, and it should be driven by Exploratory Data Analysis (EDA). EDA's primary purpose is to help elicit scientific hypotheses about observed phenomena and to indicate which statistical modelling tools are suitable for a full analysis[10]. Popular methods for environmental data are summarised in Cressie and Wikle (2011), Ch.5[4] and include: summary statistics (mean, median, quantiles, etc.), scatter plots to display the relationship between responses and covariates, plots that collapse space and/or time to reveal spatial covariances and/or temporal dynamics; space-time plots such as Hovmöller diagrams[11], empirical covariance and cross-covariance plots; and dimension-reduction tools such as Empirical Orthogonal Functions (EOFs) and Canonical Correlation Analysis (CCA). The authors then applied these methods to tropical Pacific SSTs and the El Niño phenomenon. In particular, a few leading EOFs reduced the dimension of the large, tropical Pacific Ocean SST dataset and preserved most of the variability. The remaining variability may also be important, and keeping it in the model preserves variability balance, the statistical analogue of mass

balance.

For very large spatial datasets, classical EOFs[4,12] can extract important patterns from the data, yet its eigenfunctions are often very "noisy" or exhibit patterns without physical meaning. In order to improve the identification of important spatial patterns from the noisy data, there are generalisations of the classical approach that are obtained by imposing sparseness and smoothness constraints on the eigenfunctions. For example, Wang and Huang (2016)[13] apply a regularised Principal Component Analysis method to modelling SSTs in the Indian Ocean, resulting in a much better estimate of the spatial covariance matrix for extraction of EOFs.

A key component of EDA is visualisation, which is needed at both the very early stages, straight after data collection, and at the final stages of the analysis. Initially, it provides a means of displaying the raw data to the analyst, in a form that is clear, effective, and able to show outlying or anomalous features. When the anomaly seen in the visualisation is deemed to be physically plausible, one might ask: Which covariates might explain the observed anomaly? If the anomaly is not plausible and its provenance is untraceable, the offending data may be set aside. Visualisation is also a powerful tool for hypothesis generation.

In the final stages of a scientific analysis, visualisation of results is important to (i) provide a reality check that the results make sense to a domain expert, and (ii) convey a conclusion to a non-specialist audience, such as decision-makers or the general public. Importantly, the visualisations should also convey uncertainty. Techniques that do this range from plots of prediction standard errors[14] to plots of diagnostics such as residuals at validation-data locations[8,15,16]. Overfitting a model can yield incorrect and even dangerous conclusions, which is where innovative diagnostics and visualisations can play an important role.

There are several open-source software packages that can be used for EDA. These include *R Software*[17], accompanying visualisation packages[18] (see Statistical Graphics, stat07368), and web-based interactive tools that can provide a deeper engagement with the user. When a dataset is too large for direct visualisation, which is becoming increasingly common, *cognostics* are sometimes used as a divide-and-conquer approach[19]. Visualisation using cognostics is facilitated with the *Trelliscope* tool[19].

Because environmental datasets have a strong spatial component, it is natural to use a Geographic Information System (GIS), particularly for the visualisation component of EDA. A GIS is designed to manage spatial databases, to extract information from them, and to visualise the results. *ArcGIS* is a proprietary GIS that has become a standard in mapping and map-querying applications. It is used by many companies and government agencies, for projects ranging from flood-prevention management to the mapping of Earth observations (see ArcGIS, http://www.arcgis.com/home/gallery.html, for an extensive list).

Machine Learning[20] is a subfield of computer science, aiming to construct algorithms that can learn from data and make precise predictions. The algorithms are largely classified as either *supervised learning* or *unsupervised learning*, and learning paradigms as *active* or *passive*. Machine learning provides rich tools for modelling environmental datasets, especially for interpolating and downscaling spatial data. For example, creating a dataset like the tropical Pacific Ocean SSTs on a regular $1^\circ \times 1^\circ$ grid involves tools that combine ship tracks,

buoys, and remote sensing, taken at different spatial and temporal resolutions. These tools often have a basis in statistical thinking[6].

The next section shows how one can go from information to knowledge. The "information" stage between "data" and "knowledge" in EI is deliberate, since it emphasises and crystallises uncertainty quantification along the way. For some purposes, having statistical summaries and evocative visualisation is enough, but scientific inference needs more.

3 From Information to Knowledge

Environmental Informatics for global phenomena will typically involve very large datasets, that are often spatio-temporal and initially can be hard to make sense of. The methods described in Section 2 give us the understanding to build statistical models that attempt to capture the underlying "signal" (Y) from the data (Z), where the "noise" is described by random variation. "Knowledge" comes in the form of inference on Y and on parameters θ in the hierarchical-statistical model given by (2) and (3) below.

Suppose that the data are Z and the model is written as $[Z \mid \theta]$, where $[A]$ denotes the distribution of the random quantity A, and $[A \mid B]$ denotes the conditional distribution of A given B. Further, suppose that the model parameters θ are unknown and need to be estimated. Central to this is the likelihood,

$$L(\theta; Z) \equiv [Z \mid \theta], \tag{1}$$

which is the joint probability distribution of the data, considered as a function of θ. This direct approach to modelling the marginal distribution of the data Z, is not very helpful when trying to infer the true underlying environmental process, namely the "signal" Y .

A conditional-distributional approach is to specify

$$\text{Data model:} \qquad [Z \mid Y, \theta] = [Z \mid Y, \theta_D] \tag{2}$$

$$\text{Process model:} \qquad [Y \mid \theta] = [Y \mid \theta_P], \tag{3}$$

where $\theta = (\theta_D, \theta_P)$. This is a hierarchical statistical model where there are two unknowns, Y and θ; usually, Y represents the scientific process being studied. For example, Y is the true SST in the tropical Pacific at resolutions finer than $1^\circ \times 1^\circ$, and Z is the dataset released for a given month at a $1^\circ \times 1^\circ$ resolution[5]. Notice that Y is "hidden" in the likelihood given by (1), as the following marginalisation operation shows:

$$L(\theta; Z) = \int [Z \mid Y, \theta_D][Y \mid \theta_P] \mathrm{d}Y.$$

Hypothesis testing is a fundamental part of EI. Detection and attribution of anthropogenic effects on climate change can be investigated by first testing the null hypothesis that there is no change. Temperature is an important component of climate; hence, one alternative hypothesis may be that the mean ambient air temperature (in North America, say) between 1990 and 1999 was higher than that between 1980 and 1989. The two mean

temperatures will be different but, based on climate data, hypothesis testing may be used to determine whether the null hypothesis of no difference is true, or whether the temperature difference is significant. Rejection of the null hypothesis can be treated as evidence of a mean temperature change between the two decades for the study region. Subsequent hypotheses can be formulated for smaller subregions in order to detect "hot spots" with significant temperature changes[21].

Alternatively, the null hypothesis may be that an environmental process (e.g., temperature change) exceeds some threshold value μ (e.g., 0.25°C). If the problem is formulated spatially, then a variety of local hypotheses could be tested, and the collection of accepted hypotheses would make up an estimated exceedance region. When testing multiple hypotheses, special care needs to be taken to ensure that the joint Type I error rate of testing many hypotheses is controlled at level α. Enhanced False Discovery Rate (FDR) methodology[21] or conditional-simulation and testing[22,23] achieve this.

Using Bayes' Rule, the predictive distribution for the process Y defined in (2) and (3) is given by

$$[Y \mid Z, \theta] = [Z \mid Y, \theta][Y \mid \theta]/[Z \mid \theta], \tag{4}$$

where recall that θ is made up of data-model parameters θ_D and process-model parameters θ_P, both fixed but unknown. A Bayesian hierarchical model would specify a parameter model (or prior), $[\theta]$, and then the predictive distribution becomes $[Y \mid Z]$; for more details, see Ch.2 of Cressie and Wikle (2011)[4] and (6) below.

Notice that (4) gives a complete distribution, called the predictive distribution, which can be summarised, for example, with its first two moments. However, in Section 4 it is made clear that the most appropriate summary depends on the environmental question being asked. EI is particularly interested in extreme events, not only their magnitude but where they are (or will be) located. The answer to "How extreme?" and "Where?" can be found somewhere in the predictive distribution given by (4), but it is not found in the predictive mean, which is too smooth for making inference on extremes.

The known uncertainties, including those on the parameters, can be expressed through a hierarchical statistical model. From (2) and (3) and a parameter model $[\theta] = [\theta_D, \theta_P]$, the joint probability distribution is:

$$[Z, Y, \theta_D, \theta_P] = [Z \mid Y, \theta_D][Y \mid \theta_P][\theta_P, \theta_D], \tag{5}$$

which controls all the inferences in the problem. Hence, the predictive distribution can be obtained from marginalisation of (5) over θ and then using Bayes' Theorem:

$$[Y \mid Z] = \frac{\int\int [Z \mid Y, \theta_D][Y \mid \theta_P][\theta_P, \theta_D] \mathrm{d}\theta_P \mathrm{d}\theta_D}{[Z]}, \tag{6}$$

where $[Z]$ is the so-called normalising constant obtained from marginalisation of (5) over both θ and Y.

Obtaining the predictive distribution given by (4) or (6) is often achieved via Markov chain Monte Carlo (MCMC) algorithms, where the goal is to sample from the predictive distribution and to use empirical summaries of those samples to approximate the distributional

summaries[24]. Other statistical-computing methodologies, such as importance sampling and Integrated Nested Laplace Approximations (INLA)[25], may have difficulty handling predictive inference when both the parameter space and the prediction space are large. MCMC is in principle extendable to models with more unknowns by adding more full conditionals to the Gibbs sampler (a form of MCMC algorithm), although perhaps with considerable difficulty.

The methodology based on hierarchical statistical modelling, and its predictive distribution given by (4) or (6), provides a compelling framework for scientific inference (prediction and hypothesis testing). Spatio-temporal processes can now be predicted at large regional and global scales[26,27,28]. Hierarchical statistical modelling relies on the model specifications being appropriate at each level of the hierarchy. Ultimately, one has only the data Z to check the goodness-of-fit of the model and to diagnose any of its shortcomings. In a spatial setting, model-fit can be applied globally or locally. Local Indicators of Spatial Association (LISAs) typically decompose a global diagnostic into constituent parts that indicate small regions of the spatial field that do not fall in line with the model. Cross-validation is also local, in that the model is fitted when a spatial cluster of observations is left out, predictions are made at the locations of the deleted observations, and goodness-of-fit statistics computed and assessed. While cross-validation is often considered a gold standard for diagnostics[29,30,31], it is computationally expensive and may be impractical for very large datasets. Alternatives such as "testing" datasets[32,33], importance sampling[29], simulation-based model checking[34], posterior predictive checks[35,36], and approaches that balance bias with the computational burden of cross-validation[37,38] may also be used.

Inference (hypothesis testing or prediction) is typically needed for answering a number of questions that decision-makers might have. The answers are contained somewhere in the predictive distribution, and the next section addresses how they might be extracted.

4 From Knowledge to Decisions

"Decisions" are actions taken by policy makers based on inferences (knowledge). A hierarchical statistical model expresses uncertainty probabilistically, so any action taken that could affect the environmental process Y should be made in the presence of this uncertainty. For example, if farmers growing corn in Iowa, USA, want to buy crop insurance against drought, the cost of their policy is determined by, amongst other things, a long-range weather forecast, which is determined to a large extent by projections of the El Niño phenomenon six months into the future. Insurance companies insure the individual farmers, but they make their calculations of policy premiums based on the pool of clients they insure, the projected probability of drought, and the financial losses that could occur.

EI quantifies how the probabilities of outcomes and the consequences of those outcomes. can be combined into an optimal decision-making strategy. Let $\hat{Y}(Z)$ be one of many decisions about Y based on Z. Some decisions are better than others, which can be quantified through a gain function, $G(Y, \hat{Y}(Z))$. The Bayes expected gain is $E(G(Y, \hat{Y}))$, and this is maximised with respect to $\hat{Y}$. Then, it is a consequence of decision theory (e.g., Berger

(1985)[39]), that the optimal decision is:

$$Y^*(Z) = \arg\sup_{\hat{Y}} \{E(G(Y, \hat{Y}) \mid Z)\}. \tag{7}$$

Now suppose that there is scientific interest in a summary $h(Y)$ of Y (e.g., regional averages, or regional extremes). Then an *equivariance property* of sampling implies that samples from $[h(Y) \mid Z]$ are obtained by sampling from $[Y \mid Z]$ and then simply evaluating each member of the sample at $h(\cdot)$. This equivariance property is enormously powerful, even more so when the sampling does not require knowledge of the normalising term $[Z]$ in (6) (e.g., when using MCMC).

Which summary of the predictive distribution $[h(Y) \mid Z]$ will be used to estimate the scientifically interesting quantity $h(Y)$? Too often, the posterior mean,

$$E(h(Y) \mid Z) = \int h(Y)[Y \mid Z]\mathrm{dY},$$

is chosen as a "convenient" estimator of $h(Y)$. It is an optimal estimator when the gain function is "negative squared-error." However, this gain function assumes equal consequences for under-estimation and over-estimation, which is not realistic when considering extreme events such as droughts or floods. When a science or policy question is about extreme events, asymmetric loss functions should be used[40]. The lack of certainty around making a policy decision should not result in a lack of action, since no action is itself a decision $\hat{Y}^0$ that has its own gain, $G(Y, \hat{Y}^0)$. Should the optimal decision (7) be too hard to determine numerically, the posterior expected gain, $E(G(Y, \hat{Y})|Z)$, can be computed for a number, K, of possible decisions, $\hat{Y}^0, ..., \hat{Y}^K$. Then $\hat{Y}^*$ is chosen as the one that achieves the maximum expected posterior gain. The posterior expected gain is simply a summary of the predictive distribution $[Y \mid Z]$, which we saw in Section 3 relies on the ability to build a hierarchical statistical model and to sample from its predictive distribution.

5 Conclusion

A pyramid of steps, from data to information, from information to knowledge, and from knowledge to decisions, shows how Big Data can be crystallised into Big Understanding and Wise Decisions. The approach is not unique to EI, but what to do about the present and future state of our environment makes it particularly pertinent at this time. This crystallisation is possible because the probabilistic approach allows uncertainty to be apportioned to components that we can harness and components that we can only describe. The power of a statistical approach is to determine how to use the first components in the presence of the uncertainty engendered by both. The cornerstones of Environmental Informatics are the hierarchical statistical model, computational procedures to sample from its predictive distribution, and gain functions that are expressions of the questions that policy-makers are asking.

6 Acknowledgements

This research was partially supported by a 2015-2017 Australian Research Council Discovery Grant, number DP150104576 (Cressie).

7 Related Articles

See also Data Mining; Decision Support Systems, Environmental; Environmental Statistics; Environmetrics; Geographic Information Systems (GIS), Spatial Statistics in; Hierarchical Bayesian Space-Time Analysis; Hypothesis Testing; Massive Data, Models for; Parallel Computing: Statistical and Environmetric Uses; Uncertainty Analysis; Uncertainty and Computer Models; Understanding Large-Scale Structure in Massive Data Sets.

8 References

[1] Cressie, N. (2014) Environmental informatics: Uncertainty quantification in the environmental sciences, in *Past, Present, and Future of Statistical Science* (eds X. Lin, C. Genest, D.L. Banks, G. Molenberghs, D.W. Scott, and J.L. Wang), CRC Press, Boca Raton, FL, pp. 429–449.

[2] Barnett, V. (2004) *Environmental Statistics: Methods and Applications*, Wiley, Chichester, UK.

[3] Kang, E.L., Cressie, N., and Sain, S.R. (2012) Combining outputs from the North American Regional Climate Change Assessment Program by using a Bayesian hierarchical model. *Journal of the Royal Statistical Society, Series C*, **61**, 291–313.

[4] Cressie, N. and Wikle, C. (2011) *Statistics for Spatio-Temporal Data*, Wiley, Hoboken, NJ.

[5] Berliner, L.M., Wikle, C.K., and Cressie, N. (2000) Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate*, **13**, 3953–3968.

[6] Hastie, T., Tibshirani, R., and Friedman, J.H. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn.*, Springer, New York, NY.

[7] NOAA (2016), NOAA NCEP EMC CMB GLOBAL Reyn_SmithOIv2 monthly sst: Sea Surface Temperature data. URL `http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.EMC/.CMB/.GLOBAL/.Reyn_SmithOIv2/.monthly/.sst/`.

[8] Cressie, N. (1993) *Statistics for Spatial Data, rev. edn.*, John Wiley & Sons, New York, NY.

[9] Groff, J.R., Weinberg, P.N., and Oppel, A.J. (2010) *SQL, The Complete Reference, 3rd edn.*, McGraw-Hill, New York, NY.

[10] Tukey, J.W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

[11] Hovmöller, E. (1949) The Trough-and-Ridge diagram. *Tellus*, **1**, 62–66.

[12] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441.

[13] Wang, W.T. and Huang, H.C. (2016) Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics*, **doi: 10.1080/10618600.2016.1157483**, 1–30.

[14] Rougier, J. and Zammit-Mangion, A. (2016) Visualisation for large-scale Gaussian updates. *Scandinavian Journal of Statistics*, **doi: 10.111/sjos.12234**.

[15] Sahu, S.K. and Mardia, K.V. (2005) A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society, Series C*, **54**, 223–244.

[16] Bastos, L.S. and O'Hagan, A. (2009) Diagnostics for Gaussian process emulators. *Technometrics*, **51**, 425–438.

[17] R Core Team (2016) R: A Language and Environment for Statistical Computing, Vienna, Austria.

[18] Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*, Springer-Verlag, New York, NY.

[19] Hafen, R., Gosink, L., McDermott, J., Rodland, K., Dam, K.V., and Cleveland, W.S. (2013) Trelliscope: A system for detailed visualization in the deep analysis of large complex data, in *Proceedings of the IEEE Symposium on Large Data Analysis and Visualization 2013* (eds B. Geveci, H. Pfister, and V. Vishwanath), IEEE, Piscataway, NJ, pp. 105–112.

[20] Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**, 199–231.

[21] Shen, X., Huang, H.C., and Cressie, N. (2002) Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association*, **97**, 1122–1140.

[22] French, J.P. and Sain, S.R. (2013) Spatio-temporal exceedance locations and confidence regions. *The Annals of Applied Statistics*, **7**, 1421–1449.

[23] French, J.P. and Hoeting, J.A. (2016) Credible regions for exceedance sets of geostatistical data. *Environmetrics*, **27**, 4–14.

[24] Brooks, S., Gelman, A., Jones, G., and Meng, X.L. (eds) (2011) *Handbook of Markov Chain Monte Carlo*, CRC Press/Taylor & Francis, Boca Raton, FL.

[25] Rue, H., Martino, S., and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **71**, 319–392.

[26] Wikle, C.K., Milliff, R.F., Nychka, D., and Berliner, L.M. (2001) Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, **96**, 382–397.

[27] Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014) Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, **56**, 174–185.

[28] Cressie, N., Shi, T., and Kang, E.L. (2010) Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, **19**, 724–745.

[29] Stern, H.S. and Cressie, N. (2000) Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, **19**, 2377–2397.

[30] Gelfand, A.E., Dey, D.K., and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods, in *Bayesian Statistics 4* (eds J.M. Bernardo, J.O. Berger, A.P. Dawid, and A. Smith), Oxford University Press, Oxford, UK, pp. 147–167.

[31] Marshall, E.C. and Spiegelhalter, D.J. (2003) Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649–1660.

[32] Efron, B. (1983) Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316–331.

[33] Efron, B. (1986) How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, **81**, 461–470.

[34] Dey, D., Gelfand, A., Swartz, T., and Vlachos, P. (1998) A simulation-intensive approach for checking hierarchical models. *Test*, **7**, 325–346.

[35] Gelman, A., Meng, X.L., and Stern, H.S. (1996) Posterior predictive asssessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–807.

[36] Marshall, E.C. and Spiegelhalter, D.J. (2007) Identifying outliers in Bayesian hierarchical models: A simulation-based approach. *Bayesian Analysis*, **2**, 409–444.

[37] Bayarri, M.J. and Berger, J.O. (2000) P-values for composite null models. *Journal of the American Statistical Association*, **95**, 1127–1142.

[38] Bayarri, M.J. and Castellanos, M.E. (2007) Bayesian checking of the second levels of hierarchical models. *Statistical Science*, **22**, 322–343.

[39] Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis, 2nd edn.*, Springer, New York.

[40] Zhang, J., Craigmile, P.F., and Cressie, N. (2008) Loss function approaches to predict a spatial quantile and its exceedance region. *Technometrics*, **50**, 216–227.