Assessing Quality of Spatial Models using the Structural Similarity Index and Posterior Predictive Checks

Colin Robertson[1], Jed A. Long[2], Farouk S. Nathoo[3], Trisalyn A. Nelson[2], Cameron C.F. Plouffe[1]

[1]Department of Geography & Environmental Studies, Wilfrid Laurier University, 75 University Ave West, Waterloo, ON, N2L 3C5, Canada, [2]Spatial Pattern Analysis & Research (SPAR) Laboratory, Department of Geography, University of Victoria, PO Box 3060, Victoria, BC V8W 3R4, Canada, [3]Department of Mathematics & Statistics, University of Victoria, PO Box 3060, Victoria, BC V8W 3R4, Canada

Correspondence: Colin Robertson, Department of Geography & Environmental Studies, Wilfrid Laurier University, 75 University Ave West, Waterloo, ON, N2L 3C5, Canada
e-mail: crobertson@wlu.ca

Pre-print of published version.

Reference:
Robertson, C, JA Long, FS Nathoo, TA Nelson, CCF Plouffe. 2014. Assessing quality of spatial models using the structural similarity index and posterior predictive checks. Geographical Analysis. 46(1). 53-74.

DOI:
http://dx.doi.org/10.1111/gean.12028

Abstract

Model assessment is one of the most important aspects of statistical analysis. In geographical analysis, models represent spatial processes, where variability in mapped output results from uncertainty in parameter estimates. Slight spatial misalignments can cause inflated error scores when comparing maps of observed and predicted variables using traditional error metrics at the level of individual spatial units. We conceptualize spatial model assessment as a continuous value map comparison problem, and employ methods from image analysis to score model outputs. The structural similarity index, a measure that attempts to replicate the human visual system using a local region approach is used as an exploratory map comparison statistic. The measure is implemented within a Bayesian spatial modelling framework, as a discrepancy measure in a posterior predictive check of model fit. Results are reported for simulation studies representing a variety of spatial processes in a spatial and space-time context. A case study of rainfall mapping in Sri Lanka demonstrates the proposed methodology applied to assessment of Bayesian kriging interpolations. Both simulation studies as well as the case study demonstrate that the approach reveals hidden spatial structure not uncovered by traditional methods. The spatially sensitive assessment methodology provides a diagnostic tool to support spatial modelling and analysis.

Introduction

Models of spatially varying phenomena commonly incorporate spatially distributed model parameters to account for spatial variation in the relationship between a process and modeled or missing covariates (Finley 2011). For example, space-time models of dynamic processes such as infectious disease or invasive species spread, utilize spatially distributed model parameters to account for spatial heterogeneity in epidemic waves (e.g., Smith et al. 2002, Wheeler and Waller 2008). Statistical models with spatially-varying coefficients can accommodate non-stationary relationships in a regression context, which is particularly useful for large study areas. A key advantage of spatially explicit modelling for geographical analysis is that parameter estimates can be mapped and integrated with other spatial data sets. The most common approaches for modelling spatially heterogeneous processes include Bayesian hierarchical modelling with dynamic spatial models (Gelfand et al. 2003), geographically weighted regression (GWR) (Brundson, Fotheringham, and Charlton 1996) and spatial filtering methods (Griffith 2008).

A key difficulty in developing spatial models is model assessment, which comprises two stages: the relative comparison between candidate models in development in order to assist in model selection, and the characterization of discrepancies between modeled outcomes and observed data in order to facilitate further model development, data collection and hypothesis generation. In practice, these stages often occur iteratively. Traditionally metrics employed in the model selection phase simultaneously consider a quality of fit statistic (such as the deviance) along with some penalty for model complexity to discourage overfitting (e.g., AIC–Akaike 1973, BIC–Schwarz 1978, DIC–Spiegelhalter et al. 2002, Bayes Factors–Kass and Raftery 1995). By definition, the model selection stage is relative, and provides evidence about the performance of a given model in relation to others being considered–that is, model selection indices are used to

determine the best model among a suite of candidate models, or to combine relative strengths of different models through model averaging (Burnham and Anderson 2002). The model selection stage provides little information about the overall quality of proposed models, requiring alternative objective approaches for examining goodness-of-fit. For a recent example, Finely et al. (2012) use a squared error loss function that incorporates both discrepancy in the estimated values (i.e., mean) and uncertainty (i.e., variance) in predictions as model selection criteria within a posterior predictive model-checking framework.

The second stage of model assessment examines how well a selected candidate model agrees with data in order to quantify goodness-of-fit. A more expansive treatment of model residuals is typically conducted to quantify their magnitude and structure in a separate stage through goodness-of-fit tests. Typical approaches for measuring goodness-of-fit include a comparison of observed and estimated values, such as the chi-squared (χ^2) test (e.g., Dice 1945), the root mean squared error, the mean absolute error (i.e., bias), or kappa statistics (e.g., Carletta 1996). These tests are undertaken in geographical analysis any time a researcher has observed and predicted values (e.g., spatial interpolation, hierarchical modelling, image classification). Because these methods are employed so widely in many areas of geography, we use the term 'error' here interchangeably with 'residual' rather than the more precise definition related to the difference between an estimate and its true parameter value.

With analysis of spatial data, aspatial goodness-of-fit measures can be misleading, especially in the presence of spatially dependent error structures (Lee and Ghosh 2009). Analysis of the spatial structure of regression residuals has a long history in geographical analysis (e.g., Thomas 1968). Based on the assumption that simulated data drawn from a fitted spatial model represent realizations of random spatial processes, the spatial pattern of residuals can provide

details or confidence about model fit. Cliff and Ord (1981) employ global join-counts and Moran's statistics to measure spatial autocorrelation of observed and expected maps derived from simulations of Hagerstrand's spatial diffusion model (Hägerstrand 1953). More recently, Wulder et al. (2007) use a local measure of spatial autocorrelation to assess the spatial variation in error/variance in several scenarios of a forest productivity model; this practice is becoming increasingly common (e.g., Räty and Kangas 2010). Here we are concerned specifically with this second problem, and the continued development of spatially sensitive metrics for assessing model goodness-of-fit.

A number of problems exist with measures of model fit commonly applied to traditional spatial models; for example, the true values for spatial parameters are generally unknown. Typically, assumptions and estimates of true values are based on results of field experiments taken over limited spatial scales (e.g., Turchin and Thoeny 1993). Scaling up field experiments to large-area spatial models is extremely challenging because ground truth data for models using coarsely grained units, such as pixels of 1 km by 1 km or larger, are very difficult to obtain.

A second issue specifically stems from comparing two maps, either two model outputs or a model output with validation data, for each individual spatial unit. Typically a measure of discrepancy is computed at each spatial unit and summarized in one global metric. If the spatial locations for one map are offset by even a single spatial unit, due to geo-referencing error, for example, quantitative comparison of spatial units may indicate a high level of mismatch. However, the maps may be very similar (see Hagen-Zanker 2006a). As such, pixel-by-pixel metrics may produce overly critical comparisons due to errors in spatial co-registration (Pontius 2000). A third related issue concerns spatial structure in the way that parameters themselves describe data, which could be easily missed by comparison of global summary measures.

Spatially local assessment methods, or those that characterize the spatial pattern of parameter estimates, are beneficial because they reveal systematic errors and spatial variability in errors that can be used to further refine models.

Furthermore, because some variability in predictions is expected from all models (i.e., all models are wrong), determining when the difference between two maps is more than expected given a data generating spatial process can be problematic. For instance, often two maps generated as realizations of a spatial model are expected to have variability in output values at the same spatial location (Csillag and Boots 2005). This issue has been discussed mostly in the remote sensing literature, where traditionally researchers have set subjective thresholds above which change is considered substantive (Gong and Xu 2003).

In the last decade, map comparison research has begun to overcome several of the preceding issues. Map comparison techniques are useful for evaluating similarities and differences between two map patterns. Map comparison approaches have been developed primarily for measuring agreement between categorical maps (Boots and Csillag 2006; Hagen-Zanker 2006b; Visser and DeNijis 2006); for example, land-use and land-cover change simulations (Hagen-Zanker and Martens 2008). A benefit of map comparison techniques is the use of spatial neighborhoods to assess spatial pattern similarity, rather than individual spatial units. Csillag and Boots (2005) present a framework to test the hypothesis that the spatial patterns expressed in two categorical maps are generated by the same spatial processes. Such an approach enables statistical assessment of whether generating spatial processes are similar, and overcomes the need to set arbitrary thresholds. A key insight in Csillag and Boots (2005) is the need to consider users' perceptions of when maps are different. Much less work considers the

assessment of change in continuous-value maps, although Hagen-Zanker (2006a) reviews a suite of potential methods.

A final issue with assessing spatially explicit models is that the true values for the spatial process of interest often are unknown, and so must be estimated in some fashion. Frequently, hierarchical spatial models are fit within the Bayesian paradigm with implementation based on Markov chain Monte Carlo (MCMC) techniques. Within the Bayesian context, posterior-predictive checking is a general approach to goodness-of-fit, whereby a fitted model, through its posterior distribution, is used to generate replicate data with the assumed model, where values are compared with observed data through a measure of discrepancy (Gelman, Meng, and Stern 1996). The combined methods described here represent a framework for checking spatial models, and provide a diagnostic tool to improve model-based spatial analysis.

In this paper, we employ a spatially explicit approach to evaluation of spatial models in a Bayesian model-checking framework. The primary contribution of this paper is a new model assessment methodology, based on two approaches: posterior predictive checks, and map comparison based on the structural similarity (SSIM) index, which are combined to quantify the spatial nature of a model fit. In short, fitted models are used to simulate new realizations of a process using its posterior predictive distribution, and these are compared to the observed data using spatially sensitive metrics. This approach is in line with the notion of a process-based approach to map comparison presented in Csillag and Boots (2005). The posterior predictive realizations from a fitted model can be interpreted as replicate data that could have been observed in a study area at another time, given that the data generating model is true. Therefore, comparing these realizations to the observed data provides a mechanism for assessing the fit of statistical models. A novel contribution of our work is the fusion of a Bayesian model-checking

approach with the SSIM index for checking the fit of spatial models. Because geographers typically work with data that describe one instance of spatially stochastic processes, simulation-based inference is a powerful method for understanding geographic patterns.

Methods

Two component methods for the proposed model assessment approach are posterior predictive checks and map comparison. We briefly review each of these approaches before demonstrating their role in spatial model assessment.

Posterior Predictive Model Checking

Posterior predictive model checking differs from the classical hypothesis testing framework in that emphasis is placed on measuring the discrepancy between observed data and replicate data simulated with a fitted model, rather than testing whether the model is true or false. In Bayesian statistics, all parameters are treated as stochastic and assigned a prior distribution $p(\theta)$, and inference proceeds through the posterior distribution $p(\theta|Y)$, which conditions on observed data Y. For a given model, specified by a likelihood $p(Y|\theta)$ and a prior $p(\theta)$, the posterior predictive distribution

$$p(Y_{rep} \mid Y) = \int p(Y_{rep} \mid \theta)\, p(\theta \mid Y)\, d\theta \qquad (1)$$

allows us to draw new replicate data from the fitted model. The replicate data Y_{rep} are assumed to have the same distribution as the observed data Y (which is specified as part of a model), and also assumed conditionally independent of *Y*, given θ. As described in Gelman, Meng, and Stern (1996), simulated realizations of Y_{rep} can be compared to observed data through a general discrepancy measure $T(Y_{rep},Y)$. This discrepancy measure is a specifically chosen statistic that compares observed data with replicate data generated with a model, and we focus here on a

measure for exploratory map comparison for continuous-valued spatial model assessment. In the analysis reported here, we simulate 100 Y_{rep} datasets from fitted models, and then compare the observed Y and Y_{rep} datasets (i.e., Y_{rep} vs, Y). Alternatively, in some cases the appropriate comparison is between parameter estimates obtained from posterior distributions and their true values (i.e., $\hat{\theta}$ vs. θ), to evaluate the fit of a specific model component. In all analyses, replicates were generated by randomly drawing parameter values θ_i from the posterior distribution θ, and using these to simulate new values (i.e., Y_{rep}).

Map Comparison

We selected the SSIM index (Wang et al. 2004), developed for evaluating image degradation, as an exploratory map comparison statistic for use in continuous-valued spatial model assessment. The SSIM index was originally proposed for evaluating the quality of image compression algorithms, but was later identified as a potential method for comparing continuous valued maps by Hagen-Zanker (2006a). The SSIM index is constructed to objectively make comparisons between images similar to the human visual system. A comparison of the mean square error (MSE) and SSIM index in Wang et al. (2004) shows that for a fixed MSE, vastly different image degradations are possible from a human perception standpoint. We employ the SSIM index to extend this notion to map comparison for model checking. A local region approach is ideal for simultaneously assessing similarity in spatial structure along with pixel-by-pixel correspondence in maps (Hagen-Zanker and Martens 2008).

The SSIM index considers three components for map comparison: luminance, contrast, and structure, relating to local differences in mean, variance, and correlation (Wang et al. 2004).

Three summary statistics used in calculating the SSIM index are computed for each cell on the basis of a defined local region (e.g., a 5x5 moving window).

$$\mu_a = \sum_{i=1}^{n} w_i a_i \tag{2}$$

$$\sigma_a = \left(\sum_{i=1}^{n} w_i (a_i - \mu_a)^2 \right)^{\frac{1}{2}} \tag{3}$$

$$\sigma_{ab} = \sum_{i=1}^{n} w_i (a_i - \mu_a)(b_i - \mu_b) \tag{4}$$

In the preceding three equations, *a* and *b* are two regular lattice maps, the index *i* iterates through *n* cells in a defined local region, and w_i are spatial weights that can be used to adjust the smoothness/ abruptness of the local region effect. For example, in Wang et al. (2004), an 11x11 circular local region is used with Gaussian weights. For equal weights, all w_i can be set to 1/n. The local measures are combined into the three SSIM index components–luminance (*L*), contrast (*C*), and structure (*S*)–as follows:

$$L(a,b) = \frac{2\mu_a\mu_b + c_1}{\mu_a^2 + \mu_b^2 + c_1} \tag{5}$$

$$C(a,b) = \frac{2\sigma_a\sigma_b + c_2}{\sigma_a^2 + \sigma_b^2 + c_2} \tag{6}$$

$$S(a,b) = \frac{\sigma_{ab} + c_3}{\sigma_a\sigma_b + c_3} \tag{7}$$

where the constants c_1, c_2, and c_3 are included for stability in situations where either the mean or variability is close to zero (e.g., large homogeneous patches). These constants can be related to the range of pixel values (*R*) via two additional constants, $k_1 = 0.01$ and $k_2 = 0.03$, established heuristically by Wang et al. (2004): $c_1 = (k_1R)^2$, $c_2 = (k_2R)^2$, and $c_3 = ½\ c_2$. Note that these three components (*L*, *C*, *S*) are relatively independent, and changes in one component do not

necessarily affect the others. The components *L* and *C* fall in the interval [0,1] with 1 indicating perfect agreement, and *S* falls in the interval [-1,1] (the correlation coefficient between cells in each window). A value for the SSIM index of -1 indicates perfect negative correlation among values in the locally compared regions. The three SSIM index components are multiplicatively combined to give a measure of similarity for each local region that is equal to 1 when the two maps are identical:

$$SSIM\ (a,b) = [L(a,b)]^{\alpha}\,[C(a,b)]^{\beta}\,[S(a,b)]^{\gamma} \qquad (8)$$

The exponents α, β, and γ can be used to weigh individual components, with default values taken as $\alpha = \beta = \gamma = 1$. Interpreting maps of *L*, *C*, *S* and the SSIM index values allows spatially local analysis of patterns of model fit. Alternatively, a global score of similarity (termed mSSIM) can be computed by taking the mean of all local SSIM index values. Global summaries can be calculated for each individual component. Given the structure of these formulas, the mSSIM value obtained by taking the mean of the local SSIM index values does not necessarily equal the product of the component global means.

Expected and observed maps with low similarity in L indicate a poorly estimated local mean, whereas disagreement in C may indicate that local transitions are different across estimates. For example, smoother transitions might result from a spatial smoothing parameter in a model that is not evident in the data under study (i.e., oversmoothing). Differences in local spatial structure indicate local disagreement in pattern, and are more robust to errors induced by slight spatial misalignment. Where S is high, spatial patterns tend to be similar; even if pixel magnitudes are quite different (i.e., low L). Spatial patterns in model fit, evaluated by the SSIM statistic, may highlight model mis-specification due to missing variables, terrain-induced errors, or models that are over-parameterized. When combined with simulations from a posterior

predictive distribution, the posterior-predictive-SSIM index provides a novel approach for spatial model assessment.

Visual versus Quantitative Model Assessment

Classical model assessment focuses on minimizing the error of predictions. The typical metrics for this are the root mean square error (RMSE) and the mean squared error (MSE). While error minimization is important to know how well a model describes the data, recent research in image processing reveals that in a spatial context, other aspects of image structure are very important for perception of spatial similarity in images (Wang et al. 2004; Wang and Li 2011; Brunet, Vrscay, and Wang 2012). Here we extend this notion to maps, suggesting that additional information about spatial structure may be useful in assessing the fit of spatially explicit models. Similar to categorical map comparison, where maps with the same level of composition for a given map class, but different spatial configurations of those classes, are perceived differently (Remmel and Csillag 2003), here we hypothesize that for a fixed level of error, but different spatial patterns in error, continuous-valued maps appear to be different.

To demonstrate, a Gaussian spatial process was simulated for a 100x100 lattice on the unit square. The value for each cell was distributed as Normal ($\mu = 127$, $\sigma = 50$), with spatial correlation defined by a partial sill and range of $\sigma^2 = 0.05$ and $\gamma = 2\mu$, respectively, and a Gaussian covariance function. A second Gaussian noise process was simulated with parameters $\mu=5$, $\sigma = 2$, and spatial parameters drawn from $\gamma = [100, 200, 300]$, $\sigma^2 = [0.05, 0.15, 0.30]$. Realizations of the noise process were added to the first process to create spatially variable distortion of the original map. Fig. 1 shows different realizations of the altered process where the MSE ($\Sigma[a-a_x]^2$, where a is the original process and a_x is the noise-distorted process) has been held constant. The SSIM index scores in Fig. 1 highlight different levels of apparent distortion

undetectable by the MSE alone. This result has been demonstrated many times, and is why 'structure-based' measures such as the SSIM index are used instead of the MSE in image quality assessment. The SSIM index rests on the assumption that the human visual system is adapted to extract structural information from a given view, and as such, image structure should be the basis for measures of image degradation. Similarly, we hypothesize that perceived similarity between maps also depends on spatial structure similarities. Based on this initial exploration, the SSIM index appears to be effective at discriminating between intuitive perceptions of map comparison for continuous valued maps. We contend that such additional information is helpful in spatial model assessment.

Figure 1 about here

A Simulation Study

Synthetic data were used in order to demonstrate the model checking framework under known conditions. In one experiment, a spatial model was specified to replicate a study investigating a spatial process at a snapshot in time (e.g., an economic indicator across counties). In a second experiment, a discrete space-time model was used to create synthetic data to replicate a study evaluating spatially distributed model parameters over time (e.g., the diffusion rate for the spread of a disease across a landscape). We compare both replicate data (Y and Y_{rep}) and spatial parameter estimates (i.e., $\hat{\theta}$ and θ) using the SSIM index, as well as the MSE in a posterior predictive model checking framework. The method demonstrated here is appropriate for a study area subdivided into *n* spatial units forming a regular square lattice. Fig. 2 outlines the analysis methodology.

Figure 2 about here

A Spatial Data Generating Process: The Conditional Autoregressive Model

Models that explicitly incorporate the spatial structure of a process and underlying covariates are becoming commonplace due to recent statistical developments and the relative ease of fitting these models in software such as WinBUGS (Lunn et al. 2000) or the MCMCglmm package in R (Hadfield and Kruuk 2010). When spatial effects (i.e., autocorrelations) arise from missing variables that are spatially structured and represent a theoretical population of random effects, typically a spatial random effects model is employed. Versions of such models are frequently used in disease mapping (Lawson 2009), spatial econometrics (Anselin 1988), and spatial statistics (Chun and Griffith 2013). Modelling proceeds via a Bayesian hierarchal approach, where data are specified conditional on unknown parameters, which are linked hierarchically to other parameters that are given spatial prior distributions. The spatial model we employed was an intrinsic Gaussian conditional autoregressive model with spatial random effects defined by

$$E[b_i \mid b_{-i}] = \sum_{j \neq i}^{N} b_j / m_i$$
$$Var\ [b_i \mid b_{-i}] = (m_i \tau)^{-1} \qquad (9)$$

where m_i is the number of neighbours defined in a list N of indices that define each adjacency-based neighbourhood, and τ is a parameter characterizing the conditional variance. Thus the conditional expected value in any cell is a weighted average of its neighbours, with a conditional variance inversely proportional to the number of neighbours. Neighbour relations were defined on a 40x40 grid using the queen's case contiguity; m_i is 8 for all non-edge cells. The conditional expectation of an observation in any given cell is a linear combination of observations of its neighbours.

The model fit to the simulated data (described subsequently) was a null spatial model where

$$\begin{aligned} y_i &\sim Normal \quad (\mu_i, \tau) \\ \mu_i &= b_0 + \beta X + b_i \end{aligned} \tag{10}$$

where X is a vector of covariates, and β is a vector of regression coefficients. The spatially correlated error b_i is a conditional autoregressive result, as specified previously for each grid cell, and b_0 is an intercept. Regression coefficients (β, b_0) were given Normal prior distributions (μ = 0, σ = 1000). Data were simulated to create a spatially structured dataset as follows: a 40x40 grid made up a study area where each grid cell had one covariate and coefficient values drawn from Normal distributions with spatial correlation defined by an exponential spatial covariance function [γ=10, σ^2=0.05]. This specification induced spatial structure in the covariates, and ultimately in the simulated dependent variable, which was calculated as Y=127+β*X*, yielding a dependent variable (Y), independent variable (*X*), and coefficient (*β*) at each spatial location (*n* = 1,600). For the purposes of illustrating the model assessment procedure, measurement error was excluded from the model.

The model was fit to the simulated data in WinBUGS using 100,000 iterations after a burn-in of 10,000, where convergence was examined via the Gelman-Rubin statistic (Gelman and Rubin 1992) and autocorrelation plots were inspected visually for serial autocorrelation. Parameter posterior distributions were used to generate 100 replicate datasets (Y_{rep}) from the fitted model in order to make inferences at a pseudo-significance level of 0.01. The map comparison statistic SSIM index and the MSE were used to compare each Y_{rep} to the true *Y* values yielding 100 maps of the SSIM index and each of its components (*L*, *C*, and *S*).

A Space-time Data Generating Process: The Space-time Logistic Mixed Model

To further illustrate the importance of spatial structure in model assessment, we used the SSIM index for comparison of data simulated with a space-time model having spatially local parameters. In this experiment, we specified a binary space-time process that was discrete in both space and time, analogous to the spread of an invasive species or emerging disease. Given n regions comprising a study area, we let $Z_i(t) \in \{0, 1\}$ indicate presence (e.g., of disease) in region *i, i=1,..., n,* at time *t, t=1,...,T*. Therefore $\mathbf{Z}$ represented a sequence of binary maps describing the progression of spread across the landscape. We specified a logistic model for the transition probabilities (p_{it}) of the spatial spread process conditional on the previous time period, $Pr\{\mathbf{Z}(t)|\mathbf{Z}(t–1)\}$. Following Smith et al. (2002), we assumed that occupied regions remain occupied, so that $p_{it} = 1$ if $Z_i(t–1) = 1$; if region *i* is unoccupied at time *t*–1, so that $Z_i(t–1) = 0$, the probability it becomes occupied at time *t* is defined as:

$$log\left(\frac{p_{it}}{1-p_{it}}\right) = \mu_t + \lambda_i NN_{i,t-1} \qquad (11)$$

where μ_t is a time varying parameter representing a baseline probability of occupation, $NN_{i,t-1}$ is the number of occupied neighbors of region *i* at time *t*–1, and λ_i is a spatially varying parameter quantifying the impact of occupied regions on their unoccupied neighbors. Here, a CAR prior was used for the spatially varying term λ_i, identical to that in the previous example, while a Normal prior was used for each μ. Full details of the model are given in Long et al. (2012).

This experiment focused on investigating the spatial structure of differences between the true values for λ and those estimated by the model, $\hat{\lambda}$ (as opposed to *Y* and Y_{rep}). True values for λ were simulated to represent various patterns of spatial structure. The spatial processes reported include a linear spatial trend and constant temporal trend (process 2A; Table 1), a Gaussian

Markov random field (GMRF) exhibiting spatial nonstationarity and constant temporal trend (process 2B; Table 1), and a GMRF and sinusoidal temporal trend (process 2C; Table 1). These three scenarios were used to simulate binary data (presence/absence) describing a spreading process on a 40x40 grid over 100 time periods. As such, the λ values from each of the three scenarios represent the true values with which fitted estimates ($\hat{\lambda}$) are compared.

Table 1 about here

The space-time model was fitted to the simulated datasets in WinBUGS using 100,000 iterations after a burn-in of 10,000, where convergence was confirmed via the Gelman-Rubin statistic (Gelman and Rubin 1992) and autocorrelation plots were inspected visually for serial autocorrelation. Again, 100 replicate datasets (termed Y_{rep}) were simulated from the posterior distributions of parameters to facilitate posterior predictive checking of model fit. Unlike the preceding experiment with which we compared the map of Y with the Y_{rep} data (producing 100 SSIM index comparisons), in the space-time case, we assessed the spatial structure of comparisons of the fitted estimates of the spatial diffusion parameter $\hat{\lambda}$ with the true values λ. We used the posterior means as point estimates of $\hat{\lambda}$ to compare via map comparison with the map of true λ values.

Simulation Study Results and Discussion

SSIM index and posterior predictive checking model assessment analysis was conducted in both spatial and space-time modelling contexts. Results indicate complimentary findings that spatial structure is an important part of model error and the proposed methodology reveals hidden structure in model spatial errors. .

Spatial Model Assessment

The Bayesian posterior predictive p-value revealed the posited model did fit the generated data, as the p-value based on the sample mean was 0.47 (where p-values in the tails indicate poor fit). When the average values for Y_{rep} were compared with the observed values for each cell (i.e., a pixel-by-pixel comparison), the mean square error was 0.72, for a dependent variable with mean 3.19 and standard deviation 2.31.

Figure 3 about here

Fig. 3 presents a map comparison analysis comparing the Y_{rep} to the observed data. The mean SSIM index score was 0.88, varying between 0.78 and 0.89 when comparisons were made between each of the 100 Y_{rep} and Y. The SSIM index components mapped in Fig. 3 reveal that the main component contributing to lack of fit was structure–identifying a lack of fit in pattern in certain parts of the map. This pattern is not apparent in the map of squared error (Fig. 3), where high errors appear randomly distributed. Because L and C components were uniformly very high (L = 0.99, C = 0.98, Table 2), areas of high structural dissimilarity indicate differences in pattern, where the values were still similar in terms of locally averaged error magnitude. Highlighted areas (dashed lines) in Fig. 3a-d demonstrate that on the right side, whereas L averages over local error variation, S highlights a pattern of dissimilarity in in spatial structure (i.e., correlation). This is due to overlapping windows that use the data multiple times in computation of the SSIM

index. On the left side (Fig. 3a-d), a difference in pattern is highlighted by the S component as well, while the map of errors appear to have no spatial structure.

Space-time Model Assessment

The SSIM analysis results of the space-time models are outlined in Table 2. For the simplest case, with linear spatial trend and no temporal trend in translocation, the structural similarity was poorest, with a mean SSIM index of 0.12. The components responsible for this low score were L (0.55) and C (0.56), whereas S was higher (0.70). Conversely, for patterns generated from a GMRF exhibiting spatial non-stationarity (processes 2B and 2C), L (0.95, 0.98) and C (0.76, 0.92) scored higher, whereas S scored lower (0.32, 0.74). In all cases, the global SSIM index indicates differences in spatial structure between true model parameters and those estimated by the model. Of all models examined, process 2A had the lowest scores in the SSIM analysis. Threshold-like behaviour is evident, whereby values for λ below 1 could not be estimated accurately. Dependence between L and C is evident, as those areas that scored low in L had higher C scores. The S component was high due to the simple nature of the pattern (i.e., a linear vertical trend). The best fitting model spatially was for process 2C, with spatial nonstationarity for local spread and sinusoidal temporally random spread (Fig. 4). S and C were components dominating spatial estimation error in this more complex process. The small luminance errors were smoothed over by the moving window, yielding high L across the study area. Dashed areas in Fig. 4 highlight both S and C as contributing components. Interestingly, the high error hotspot evident in the map of squared errors is completely smoothed over in the SSIM maps. Because the smoothing is a function of the window size (5x5) and the smoothing function (none used here), SSIM results must be interpreted relative to these two parameters.

Table 2 about here

Figure 4 about here

A Case Study: Bayesian Kriging

The simulation study demonstrates the use of the SSIM index in a context in which we have a known value for a spatial variable, and we estimate a model, and we make comparisons. However, practical spatial modelling encounters few situations where the true values of a spatial variable are known. Typically when building spatial models, the only data available to assess model fit are used to build a model. As discussed previously, posterior predictive checking provides a framework for using a model to draw new simulations from a posited process, and these can be compared with the original data. In this case, model assessment takes one of two forms: 1) check different realizations of spatial models against one thought to be the best, or 2) compare all iterations of models against each other. We employ the former strategy in an assessment of precipitation modelling in Sri Lanka.

The Study Area and Data

Sri Lanka is situated in the Indian Ocean, off the southeastern tip of the Indian subcontinent. The climate is tropical, and weather is characterized by two seasonal monsoons. The northeast (maha) monsoon typically lasts from October until March, whereas the southwest (yala) monsoon lasts from April until September. The southwest area of Sri Lanka generally receives significant rainfall during all seasons, while the northern and eastern regions of the country become arid and dry during the southwest monsoon season. Our long-term research interests focus on identifying associations between rainfall and incidence of waterborne infectious diseases.

Precipitation data were obtained from the Department of Meteorology of Sri Lanka. They include daily rainfall measurements (millimeters) from a network of 361 small-scale agro-ecological weather monitoring stations (Fig. 5). The spatial distribution of the station network varies considerably with population, climate, and landuse. For this research, daily rainfall measurements were aggregated into total monthly rainfall. A subset of these data was extracted for the month of December 2008.

Figure 5 about here

Daily rainfall data were obtained from 20 official meteorological stations operated by the Department of Meteorology of Sri Lanka. This data set was aggregated into monthly rainfall for December 2008. The official meteorological station data (n = 20) was used to validate the interpolations generated using the larger (n = 361) agro-ecological monitoring data.

Methods

Bayesian kriging was used to interpolate rainfall values across Sri Lanka onto a regular spatial lattice (1 km grid cells). Bayesian kriging differs from ordinary kriging in that priors are put on parameters of the semivariogram, and estimation yields a posterior distribution for each of the parameters (range, sill, and nugget). Random draws from the posteriors were used to generate predictive simulations (i.e., Y_{rep}) and these were compared to the simulation based on posterior means using the SSIM index. Similar to the previous analysis, the variability expressed in the SSIM analysis reveals uncertainty in the fit of the predicted interpolations. A Gaussian spatial linear mixed model was used to perform Bayesian kriging. The R package geoR (Ribiero and Diggle 2001) was used to generate posterior predictive distributions. Because the initial rainfall values used to perform Bayesian kriging were not normally distributed, a Box-Cox transformation was performed on the rainfall variable to satisfy model assumptions. The Box-

Cox transformation searches for an exponent value which is applied to each observation in order to make the shape of the distribution more Normal. The interpolated data then were back-transformed for analysis and interpretation. Posterior predictive simulations (n = 99) were obtained with kriging by using random draws from posterior distributions of the semivariogram parameters. Each interpolation was compared to the posterior mean interpolation using the SSIM index.

Results and Discussion

Fig. 6 presents a frequency distribution of mSSIM values from 99 Y_{rep} compared to the posterior mean. The majority of the mSSIM values were between 0.26 and 0.28. These values suggest that the interpolation generated from posterior predictive simulations were not very similar spatially to the posterior mean. The highest mSSIM value attained was 0.31, while the lowest was 0.22 (Table 2). Fig. 7 displays the SSIM map outputs attributed to both of these mSSIM values together with the posterior mean raster. While L was very high for both simulations (0.99), both C (high mSSIM: 0.68, low mSSIM: 0.61) and S (high mSSIM: 0.49, low mSSIM: 0.38) were notably lower, and can be thought to be responsible for the low mSSIM values. The spatial pattern of the distribution of rainfall in Sri Lanka is quite complex, which is reflected in the low S scores for both simulations.

Figure 6 about here

Southern Sri Lanka attained much higher SSIM index scores than did northern Sri Lanka for both simulations, which most likely results from the much sparser sampling of rainfall values in the north vis-a-vis the south. Because far fewer observed rainfall values support interpolation in northern Sri Lanka, the posterior variances in these areas are likely to be much higher. Therefore, the simulation that attained the highest mSSIM value did so largely as a result of its

structural similarity in the southwest region of Sri Lanka (the most densely sampled area). This outcome suggests that the map of local SSIM index values is a much more important diagnostic tool in interpolation assessment than the global score.

Figure 7 about here

Appendix A summarizes a comparison of the observed rainfall measurements at the same 20 meteorological station locations to the predictions attained from the posterior mean from Bayesian kriging, as well as a variety of other different spatial interpolation methods (ordinary kriging, inverse distance weighting, and spline interpolation). While the mSSIM scores of comparisons between posterior predictive distribution simulations and the posterior mean interpolation were quite low, suggesting that visually, substantial change exists in model output in simulations from the posterior predictive distribution, Bayesian kriging attained the lowest mean absolute error (39.91 mm) over all 20 locations of any of the interpolation techniques.

Discussion

Map comparisons revealed that with both models, even when a model appeared to describe data well, spatial irregularities existed in how a model fits to data. The SSIM index identifies local differences in mean, variance, and correlation, providing information about spatial context and differences in each that can be further explored to reveal systematic deficiencies in a model specification. As Cliff and Ord (1981) show with global spatial measures, our analysis here demonstrates the importance of spatial model assessment when working with spatially detailed models. However, our work differs from previous attempts at spatially-oriented model validation in a number of important ways. First, the assessment of models using the SSIM index provides spatially local information not available in global autocorrelation measures. From a diagnostic perspective, this approach attempts to replicate the human visual system and determine models

that look right. But we are not advocating the use of the SSIM index as a replacement for error based approaches; rather, we suggest that map comparison metrics should play a complimentary role in model assessment for spatial models. Fig. 1 highlights the need for this type of metric.

Two simulation experiments indicate that the SSIM index analysis provides additional information for model assessment purposes. However we have not defined any explicit criteria for distinguishing between good and bad model fit based on the SSIM index alone (e.g., SSIM index > 0.5 = good fit). The SSIM index, like many model assessment tools, is best used as a comparative measure of model agreement. The overall SSIM index is sensitive to three components: L, C, and S, which are largely independent of each other. The SSIM index should be considered together with these three components, because each component provides unique information valuable for model assessment. Given that local SSIM index values are the product of local L, C, and S, a small SSIM index value may be the result of lower scores in all three components, or a mixture of high and low scores in various components. This relationship becomes further complicated as local values are averaged across a map to provide a global score. For example, comparing the individual component results from the space-time simulation example (e.g., process 2A, SSIM = 0.12, L = 0.55, C = 0.56, S = 0.70; Table 2) with those from the Bayesian kriging case-study (SSIM = 0.27, L = 0.99, C = 0.65, S = 0.44; Table 2) illustrates the different ways component scores can combine to form global SSIM index values. From this result we can identify which components show better agreement, and use this extra information in assessment and model improvement. Finally, the value of any spatially local analyses is the resulting spatial information that is most effectively portrayed with a map. Thus, the local SSIM index maps (e.g., Fig. 7d, e) provide invaluable information about the spatial structure of model

agreement. Such maps can be used to examine where and how well the spatial structure of model output matches some true map or expectation, and where that model fails.

Although the analysis here employs the SSIM index as the local spatial measure, this is by no means the only available option, and using this approach has its disadvantages. The SSIM index method is wholly dependent on the window size parameters (set here to a 5x5 local window on a 40x40 grid). Further, recently the whole notion of measuring perceived error using the SSIM has been questioned. Dosselmann and Yang (2011) suggest that the SSIM index is directly related to the MSE, and its formulation (a product of means, variances, and correlations) is too simple to actually model the human visual system. Regardless, while the mechanism accounting for its performance requires further investigation (e.g., luminance adaptation, textural masking), studies demonstrate its ability to accurately reflect subjective mean opinion scores (Wang et al. 2004). A major limitation in our application to maps is that the SSIM index method is implemented and tested only on maps defined using a regular spatial lattice, which currently limits its application in model checking to spatial models based on raster data or a regular square grid. Future work could explore the properties of this statistic for more complex spatial lattice structures; however in such cases, care must be taken in the definition of the spatial neighbourhood matrix. The framework presented here is easily extendible to other spatial measures (e.g., Getis and Ord 1992; Anselin 1995), and we suggest that further research in this area is warranted. What is lacking in this analysis is a formal hypothesis test to determine significantly similar or different spatial models (Csillag and Boots 2005). However, before such hypotheses can be realized, a full analysis is required of the statistical power of the SSIM index for different spatial model specifications. How the SSIM index values relate to their data scale units remains unclear, and more work needs to be done to understand this relationship before the

SSIM index can be used in geographical analysis beyond relative comparisons. A final limitation is that the SSIM approach is scale dependent, and as in any pattern analysis, multiscale approaches are critical to identify the scale sensitivity to observed patterns. This limitation is also true in the context of model assessment. Wavelet-based methods in particular may have potential for comparison of continuous valued maps, and thus spatial model assessment.

Conclusion

Typically, with spatial and space-time models, evaluation of model fit is based solely on traditional aspatial model diagnostics. Here we advocate for a spatially explicit approach to testing how some mapped spatial output from a model differs from an expectation as a complimentary model diagnostic tool. Specifically, the SSIM index is a useful tool for spatial model assessment because it calculates local differences in mean, variance, and correlation (spatial structure) between two maps using a spatial neighbourhood-based approach. Given the rate at which spatial models are now implemented in a range of applications, the inclusion of spatial measures for model assessment provides invaluable spatial information that can be used to examine the qoodness-of-fit of a chosen model.

Appendix A

Table A1 Bayesian kriging predictions compared with multiple interpolation methods' predictions and observed rainfall values at 20 official meteorological station locations

Observed Rainfall (mm)	Bayesian Kriging (mm)	Ordinary Kriging (mm)	Inverse Distance Weighting (mm)	Splines (mm)
259.57	265.65	242.57	305.78	355.68
96.52	110.76	117.77	102.21	104.02
146.71	130.85	127.61	137.72	27.24
160.92	195.28	143.87	129.11	109.91
91.31	115.82	82.94	81.25	87.58
251.81	182.12	150.87	143.18	-269.53
208.71	87.68	76.91	78.56	64.97
131.73	79.33	120.6	103.46	65.65
181.42	187.97	255.79	165.37	184.61
147.34	69.27	111.13	77.89	93.42
94.52	74.69	100.65	78.48	83.74
47.55	27.81	85.09	77.77	-56.74
304.12	254.76	206.55	280.13	302.9
57.34	35.59	58.62	39.55	42.44
276.23	312.12	206.95	277.32	281.86
321.94	256.53	251.12	216.97	273.74
313.3	283.09	198.65	282.93	352.18
244.92	294.18	235.45	206.31	215.22
166.84	171.85	123.34	46.63	-33.05
184.54	105.66	132.56	81.98	83.62
Average Absolute Error:	39.91	46.97	47.06	81.03

Acknowledgements

The authors thank the Department of Meteorology, Sri Lanka for facilitating access to rainfall data. The analysis presented here does not necessarily reflect the views of the Department. We also thank both GEOIDE and the Social Sciences and Humanities Research Council of Canada for funding.

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, 1:267–281. Springer Verlag.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. New York: Springer.

Anselin, L. (1995). "Local Indicators of Spatial association-LISA." Geographical Analysis 27(2), 93–115.

Boots, B., and F. Csillag. (2006). "Categorical Maps, Comparisons, and Confidence." *Journal of Geographical Systems* 8(2), 109–18.

Brunet, D., R. Vrscay, and Z. Wang. (2012). "On the Mathematical Properties of the Structural Similarity Index." *IEEE Transactions on Image Processing* (99): 1488–1499.

Brundson, C., A.S. Fotheringham, and M. Charlton. (1996). "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity." *Geographical Analysis* 28(4), 281–98.

Burnham, K. P., and D. R. Anderson. (2002). *Model Selection and Multi-model Inference: a Practical Information-theoretic Approach*. New York: Springer.

Carletta, J. (1996). "Squibs and Discussions Assessing Agreement on Classification Tasks: The Kappa Statistic." *Computational Linguistics* 22(2), 249–54.

Chun, Y., and Griffith, D.A. (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks: SAGE.

Cliff, A., and J. K. Ord. (1981). *Spatial Processes Models and Applications*. London: Pion Limited.

Csillag, F., and B. Boots. (2005). "A Framework for Statistical Inferential Decisions in Spatial Pattern Analysis." *The Canadian Geographer* 49(2), 172–79.

Dice L.R., (1945). "Measures of the Amount of Ecologic Association between Species." *Ecology* 26(3), 297–302.

Ribeiro Jr., P.J. and P.J. Diggle. (2001). "geoR: A package for geostatistical analysis." *R-NEWS* 1(2),15-18.

Dosselmann, R., and X. D. Yang. (2011). "A comprehensive assessment of the structural similarity index." *Signal, Image and Video Processing* 5(1), 81–91.

Finley, A. O. (2011). "Comparing Spatially Varying Coefficients Models for Analysis of Ecological Data with Non Stationary and Anisotropic Residual Dependence." *Methods in Ecology and Evolution* 2(2), 143–54.

Finley, A.O., S. Banerjee, and A.E. Gelfand. (2012). "Bayesian Dynamic Modeling for Large Space-time Datasets Using Gaussian Predictive Processes." *Journal of Geographical Systems* 14 (1): 29–47.

Gelfand, A. E., H.J. Kim, C. F. Sirmans, and S. Banerjee. (2003). "Spatial Modeling with Spatially Varying Coefficient Processes." *Journal of the American Statistical Association* 98(462), 387–96.

Gelman, A., and D. B. Rubin. (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4), 457–72.

Gelman A, X.L. Meng XL, and H. Stern. (1996). "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica* 6,733–59.

Getis, A., and C. Ord. (1992). "The Analysis of Spatial Association by use of Distance Statistics." *Geographical Analysis* 24(3), 189–206.

Gong, P., and B. Xu. (2003). "Remote Sensing of Forests over Time: Change Types, Methods, and Opportunities." In *Remote Sensing of Forest Environments: Concepts and Case Studies*, 301–34, edited by M.A. Wulder, and S.E. Franklin. Norwell: Kluwer Academic Publishers.

Griffith, D. A. (2008). "Spatial-filtering-based Contributions to a Critique of Geographically Weighted Regression (GWR)." *Environment and Planning A* 40(11), 2751–69.

Hadfield, J. D, and L. E. B. Kruuk. (2010). "MCMC Methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R Package." *Journal of Statistical Software* 33(2), 1–22.

Hagen-Zanker, A. (2006a). "Comparing Continuous Valued Raster Data: A Cross Disciplinary Literature Scan." Netherlands Environmental Assessment Agency. Maastricht: Research Institute for Knowledge Systems.

Hagen-Zanker, A. (2006b). "Map Comparison Methods that Simultaneously Address Overlap and Structure." *Journal of Geographical Systems* 8(2), 165–85.

Hagen-Zanker, A., and P. Martens. (2008). "Map Comparison Methods for Comprehensive Assessment of Geosimulation Models." *Lecture Notes in Computer Science* 5072, 194–209.

Hagerstrand, T. (1953). "On Monte Carlo simulation of diffusion" in *Economic and Cultural Topics (Quantitative Geography, Part I,* edited by W L Garrison, D F Marble, [2 vols.; Evanston, Illinois: Northwestern University Department of Geography, 1967]), 1-32.

Kass, R. E., and A. E. Raftery. (1995). "Bayes Factors." *Journal of the American Statistical Association* 90 (430), 773–795.

Lawson, A. B. (2009). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology,* Boca Raton: Chapman & Hall/CRC.

Lee, H., and S. K. Ghosh. 2009. “Performance of Information Criteria for Spatial Models.” *Journal of Statistical Computation and Simulation* 79 (1): 93–106.

Long, J. A., C. Robertson, F.S. Nathoo, and T.A. Nelson. (2012) “A Bayesian Space–time Model for Discrete Spread Processes on a Lattice. *Spatial and Spatio-temporal Epidemiology* 3, 151-162.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter. (2000). “WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility.” *Statistics and Computing* 10 (4), 325–37.

Pontius Jr, R.G., (2000). “Quantification Error versus Location Error in Comparison of Categorical maps.” *Photogrammetric Engineering and Remote Sensing* 66,1011–16.

Räty, M., and A. Kangas. (2010). “Segmentation of Model Localization Sub-areas by Getis Statistics.” *Silva Fenn* 44 (2), 303–17.

Remmel, T., and F. Csillag. (2003). “When Are Two Landscape Pattern Indexes Significantly Different.” *Journal of Geographical Systems* 5, 331–51.

Schwarz, G. 1978. “Estimating the Dimension of a Model.” *The Annals of Statistics* 6 (2): 461–464.

Smith D.L., B. Lucey, L.A.Waller, J.E. Childs, and L.A. Real. (2002). "Predicting the Spatial Dynamics of Rabies Epidemics on Heterogeneous Landscapes." *Proceedings of the National Academy of Sciences* 99, 3668–72.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (4): 583–639.

Turchin, P., and W. Thoeny. (1993). "Quantifying Dispersal of Southern Pine Beetles with Mark-recapture Experiments and a Diffusion Model." *Ecological Applications* 3(1), 187–98.

Thomas, E. N. (1968). *Maps of Residuals from Regression: Their Characteristics and Uses in Geographic Research*. Ann Arbor: University Microfilms, A Xerox Company.

Visser, H., and T. de Nijs. (2006). "The Map Comparison Kit." *Environmental Modelling & Software* 21(3), 346–58.

Wang Z., A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing* 13, 600–11.

Wang, Z., and Q. Li. (2011). "Information Content Weighting for Perceptual Image Quality Assessment." *IEEE Transactions on Image Processing* 20 (5), 1185–98.

Wheeler D.C., and L.A.Waller. (2008). "Mountains, Valleys, and Rivers: The Transmission of Raccoon Rabies over a Heterogeneous Landscape." *Journal of Agricultural, Biological, and Environmental Statistics* 13, 388–406.

Wulder, M. A., J.C.White, N.C. Coops, T. Nelson, and B. Boots. (2007). "Using Local Spatial Autocorrelation to Compare Outputs from a Forest Growth Model." *Ecological Modelling* 209(2–4), 264–76.

Table 1 Spatial processes used in the simulation experiments

Name	Model Parameters	Spatial Parameter
1	μ, σ	σ – Exponential distance decay
2A	Λ, M	Λ – Linear vertically increasing
2B	Λ, M	Λ – Gaussian Markov Random Field
2C	Λ, M	Λ - Gaussian Markov Random Field

Table 2 Average SSIM index values, component values and MSE for 100 Y replicate data obtained from a conditional autoregressive model (Process 1) and three space-time models (process 2A, 2B, 2C). The range of SSIM index and component values obtained from a Bayesian kriging case study of spatial interpolation of rainfall in Sri Lanka (99% credible interval) .

Measure	Process 1 [CAR model]	Process 2A[L1M2]	Process 2B[L3M2]	Process 2C[L3M3]	99% Credible Interval
SSIM	0.88	0.12	0.47	0.67	0.22-0.31
L	0.99	0.55	0.95	0.98	0.99-0.99
C	0.98	0.56	0.76	0.92	0.61-0.68
S	0.87	0.70	0.32	0.74	0.38-0.49
MSE	0.72	0.51	0.42	0.22	75.80-51.20
P-value	0.48	n/a	n/a	n/a	n/a

Figure 1. Examples of the structural similarity (SSIM) index measures of mapped patterns on a 100x100 grid. Maps are realizations of a) a Gaussian spatial process, with b) and c) Gaussian noise added. Distorted maps have similar mean squared errors (MSE) when compared with the reference map, but different SSIM index values.

Figure 2. An overview of the analysis methodology used in this paper, including simulation studies (a-b) and, c) a case study of rainfall interpolation in Sri Lanka.

Figure 3. Spatial model results comparing a reference map (observed data) with one replicate dataset simulated by a random draw from posterior distributions of model parameters (replicate data). The SSIM index, squared error, and SSIM components are presented. Dashed lines indicate areas of spatial discrepancy based on SSIM index analysis.

Figure 4. Space-time model results comparing a reference map (true λ) with posterior mean estimates for spatial diffusion parameters. The SSIM index, squared error, and SSIM components are presented. Dashed lines indicate areas of spatial discrepancy based on SSIM index analysis.

Figure 5. Locations of official meteorological stations and small scale weather stations in Sri Lanka. 2008.

Figure 6. Histogram of mSSIM values from 99 simulations drawn from the posterior distribution.

Figure 7. A case study–Bayesian kriging of rainfall in Sri Lanka. Results comparing posterior predictive distribution simulations with highest and lowest mean SSIM index values with the posterior mean. The SSIM indices for a sample area of each simulation also are presented.