

Quantifying uncertainty associated with a numerical model output

La quantificazione dell'incertezza associata all'output da modello numerico

Lucia Paci, Daniela Cocchi and Alan E. Gelfand

Abstract Environmental numerical models are deterministic tools widely used to simulate and predict complex systems. However, they are unsatisfying since they do not provide information about the uncertainty associated with their predictions. Conversely, uncertainty assessment of model outputs can be useful to guide environmental agencies in improving computer models. We propose a Bayesian hierarchical model to obtain spatially varying uncertainty associated with a numerical model output. We show how we can learn about such uncertainty through suitable stochastic data fusion modeling using some external validation data. The model is illustrated by providing the uncertainty map associated with a temperature output over the northeastern United States.

Abstract *I modelli numerici deterministici sono ampiamente usati per simulare e prevedere sistemi complessi. Tuttavia, risultano insoddisfacenti in quanto non considerano l'incertezza associata alle loro previsioni. Al contrario, la valutazione dell'incertezza del loro output potrebbe aiutare a migliorarli. Si propone un modello gerarchico Bayesiano per ottenere una misura di incertezza associata a tale output che vari nello spazio tramite un appropriato modello stocastico di data fusion che impiega dati di validazione. Si illustra il modello con la mappa di incertezza associata ad un output deterministico di temperatura per la regione nord-est degli USA.*

Key words: Hierarchical modeling, data fusion, measurement error model, Bayesian inference, logCAR prior

Lucia Paci

Department of Statistical Sciences 'P. Fortunati' - University of Bologna, Via delle Belle Arti, 41
Bologna, e-mail: lucia.paci2@unibo.it

Daniela Cocchi

Department of Statistical Sciences 'P. Fortunati' - University of Bologna, Via delle Belle Arti, 41
Bologna, e-mail: daniela.cocchi@unibo.it

Alan E. Gelfand

Department of Statistical Science - Duke University, Box 90251, Durham, NC, e-mail:
alan@stat.duke.edu

1 Introduction

Numerical models are playing an increasing role as tools to understand and predict complex phenomena such as air pollution or meteorological events. Since computer models are thought under a deterministic framework, they do not provide any measure of uncertainty associated with their output. However, large sources of uncertainty in constructing and employing numerical models do exist and it is worth accounting for such uncertainty.

The Bayesian approach represents a very natural way to account for different uncertainty sources and several methods have been developed to deal with the uncertainty analysis in computer models under this perspective [6, 4, 2]. However, not much has been said about statistical methods for attaching uncertainty to model outputs when we do not have information about how such deterministic predictions are created. Indeed, our proposal builds upon the notion of uncertainty introduced by [3] when numerical models are unavailable, rather only deterministic outputs at some spatial resolution are provided. Paper [3] introduces a general Bayesian approach to associate uncertainties with deterministic interpolated surfaces, using some external *validation* data collected independently over the same spatial domain. Although numerical models produce deterministic surfaces, the output will not ever be the “true” value of the process since any model is a simplification of reality. So, given the truth and the model output, the associated error is not stochastic. But, under suitable stochastic modeling, this error can be reinterpreted as a random unknown which we can infer about using a Bayesian specification within the data fusion setting.

The contribution of the present work is to develop a Bayesian hierarchical model to provide spatially smoothed uncertainty associated with numerical model outputs, without information about how they were created. We show how we can learn about such uncertainty through stochastic data fusion modeling using some external validating data. We also take into account the change of support problem (COSP) which arises from the spatial misalignment between the numerical model output (averages over grid cells) and the validation data (point-level measurements), even if our objective is not the calibration of the numerical model output.

We apply our Bayesian model to obtain the uncertainty map associated with the temperature output provided by the Rapid Update Cycle (RUC) weather model over the northeastern United States (U.S.). The validation data set consists of temperature measurements collected at monitoring stations operating in the same study region.

2 Data fusion modeling

Let $R(A_i)$ denote the numerical model output over grid cell A_i , ($i = 1, \dots, I$). First, we specify a measurement error model for $R(A_i)$ relative to the truth, that is:

$$R(A_i) = \tilde{R}(A_i) + \varepsilon_r(A_i) \quad (1)$$

where $\tilde{R}(A_i)$ is the underlying process representing the ‘true’ average value for A_i and we assume $\varepsilon_r(A_i) \sim N(0, \sigma_r^2(A_i))$ independently $\forall i = 1, \dots, I$. The true average value $\tilde{R}(A_i)$ arises from a Gaussian Markov Random Field (GMRF) equipped with a conditionally autoregressive structure (CAR), that is:

$$\tilde{R}(A_i) | \{\tilde{R}(A_{i'}) : i' \neq i\} \sim N\left(\sum_{i' \sim i} \frac{\tilde{R}(A_{i'})}{w_i}, \frac{\tau^2}{w_i}\right) \quad (2)$$

where $i' \sim i$ identifies the cell $A_{i'}$ adjacent to cell A_i and w_i is the number of neighbors of cell A_i .

Let $V(\mathbf{s}_j)$ be the validation data at location $\mathbf{s}_j$, ($j = 1, \dots, n$) gathered from independent station data over the same region as the output, and $\tilde{V}(\mathbf{s}_j)$ denotes the true value at $\mathbf{s}_j$. We specify a spatial measurement error model for $V(\mathbf{s}_j)$, that is:

$$V(\mathbf{s}_j) = \tilde{V}(\mathbf{s}_j) + \varepsilon_v(\mathbf{s}_j) \quad (3)$$

where $\varepsilon'_v = (\varepsilon_v(\mathbf{s}_1), \dots, \varepsilon_v(\mathbf{s}_n))$ is a zero-mean Gaussian process equipped with a spatial exponential correlation function, i.e. $\varepsilon_v \sim N(\mathbf{0}, \sigma_v^2 H(\phi))$ with $(H(\phi))_{ij} = \exp\{-\phi \|\mathbf{s}_i - \mathbf{s}_j\|\}$.

Finally, we address the change of support problem between the station data and the numerical model output by employing the downscaling approach [1, 5] and assuming a further measurement error model for $\tilde{V}(\mathbf{s}_j)$. Then, for each $j = 1, \dots, n$ belonging to grid cell A_i we have:

$$\tilde{V}(\mathbf{s}_j) = \tilde{R}(A_i) + \varepsilon_{\tilde{v}}(\mathbf{s}_j) \quad (4)$$

where $\varepsilon_{\tilde{v}}(\mathbf{s}_j)$ are independent $N(0, \sigma_{\tilde{v}}^2)$.

The hierarchical model is completed by the specification of the prior distributions for all the unknown parameters. Since it is not possible to consistently estimate the decay and variance parameter in a spatial model with a covariance function belonging to the Matérn family [8] as the exponential covariance function above, we fix the decay parameter and we put a prior distribution on $\sigma_{\tilde{v}}^2$. We place conjugate inverse gamma priors $IG(a_\sigma, b_\sigma)$, on the variance parameters σ_v^2 and $\sigma_{\tilde{v}}^2$ where we take a_σ and b_σ according to [3]. The prior distributions for τ^2 and τ_*^2 are specified as independent proper inverse gamma distributions $IG(a_\tau, b_\tau)$ with $a_\tau = 2$ and $b_\tau = 1$, implying that these variance components have prior mean 1 and infinite variance. Finally, a prior distribution for the variances $\sigma_r^2(A_i)$ is needed. The specification of such prior represents an important step to learn about the uncertainty associated with model outputs as we clarify in the next section, which starts by illustrating what we mean by uncertainty when dealing with a deterministic output.

3 Defining and modeling uncertainty in a model output

Let's concentrate about the “true” error, say $R(A_i) - \tilde{R}_{true}(A_i)$, where $\tilde{R}_{true}(A_i)$ is the true average value for the numerical model output over cell A_i . When this error is small for a grid cell, it implies small uncertainty associated to the numerical model prediction. Conversely, if the error is large then we would state high uncertainty for such cell.

To inform about the true error, we might compare the numerical model output with the validation data for each grid cell that contains a site, i.e. the *observed* residual $R(A_i) - V(\mathbf{s}_j)$. Then, high “disagreement” between $R(A_i)$ and $V(\mathbf{s}_j)$ for $\mathbf{s}_j \in A_i$ suggests high uncertainty in A_i . Conversely, we expect small uncertainty at grid cells where the disagreement between the numerical model output and the observation is low. However, the comparison between the average $R(A_i)$ with the point-level measurement $V(\mathbf{s}_j)$ is unfair because of the different spatial support of the two data sources. Moreover, the observed residuals are available only for grid cells where sites lie, while uncertainty should be attached to every grid cell.

To accomplish that, we consider the so-called *realized* residual, that is

$$\varepsilon_r(A_i) = R(A_i) - \tilde{R}(A_i) \quad (5)$$

which enables to reinterpret model (1). To further clarify, the true error for $R(A_i)$ is not known and, as usual within the Bayesian framework, we model unknowns as random quantities and look at their posterior distributions for inference. Under the specification above, we take $\tilde{R}(A_i)$ as the model for the truth and we look at the posterior distribution of each realized residual, $[\varepsilon_r(A_i) | Data]$.

The posterior variance $var(\varepsilon_r(A_i) | Data)$ provides the desired uncertainty, varying across grid cells. We can obtain our local uncertainties by composition sampling, i.e. drawing posterior samples of $\varepsilon_r(A_i)$ and then compute their variance. Equivalently, we compute the desired uncertainties as the posterior means $E(\sigma_r^2(A_i) | Data)$.

3.1 Spatial smoothing

Focusing on the posterior variance of the realized residuals (5), we attach higher uncertainty to grid cells for which we suppose larger differences between the model output and the true value. In addition, for a large realized residual at grid cell A_i , we expect a similar behavior in its neighborhood. In this way, we figure out some spatial smoothness of the uncertainties associated with the model output, based on the neighborhood structure of the grid cells. We formalize this belief assuming a CAR process for the logarithm of the latent variances $\sigma_r^2(A_i)$ of (5), that is

$$\log(\sigma_r^2(A_i)) | \{\log(\sigma_r^2(A_{i'})) : i' \neq i\} \sim N\left(\sum_{i' \sim i} \frac{\log(\sigma_r^2(A_{i'}))}{w_i}, \frac{\tau_*^2}{w_i}\right) \quad (6)$$

where, again, $i' \sim i$ identifies the cell $A_{i'}$ adjacent to cell A_i and w_i is the number of neighbors of A_i . The logCAR prior model in (6) is analogous to the spatial stochastic volatility approach [7] to capture spatial clustering in heteroscedasticity. Model (6) enables the borrowing of strength across grid cells inducing local spatial smoothing to uncertainty estimates towards their neighboring grid cells.

Along with the prior distributions for all the unknown parameters, the Bayesian hierarchical model is completely specified. The model is fitted using Markov Chain Monte Carlo (MCMC) algorithm.

4 Attaching uncertainty to the RUC output

We illustrate the data fusion model of Section 2 by quantifying the uncertainty associated with the RUC weather model output over the northeastern U.S., according to the proposal of Section 3. The RUC model (<http://ruc.noaa.gov/.RUC>) is a regional short-term weather forecast model of the Continental U.S. developed by the National Centers for Environmental Prediction (NCEP) to serve users needing frequently updated short-term weather forecasts. The RUC is run every hour producing 12-hour forecasts at 13 km spatial resolution.

We consider the daily forecast map on August 7th, 2011 obtained as average of 24 hourly temperature forecasts ($^{\circ}\text{F}$) provided by the RUC model from 00:00 to 23:00 on August 7th over the northeastern U.S., see left panel of Figure 1. There are 3,862 RUC grid cells spanning our study region. Moreover, land-based station data over U.S. is available. Here, we consider 24-hour averages of hourly temperature collected from 163 stations operating in the study region for the same period (Figure 1, left panel).

We fit model (1) - (4) along with prior specification (6). Posterior summaries of the unknown parameters are presented in Table 1. The main result of our work is shown in the right panel of Figure 1 which presents the estimated standard deviations deriving from $E(\sigma_r^2(A_i)|Data)$. The figure provides the uncertainty map associated with the RUC output and reveals the spatial variation of such uncertainty. We also note, as expected, that higher uncertainties are not associated with higher temperatures; in fact, we have no reason to believe that larger variances should be associated with larger responses.

As a conclusion, we have developed a hierarchical model to fuse a deterministic output and validation data in order to quantify the uncertainty associated with the model output, allowing for spatially smoothed error variances via a logCAR prior model. In this way we suggest how to enrich an established deterministic framework via an uncertainty assessment that may be appreciated by decision makers.

Acknowledgements The research of the first author was funded by a FIRB 2012 grant (project no. RBFR12URQJ) for research projects of national interest that was provided by the Italian Ministry of Education, Universities and Research.

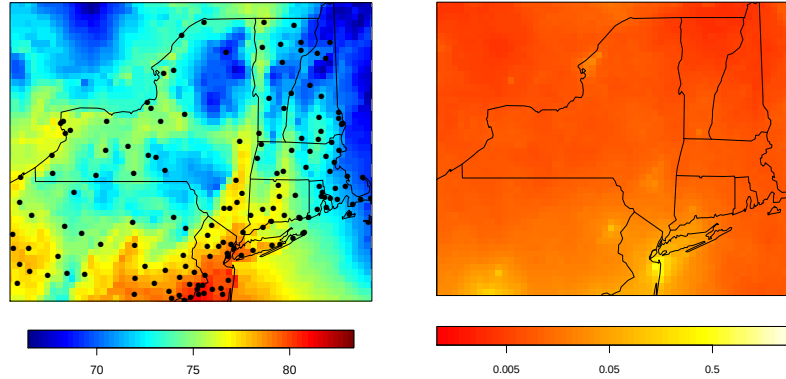


Fig. 1 Left panel: temperature stations (black dots) and daily RUC output on August 7th, 2011 over Northeastern US. Right panel: estimated standard deviations associated with the RUC output.

Table 1 Posterior summary of model parameters.

Parameters	Posterior mean and 95%CI
σ_v^2	0.902 (0.283, 2.118)
σ_v^2	1.701 (1.231, 2.242)
τ^2	0.728 (0.694, 0.764)
τ_s^2	10.868 (7.428, 14.91)

References

1. Berrocal, V.J., Gelfand, A.E., Holland, D.M.: Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*. **68**, 837–848 (2012)
2. Di Narzo, A.F., Cocchi, D.: A Bayesian hierarchical approach to ensemble weather forecasting. *J. R. Stat. Soc. C Appl. Stat.* **59**, 405–422 (2010)
3. Ghosh, S., Gelfand, A.E., Mølhave, T.: Attaching uncertainty to deterministic spatial interpolations. *Stat. Methodol.* **9**, 251–264 (2012)
4. Oakley, J.E., OHagan A.: Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*. **89**, 769–784 (2002)
5. Paci, L., Gelfand, A.E., Holland, D.M.: Spatio-temporal modeling for real-time ozone forecasting. *Spat. Stat.* **4**, 79–93 (2013)
6. Raftery, A.E., Givens, G.H., Zeh, J.E.: Inference from a deterministic population dynamics model for bowhead whales. *J. Am. Stat. Assoc.* **90**, 402–416 (1995)
7. Yan, J.: Spatial stochastic volatility for lattice data. *J. Agric. Biol. Environ. Stat.* **12**, 25–40 (2007)
8. Zhang, H.: Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Stat. Assoc.* **99**, 250–261 (2004)