

# Comments on: Comparing and selecting spatial predictors using local criteria

Finn Lindgren

Published online: 29 November 2014  
© Sociedad de Estadística e Investigación Operativa 2014

**Abstract** Large spatial data sets require innovative techniques for computationally efficient statistical estimation. In this comment some aspects of local predictor selection are explored, with a view towards spatially coherent field prediction and uncertainty quantification.

**Keywords** Stochastic spatial models · Non-stationarity · Stochastic partial differential equation · Predictive scoring rules · Local model selection

## 1 Motivation

The paper by [Bradley et al. \(2014\)](#) investigates the use of local selection of spatial predictors to aid the analysis of large spatial data sets. The idea is that, even if globally constructed predictors may not be locally optimal individually, given a set of different predictors one can select the locally optimal predictor for each location, based on a validation criterion. In this comment, I will discuss some possible generalisations, aimed at the more difficult problem of constructing spatially consistent representations of uncertainty. As the authors rightly note, the sheer size of a data set does not imply that it is also necessarily spatially dense, so Bayesian process prior models (or essentially equivalent loss-function regularisation methods) are still useful and sometimes necessary tools for practical data analysis.

The computation time for the SPD and LKR methods, and likely similarly for the other methods considered in the paper, is dominated by the numerical optimisation

---

This comment refers to the invited paper available at doi:[10.1007/s11749-014-0415-1](https://doi.org/10.1007/s11749-014-0415-1).

---

F. Lindgren (✉)  
Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK  
e-mail: [f.lindgren@bath.ac.uk](mailto:f.lindgren@bath.ac.uk)

of a very flat likelihood. While the full SPD estimations that we will explore here take on average almost 3 min each, when using 20,000 observations to predict onto 26,002 other locations using the INLA R package (Rue et al. 2013), each individual kriging field evaluation takes less than 1 s. For non-stationary models, global parameter estimation completely dominates the computational effort (Aune et al. 2014), and local methods become attractive, since estimating several smaller models can be faster than estimating a single large model. It is therefore slightly surprising that the paper does not consider that additional step, but only uses the global estimates to do local selection. With the current popularity of quantifying uncertainty with spatially coherent samples from conditional distributions, which was already a natural thing to do in Bayesian settings, the problem of estimating probabilistic non-stationary models therefore remains. However, the simplicity of the local predictor selection approach makes it an attractive starting point for model based methods, both parametric and non-parametric. The aim in this comment is to (1) explore selection criteria using predictive distribution information, and (2) assess to what extent the selected local predictors associate with a true non-stationary random process model.

## 2 A constructed test model

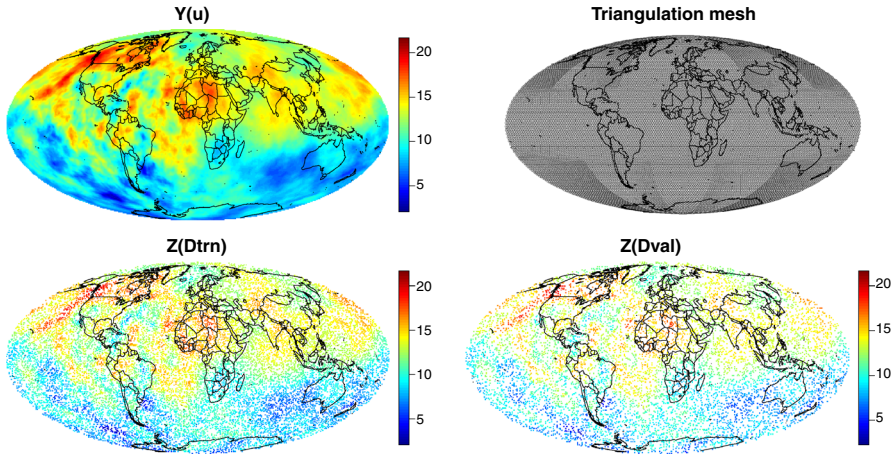
The term *white noise* used in the paper for the measurement noise process  $\varepsilon(\mathbf{u})$  is slightly problematic, since it typically implies a spatially defined spectral measure representation, which the measurement noise process does not have. In spatially continuous contexts, white noise is typically defined precisely as a spectrally white random measure, on  $\mathbb{R}^d$  informally identified with the derivative of a Brownian sheet, which does not have a practical point-wise meaning. The distinction becomes important when we now consider a version of the stochastic partial differential equation used to construct the SPD spatial predictor in the paper. As shown by Whittle (1954, 1963), the Matérn correlation with spatial scale parameter  $\kappa$  can be identified with the solutions to a fractional stochastic partial differential equation, which in turn can be closely approximated by an expansion in compactly supported basis functions with Markov-dependent coefficients (Lindgren et al. 2011). A non-stationary extension to the sphere is given by

$$\left(\kappa(\mathbf{u})^2 - \nabla \cdot \nabla\right) Y(\mathbf{u}) \, d\mathbf{u} = \kappa(\mathbf{u}) \mathcal{E}(d\mathbf{u}), \quad \mathbf{u} \in \mathbb{S}^2, \quad (1)$$

where  $\mathcal{E}(\cdot)$  is a zero mean Gaussian random measure such that for any pair of measurable sets  $A, B \subseteq \mathbb{S}^2$ ,  $\text{cov}(\mathcal{E}(A), \mathcal{E}(B)) = \frac{4\pi}{\tau} |A \cap B|$ . Note that (1) should be interpreted only as shorthand notation for a proper stochastic integral equation. The parameter  $\tau$  is nominally the precision (inverse variance) of  $Y(\mathbf{u})$ , but the true variance will depend on  $\kappa(\cdot)$ . Small  $\kappa(\cdot)$  increases the variance because of the spherical topology, and the variance is also somewhat dependent on the derivatives of  $\kappa(\cdot)$ .

As in the CO<sub>2</sub> example in the paper, we consider the model structure

$$Z(\mathbf{u}) = Y(\mathbf{u}) + \varepsilon(\mathbf{u}),$$



**Fig. 1** The true field  $Y(\cdot)$ , the computational triangulation mesh for the SPDE/GMRF construction, the training data  $Z(\mathbf{s}_i)$ ,  $\mathbf{s}_i \in D^{trn}$ , and the validation data  $Z(\mathbf{s}_i)$ ,  $\mathbf{s}_i \in D^{val}$

where  $Z(\mathbf{u})$  is observed at a set of locations  $D$ , which we split into  $n = 20,000$  training locations  $D^{trn}$  and  $m = 10,000$  validation locations  $D^{val}$ . The process  $Y(\cdot)$  is interpreted as a hidden, or *latent*, Gaussian random field, and  $\varepsilon(\cdot)$  is interpreted as independent zero mean Gaussian measurement noise with location dependent variance  $\text{var}(\varepsilon(\mathbf{u})) = \sigma_\varepsilon^2 v(\mathbf{u})$ . For simplicity, we assume that  $v(\cdot)$  is known. Note that, in reality,  $\varepsilon(\cdot)$  lives only on  $D$ , but we treat its *potential* value at arbitrary locations as real to simplify the predictive distribution formulations. The latent process  $Y(\mathbf{u})$  is modelled as the sum of a weighted sum of rotationally invariant spherical harmonics up to order 2 and a realisation of the SPDE in (1).

The simulated test case is constructed so that  $\log(v(\cdot))$  varies linearly in  $\sin(\text{lat})$ , starting at  $-\log(16)$  at the north pole, and increasing to  $\log(16)$  at the south pole. Similarly, the spatial range is smoothly varying in  $\sin(\text{lon})$ , with minimum 15 at (90W, 0N) and maximum 60 at (90E, 0N). The true field  $Y(\cdot)$ , the computational triangulation mesh, the training data, and the validation data, are shown in Fig. 1. The triangulation is a quasi-regular mesh with 16,002 nodes based on a subdivided icosahedron, and is used to define the finite elements for the Gaussian Markov random field approximation at the heart of the SPDE/GMRF modelling approach. Further details of the simulation study will be given in Sect. 4.

### 3 Alternative selection criteria

The paper uses *punctured* local predictor selectors, constructed so that the behaviour is different if for some reason one wants to predict at one of the validation data locations. The reason why this construction is used appears to be to guarantee that the locally selected predictor improves on the sum of the validation errors. However, when treating the hidden process  $Y(\mathbf{u})$  as a field on a continuous domain, a more relevant quantity might be the full spatial average of the expected prediction error, which is unaffected by changes on a null set.

From this point of view, the simple local predictor, SLS, is not meaningfully different from the global selection method GSP, and the similar puncture of the MWS and NNS methods only serves to retain some of the discontinuities they were meant to remove. Here we will instead consider a non-punctured version of the moving-window predictor, with added distance weighting to further stabilise the local predictor.

### 3.1 Distance weighted selectors

Let  $W(\mathbf{u}, \mathbf{s})$  be a non-negative weighting function, defined for all  $\mathbf{u}$  in the continuous domain, and all  $\mathbf{s} \in D^{\text{val}}$ . The local validation sets  $H \subseteq D^{\text{val}}$  and associated locally renormalised weights  $\bar{W}$  can then be defined through

$$H(W, \mathbf{u}) = \{\mathbf{s} : \mathbf{s} \in D^{\text{val}}, W(\mathbf{u}, \mathbf{s}) > 0\}$$

$$\bar{W}(\mathbf{u}, \mathbf{s}) = \frac{W(\mathbf{u}, \mathbf{s})}{\sum_{\mathbf{s}' \in H(W, \mathbf{u})} W(\mathbf{u}, \mathbf{s}' )}$$

With the exclusion of the puncturing, the unweighted MWS method in the paper corresponds to using the weight function

$$W_0(\mathbf{u}, \mathbf{s}) = \begin{cases} 1, & \text{if } \|\mathbf{s} - \mathbf{u}\| \leq w, \\ 0, & \text{otherwise,} \end{cases}$$

with  $\bar{W}_0(\mathbf{u}, \mathbf{s}) = 1/|H(W_0, \mathbf{u})|$  for all  $\mathbf{s} \in H(W_0, \mathbf{u})$ . We now introduce the distance weighting

$$W_1(\mathbf{u}, \mathbf{s}) = \max(0, 1 - \|\mathbf{s} - \mathbf{u}\|/w)$$

as a simple alternative, that gives a spatially less abrupt reaction to outlier observations. A similarly weighted version of the  $g$ -nearest-neighbour method, NNS, can also be formulated in this manner, but we refrain from doing that here, and note that the Voronoi method, VPS, is identical to NNS with  $g = 1$ .

### 3.2 Alternative scoring rules

With the understanding that all the local selection criteria only consider the validation data set, we can consider alternative measures of validation error. As noted in the discussion section of the paper, using the standard errors of the selected predictor, when available, as estimates of the standard errors of the resulting LSP may lead to underestimating the uncertainty. Also, in order to use local selection as part of a local model selection procedure, it seems reasonable to consider more aspects of the predictors than their point estimate of the hidden fields. Probabilistic prediction estimates, such as those based on Bayesian hierarchical models, contain additional information that can be used to inform the local selector. We therefore turn to [Gneiting et al. \(2007\)](#) for inspiration on alternative proper scoring rules that are able to utilise

such information. Note that this does not in itself require distributional modelling assumptions to be made, but it does make sure that the scoring rules are consistent with distributional aspects of the spatial predictions.

First, we reformulate the LSVE metric from the paper into an equivalent root mean squared error LRMSE, and introduce a similar *mean absolute error* LMAE:

$$\begin{aligned} \text{LRMSE}_Z(\mathbf{u}; W, \hat{Y}^{(k)}) &= \left\{ \sum_{\mathbf{s} \in H(W, \mathbf{u})} \bar{W}(\mathbf{u}, \mathbf{s}) \left( Z(\mathbf{s}) - \hat{Y}^{(k)}(\mathbf{s}) \right)^2 \right\}^{1/2} \\ \text{LMAE}_Z(\mathbf{u}; W, \hat{Y}^{(k)}) &= \sum_{\mathbf{s} \in H(W, \mathbf{u})} \bar{W}(\mathbf{u}, \mathbf{s}) \left| Z(\mathbf{s}) - \hat{Y}^{(k)}(\mathbf{s}) \right| \end{aligned}$$

For methods that generate prediction distributions, let  $\hat{Y}(\mathbf{u})$  and  $\hat{V}_Z(\mathbf{u})$  be the data predictive expectation and variance. In a Bayesian setting, take for example

$$\begin{aligned} \hat{Y}(\mathbf{u}) &= E(Y(\mathbf{u}) \mid Z(\mathbf{s}_i), \mathbf{s}_i \in D^{\text{tm}}), \\ \hat{V}_Z(\mathbf{u}) &= \text{var}(Y(\mathbf{u}) \mid Z(\mathbf{s}_i), \mathbf{s}_i \in D^{\text{tm}}) + \sigma_\varepsilon^2 v(\mathbf{u}), \end{aligned}$$

which is readily available in the estimation output of the INLA package. In frequentistic settings, the Markov representation of the SPDE model provides an efficient way to calculate the kriging variances, which then replace the posterior variances in the Bayesian formulation. Following the treatment by [Gneiting et al. \(2007\)](#), the negatively orientated *continuous ranked probability score* (CRPS) is given by

$$\text{CRPS}^*(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}(y \geq x))^2 dy$$

for the cumulative probability function  $F$  of a probabilistic forecast of an observation  $x$ . When  $F$  describes a pure point estimate, the CRPS is equal to the absolute error, and acts as a natural generalisation for probabilistic forecasts. The CRPS has a simple closed form expression in the Gaussian case, and we can define a local predictor selector criterion via

$$\text{LCRPS}_Z(\mathbf{u}; W, \hat{Y}^{(k)}) = \sum_{\mathbf{s} \in H(W, \mathbf{u})} \bar{W}(\mathbf{u}, \mathbf{s}) \text{CRPS}^*\left(N\left(\hat{Y}^{(k)}(\mathbf{s}), \hat{V}_Z^{(k)}(\mathbf{s})\right), Z(\mathbf{s})\right).$$

Another option is derived from the moment based logarithmic score,

$$\text{LOGS}^*((\mu, \sigma^2), x) = \frac{(x - \mu)^2}{\sigma^2} + \log \sigma^2,$$

which favours predictive distributions where  $\sigma^2$  is close to  $(x - \mu)^2$ .

## 4 Results

In the CO<sub>2</sub> example in the paper, the spatial predictors used as input to the local selection procedure all produced similar spatial fields, with the exception of the SPD estimate, which smoothed out the all fine scale structure. Despite this, the SPD predictor was chosen nearly as often as the FRK predictor. However, despite extensive testing, I have been unable to construct a test case producing such overly smooth SPD predictors, which indicates a problem with the parameter optimisation settings used in the paper. One clear difference between the FRK and SPD results is that the variable observation noise model was not used in the SPD case, possibly leading to an unreasonably high overall estimated noise-to-signal ratio, as also indicated by the small Lag-1 semivariogram for the SPD estimate (Bradley et al. 2014, Table 3). Speaking against this hypothesis is that the LKR estimate also did not use the full noise model, and was seemingly unaffected. In the simulation test case here I used a spatially variable observation noise model both for the SPD and LKR predictors, via `inla(..., scale=1/v)` in INLA and `LKrig(..., weights=1/v)` in LatticeKrig (Nychka et al. 2013, 2014), for a known weight function  $v(\cdot)$ . One could conceivably include a semi-parametric estimate of  $v(\cdot)$  by applying the full force of the general latent Gaussian model structure available in INLA, since such a model can be programmed as a special case of the existing internal representation of non-stationary SPDE precision models.

In order to evaluate the local selection procedure on the simulated model from Sect. 2, seven global predictors were constructed using the training data set:

$$\widehat{Y}^{(0)}(\cdot) = \text{INLA of the true model,}$$

$$\widehat{Y}^{(k)}(\cdot) = \text{INLA of stationary models, for ranges } r_k = 7.5, 15, 30, 60, \text{ and } 120,$$

$$\widehat{Y}^{\text{LKR}}(\cdot) = \text{LatticeKrig estimate, with default long range.}$$

The ranges  $r_k$  were chosen so that the true model range lies inside the span of  $r_2, r_3$ , and  $r_4$ , with  $r_1$  being clearly shorter than the smallest range, and  $r_5$  being clearly larger than the largest. The resulting spatial prediction fields are shown in Fig. 2 (left). Since  $\widehat{Y}^{\text{LKR}}(\cdot)$  was similar to the longer range SPD estimates, and the current standard error implementation in LatticeKrig is comparatively slow, it was excluded from further analysis, to allow fair comparisons for the scores based on predictive distributions, CRPS and LOGS.

The aim is to compare the behaviour of the LSP predictor based on the stationary models used for  $\widehat{Y}^{(k)}, k = 1, \dots, 5$ , with the predictor based on the full non-stationary model,  $\widehat{Y}^{(0)}$ . For  $j = 0$  and 1 (for the two weighting schemes  $W_0$  and  $W_1$  in Sect. 3, with radius  $w = 5^\circ$ ), the LSP construction proceeded as follows:

$$\widehat{k}_{\text{LRMSE}(W_j)}(\mathbf{u}) = \arg \max_{k \in \{1, 2, 3, 4, 5\}} \text{LRMSE}_Z(\mathbf{u}; W_j, \widehat{Y}^{(k)}),$$

$$\widehat{Y}_{\text{LRMSE}(W_j)}(\mathbf{u}) = \widehat{Y}^{(\widehat{k}_{\text{LRMSE}(W_j)})}(\mathbf{u}).$$

The procedure was then repeated for LMAE, LCRPS, and LLOGS, generating a total of eight LSP estimates, all based on  $\widehat{Y}^{(k)}(\cdot)$ ,  $k = 1, \dots, 5$ , only. The resulting predictor indices  $\widehat{k}(\cdot)$  are shown in Fig. 2 (right). In contrast to the CO<sub>2</sub> results in the paper, these results exhibit a much stronger spatial coherence. This is to be expected, as the input predictors were chosen to have fixed ranges covering the true model ranges, and the overall effect is that the model with range  $r_1$  was chosen in 33 % of the locations, and  $r_5$  was chosen in 55 %, in a pattern matching the transition from short range in the western hemisphere to long range in the eastern hemisphere. Note however that these two predominantly chosen models are both outside the spread of the true model range function, so even though they produced the lowest scores, they should not be mistaken for good estimates of the true model.

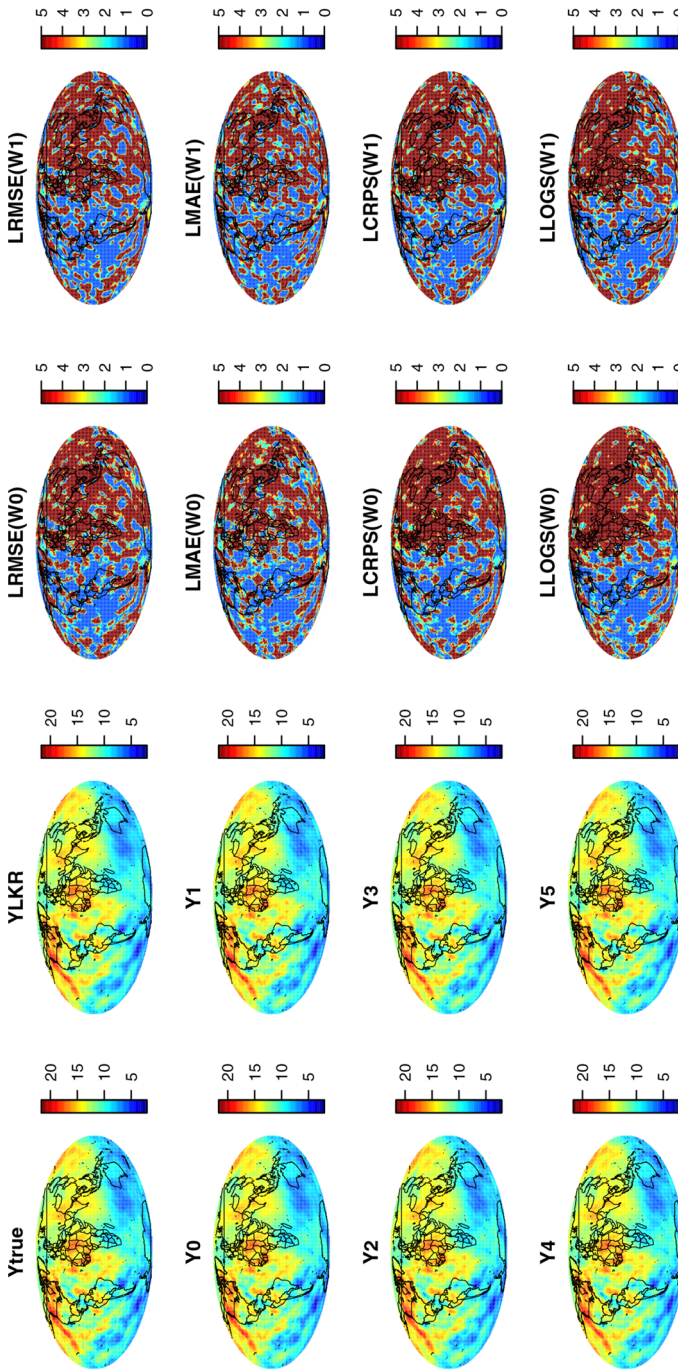
In order to evaluate the behaviour of the LSP under the alternative scoring rules, the validation scores of the final predictors were calculated. Figure 3 (left) shows the differences between the scores for each LSP and the full model predictor for  $Z(\cdot)$ , and Table 1 (left) show the globally averaged scores. The scores are very close, giving the appearance that the LSP method was able to construct reasonable predictions under each scoring rule. However, in a practical application the focus would normally be on predicting the hidden process  $Y(\cdot)$  itself, and *not* on predicting noisy data. As shown in the rightmost parts of Fig. 3 and Table 1, the global non-stationary model is clearly better than the LSP at producing  $Y(\cdot)$ -predictions, in particular with respect to the scores sensitive to the full predictive distributions, LCRPS and LLOGS. The effect is most clearly seen in the eastern hemisphere, where the true model has long spatial correlation range.

Finally, since the fine-scale detail made visual assessment of some of the aspects of the estimates difficult, the procedure was repeated using weighting windows of radius  $w = 10$ , which revealed that the distance weighting scheme,  $W_1$ , as intended is indeed less abruptly sensitive to outliers than the flat weighting  $W_0$ . Also as expected, the spatial coherence in the predictor selections increased, and the scoring behaviour was similar to the result presented here.

## 5 Discussion

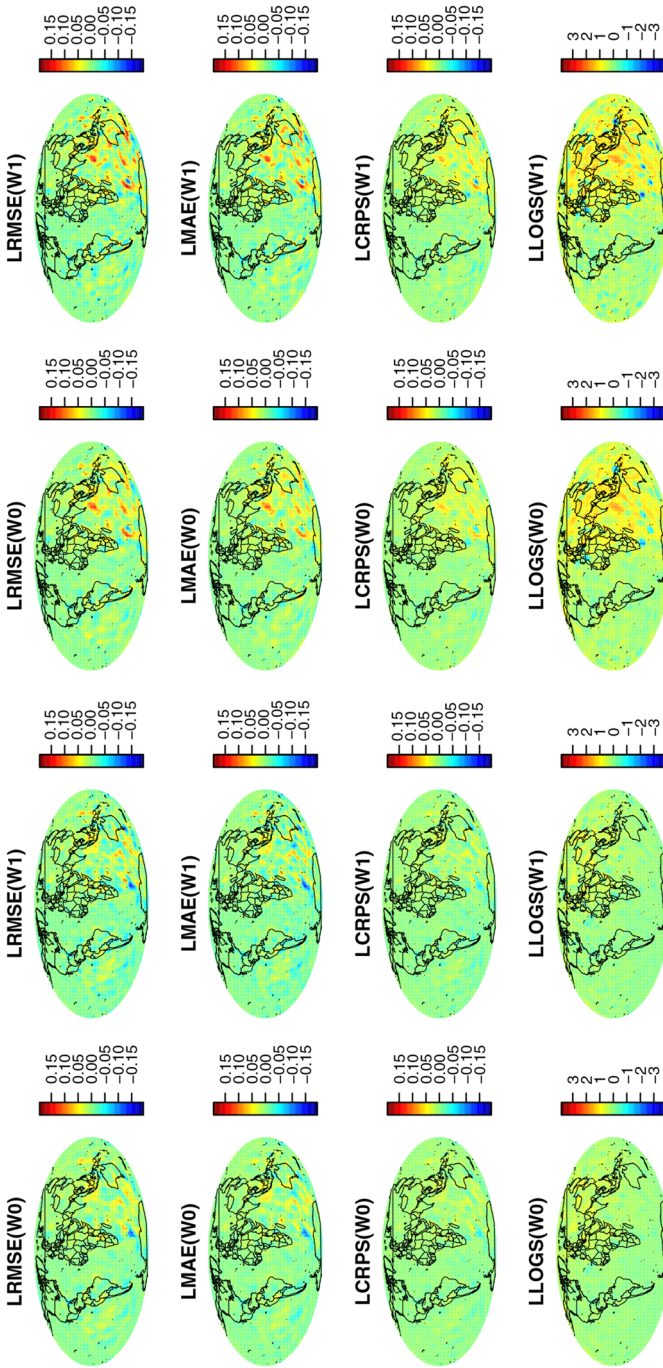
As observed in Sect. 4, the gain in local prediction error using the LSP method can be very small, compared with using a more problem adapted model, but does show great promise for cases when such models are too computationally expensive. One benefit is robustness to mis-specified or overly simple global prediction estimators, and using a wide variety of simple predictors may be faster than using a more complex model. However, the results also show that the LSP in its current form may not be adequate for generating suitable uncertainty estimates, an issue touched upon briefly in the final discussion of the [Bradley et al. \(2014\)](#) paper.

A worthwhile direction to explore is to replace the simple global estimators with equally simple but *local* estimators as input to the LSP, that may have a better chance of capturing non-stationary behaviour in both mean and variance, as well as being computationally more efficient. Selection criteria based on local joint predictive distributions may also be necessary to capture the spatially coherent structure of the



**Fig. 2** *Left panel* true spatial field  $Y(\cdot)$  and spatial prediction fields  $\hat{Y}^{\text{LKR}}(\cdot)$ ,  $\hat{Y}^{(k)}(\cdot)$ ,  $k = 0, \dots, 5$ . *Right panel* selected predictor indices  $\hat{k}(\cdot)$  for each of the 8 LSP methods, all based on  $\hat{Y}^{(k)}(\cdot)$ ,  $k = 1, \dots, 5$





**Fig. 3** Local validation score differences for prediction of  $Z(\cdot)$  (left) and  $Y(\cdot)$  (right). The first plot shows  $\text{LRMSE}_Z(\mathbf{u}; W_0, \hat{Y}_{\text{LRMSE}}) - \text{LRMSE}_Z(\mathbf{u}; W_0, \hat{Y}^{(0)})$ , and analogously for the other plots. The scores are comparable for  $Z(\cdot)$ -prediction, but for  $Y(\cdot)$ -predictions there is a benefit to using the global non-stationary model

**Table 1** Global average validation scores for prediction of  $Z(\cdot)$  and  $Y(\cdot)$ 

Type	$Z \mid \hat{Y}^{(0)}$	$Z \mid \hat{Y}_{LSP}$	$Y \mid \hat{Y}^{(0)}$	$Y \mid \hat{Y}_{LSP}$
LRMSE( $W_0$ )	0.321	0.323	0.170	0.176
LRMSE( $W_1$ )	0.319	0.318	0.168	0.172
LMAE( $W_0$ )	0.259	0.260	0.134	0.140
LMAE( $W_1$ )	0.259	0.257	0.134	0.138
LCRPS( $W_0$ )	0.183	0.186	0.095	0.105
LCRPS( $W_1$ )	0.183	0.184	0.095	0.104
LLOGS( $W_0$ )	-1.694	-1.569	-2.852	-2.435
LLOGS( $W_1$ )	-1.693	-1.580	-2.856	-2.441

The first value of the top row shows  $\text{LRMSE}_Z(\mathbf{u}; W_0, \hat{Y}^{(0)})$ , and analogously for the other entries. The scores are comparable for  $Z(\cdot)$ -prediction, but for  $Y(\cdot)$ -predictions there is a benefit to using the global non-stationary model, in particular for LCRPS and LLOGS

hidden process. Of the scores investigated in this comment, the logarithmic score generalises naturally to multivariate distributions, and even generalises to fully Bayesian settings, in the form of a negated log-posterior density.

The analysis code is available as online supplementary material. All the computations and timings (3 min each for the 6 full SPD estimates, <1 s per SPD kriging evaluation, and 9 min for a full LatticeKrig estimate, 2 min in total for computing all LSPs and diagnostic scores) were generated on a quad core 2.2 GHz Intel Core i7-4702HQ laptop, with 16 GB memory.

**Acknowledgments** I want to thank the editors for the invitation to comment, and Jonathan Bradley, Noel Cressie, and Tao Shi for producing a paper that it was well worth the effort on which to comment.

## References

- Aune E, Simpson DP, Eidsvik J (2014) Parameter estimation in high dimensional Gaussian distributions. *Stat Comput* 24(2):247–263
- Bradley JR, Cressie N, Shi T (2014) Comparing and selecting spatial predictors using local criteria. *TEST*. doi:[10.1007/s11749-014-0415-1](https://doi.org/10.1007/s11749-014-0415-1)
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102(477):359–378
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *J R Stat Soc Ser B Stat Methodol* 73:423–498
- Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2014) A multi-resolution Gaussian process model for the analysis of large spatial data sets. *J Comp Graph Stat*. doi:[10.1080/10618600.2014.914946](https://doi.org/10.1080/10618600.2014.914946) (available online)
- Nychka D, Hammerling D, Sain S, Lenssen N (2013) LatticeKrig: multiresolution kriging based on Markov random fields. <http://cran.r-project.org/web/packages/LatticeKrig/>. Accessed 2 Aug 2014
- Rue H, Martino S, Lindgren F, Simpson D, Riebler A (2013) R-INLA: approximate Bayesian inference using integrated nested Laplace approximations, Trondheim, Norway. <http://www.r-inla.org/>. Accessed 2 Aug 2014
- Whittle P (1954) On stationary processes in the plane. *Biometrika* 41:434–449
- Whittle P (1963) Stochastic processes in several dimensions. *Bull Int Stat Inst* 40:974–994