Spatial Statistics 32 (2019) 100368

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Evaluation of empirical Bayesian kriging

Konstantin Krivoruchko[1], Alexander Gribov *

Esri, 380 New York Street, Redlands, CA 92373-8100, USA

ARTICLE INFO

Article history:
Received 23 January 2019
Received in revised form 19 May 2019
Accepted 13 June 2019
Available online 27 June 2019

Keywords:
Interpolation
Empirical Bayesian kriging
Prior distribution
Simulation experiment
Big data

ABSTRACT

Using Bayesian bootstrap as the informative prior distribution is a key feature of the empirical Bayesian kriging method. We evaluate correctness of the method using two large simulation experiments. We also compare empirical Bayesian kriging performance with several popular geostatistical models implemented in R using two real datasets: the average annual precipitation in South Africa and the daytime land surface temperature satellite data. Results show that empirical Bayesian kriging performs better than other predictors, and increasingly so with data complexity.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Online automatic mapping and big data interpolation are standard requirements for modern geographic information system (GIS) technology, and Esri's ArcGIS Geostatistical Analyst (Esri, 2019b) provides a series of empirical Bayesian kriging (EBK) models that do not require specification of the prior distributions for the model parameters; allow moderate local and large global data non-stationarity; locally transform data to Gaussian distribution, if needed; allow for varying measurement error; use covariates; and work reasonably fast and produce reliable outputs with default parameters.

Available documentation on EBK (Esri, 2019a,c; Krivoruchko, 2012; Krivoruchko and Frączek, 2015) largely focuses on the software usage, and there is a need for a more detailed description of the EBK algorithms. In this paper, we discuss the key feature of the EBK method—construction and use of informative prior distribution, emphasizing the empirical comparison of the prediction accuracy with several commonly used alternatives. Other EBK features are discussed in the companion papers (Gribov and Krivoruchko, 2019; Krivoruchko and Gribov, 2019).

* Corresponding author.
E-mail addresses: kkrivoruchko.work@gmail.com (K. Krivoruchko), agribov@esri.com (A. Gribov).
[1] Permanent address: PO Box 9972, Redlands, CA 92375-3172, USA.

https://doi.org/10.1016/j.spasta.2019.100368
2211-6753/© 2019 Elsevier B.V. All rights reserved.

EBK consists of two geostatistical models: the intrinsic random function kriging (IRFK, see Yaglom (1955), Kitanidis (1993, 1997, 1999), Chilès and Delfiner (1999)) and linear mixed model (LMM, in geostatistical literature, it is also called simple kriging with external trend, see Banerjee et al. (2003), Schabenberger and Gotway (2004), Diggle and Ribeiro (2007)). Since the same algorithms are used for IRFK and LMM model fitting, we combined two different processes (stationary and stationary increments) in one computational model.

$$z_i = y(s_i) + \epsilon_i, i = \overline{1 \dots K}, \tag{1}$$

where z_i is the measured value at the observed location s_i, $y(s)$ is the Gaussian process under study at the location s, ϵ_i is the measurement error, and K is the number of measurements. Note that in the text below $\mathbf{z}$ is a vector formed by z_i.

In the case of LMM, the mean of the process $y(s)$ is a constant or it is described by covariates.

The assumption that the observed process follows Gaussian distribution is often violated; therefore, in the case of LMM, transformation to Gaussian distribution can be used. More general model with empirical data transformation is discussed in details in Gribov and Krivoruchko (2019).

The spatial process in (1) is described by generalized covariance or covariance parameters, regression coefficients, variances of the measurement errors, and the data transformation parameters, if required.

The goal is to estimate the process without measurement error $y(s)$. EBK modeling consists of the following internal steps:

1. Parameters of the spatial process Θ, including the semivariogram model, are estimated from the data.
2. Using Θ, new values are unconditionally simulated at each of the input data locations K_{sim} times.
3. New parameters Θ_i, $i = \overline{1 \dots K_{sim}}$, are estimated from the simulated data. A histogram of Θ_i can be used as an approximation of prior distribution.
4. Call Θ_i, $i = \overline{1 \dots K_{sim}}$, the empirical prior distribution. We assume that the model parameters can take only Θ_i values, that is $f\left(\Theta'|\mathbf{z}\right) = 0$ for $\Theta' \neq \Theta_i$.
5. A weight for each simulated model is calculated using Bayes' rule, $w_i \propto f(\mathbf{z}|\Theta_i)$, $w_i = \frac{f(\mathbf{z}|\Theta_i)}{\sum_{i=1}^{K_{sim}} f(\mathbf{z}|\Theta_i)}$, $\sum_{i=1}^{K_{sim}} w_i = 1$, where $f(\mathbf{z}|\Theta_i)$ is the conditional probability of the data $\mathbf{z}$ given the model parameters Θ_i. It shows how likely the observed values can be generated from the estimated model.
6. Predictions and prediction standard errors are produced at the specified locations using the following formulas:

$$\mathrm{E}\left[y(s)|\mathbf{z}\right] = \hat{y}(s) = \sum_{i=1}^{K_{sim}} \left(w_i \cdot \mathrm{E}\left[y(s)|\mathbf{z}, \Theta_i\right]\right),$$

$$\mathrm{Var}\left[y(s)|\mathbf{z}\right] = \mathrm{E}\left[\left(y(s) - \hat{y}(s)\right)^2 |\mathbf{z}\right] = \sum_{i=1}^{K_{sim}} \left(w_i \cdot \left(\mathrm{Var}\left[y(s)|\mathbf{z}, \Theta_i\right] + \left(\mathrm{E}\left[y(s)|\mathbf{z}, \Theta_i\right] - \hat{y}(s)\right)^2\right)\right),$$

where $\mathrm{E}\left[y(s)|\mathbf{z}, \Theta_i\right]$ and $\mathrm{Var}\left[y(s)|\mathbf{z}, \Theta_i\right]$ are given by kriging equations (Gribov and Krivoruchko, 2019).

The default univariate EBK model is the IRFK of order zero, and the spatial correlation model is the power model $n - b|h|^{\alpha}$, where n is the nugget, b is the slope coefficient, h is the distance between the locations of the process under study, and $\alpha \in (0.25, 1.75)$ is the power parameter. This correlation model corresponds to fractional Brownian motion (Chilès and Delfiner, 1999). Note that theoretical margins for α parameter are $[0; 2)$. We use narrower interval because for very small parameter value the number of pairs can be too small for accurate parameter estimation and value close to the theoretical maximum may cause calculation instabilities, see also (Gribov and Krivoruchko, 2019).

In this paper, we use both IRFK and LMM. In the latter case, spatial correlation of $y(s)$ is described by the K-Bessel covariance function (Johnston et al., 2001) (it is also called Matern covariance in statistical literature).

When the dataset is large (in our implementation, when the number of observations is larger than 1000) the software creates subsets with a specified number of samples (it was decided that the number varies from 20 to 1,000), as shown in Gribov and Krivoruchko (2019, section 2.1). A user defined data subset option is also provided. Model fitting in the EBK is performed for data subsets, and predictions are made using the weighted sum of the models from the possibly overlapping or disjoint nearby subsets.

EBK predicts values at the specified locations using a limited number of the nearest observations. When the prediction search neighborhood includes observations from several data subsets, the data contribution to the prediction is divided among the subsets (Gribov and Krivoruchko, 2019).

EBK depends essentially on step 4, above: *Call* Θ_i, $i = \overline{1 \dots K_{sim}}$, *the empirical prior distribution*. The goal of this paper is to compare EBK informative prior distribution with several alternatives using two large simulation experiments (Sections 2 and 3). Two possible alternatives are suggested by modification of the EBK steps above:

1. Simply stop at the first step and use parameters of the spatial process Θ estimated with the restricted maximum likelihood (REML) algorithm. In the models' comparison below, we call the resultant model estimated maximum likelihood (EML).
2. Call Θ_i, $i = \overline{1 \dots K_{sim}}$, the empirical *posterior* distribution, as suggested in Pilz and Spöck (2008), and skip step 5. We will call the resultant model empirical Bayesian bootstrap (EBB).

Note that the EBK implementation does not provide EML and EBB models.

In Section 3, performance of EBK, EML, and EBB models with the same parameters is compared with three models with better expected performance. Other alternatives can be found in the kriging models programmed in freeware R software, and in Section 4 we use modified predictive processes (MPP) (Finley et al., 2015) (implemented in the R package (Finley and Banerjee, 2017)) and a stochastic partial differential equation (SPDE) method (Lindgren et al., 2011) (implemented in the R package (Rue et al., 2012)), two packages in which kriging models assume noninformative prior distribution. Additionally, we use three non-Bayesian models allowing medium-sized data interpolation, namely, lattice kriging (LTK) (Nychka et al., 2016) (implemented in the R package LatticeKrig (Nychka et al., 2017)), fixed ranked kriging (FRK) (Cressie and Johannesson, 2008) (implemented in the R package (Zammit-Mangion and Cressie, 2017)), and local polynomial interpolation (LPI) (Gribov and Krivoruchko, 2011) (implemented in ArcGIS Geostatistical Analyst (Esri, 2019d)). LPI is conceptually similar to the universal kriging model when all spatial variation is described by local trend and the semivariogram model is a nugget effect. In the case study in Section 4, the LPI model uses the x- and y-coordinates as statistically independent variables (the first-order local polynomials). Finally, in Section 5, we demonstrate that EBK models do a good job of providing both self-tuned and hand-tuned outcomes in comparison with 12 other hand-tuned kriging model outcomes shown in the very recently published satellite data prediction comparison experiment (Heaton et al., 2018).

The data used for the models comparison in Sections 3, 4, and 5 are provided in the supplementary materials with instructions on how to make predictions using one kriging model implemented in R. Instructions for automating EBK usage with Python are also provided.

Fig. 1 shows a diagram of the EBK usage. First, the user selects the observations (the dependent variable in the case of LMM with covariates) and their individual measurement errors, if available. Next, user-defined subsets are selected or parameters for new subsets creation are accepted or defined. Then the user can run the default EBK model or change the generalized covariance model, the parameters of the search neighborhood, the number of simulations, and the model's output before running the model. The software creates a geostatistical layer, a data structure that stores the path to the source data and the parameters used to interpolate values. The geostatistical layer is used to create maps (prediction, prediction standard error, quantile, and probability that a specified threshold value is exceeded) and for evaluating (or updating) previously created models in the Geostatistical Wizard. Additionally, model diagnostics for point locations and a summary for each data

subset can be produced. Two important options, the data transformation to Gaussian distribution and the chordal distance metrics instead of straight lines are used automatically when the LMM is used without covariates and the data coordinates are latitude and longitude, correspondingly. The EBK implementation is largely discussed and illustrated in Gribov and Krivoruchko (2019).

2. Mean and standard deviation reconstruction from a small amount of data generated from normal distribution

In this section, we investigate the key feature of the EBK method: the use of the informative data-dependent empirical prior distribution. We compare EBK method with the following approaches that also do not require specification of a prior distribution: the maximum likelihood estimation, noninformative prior distribution, and the empirical posterior distribution (Pilz and Spöck, 2008).

Four methods are used to estimate the mean and the standard deviation when n random variables are generated from the standard normal distribution. In the first experiment, the mean value m of the normal distribution is estimated, while the standard deviation is known; and in the second experiment, the standard deviation σ is estimated, while the mean value is known.

In our experiments, the noninformative prior distribution of the mean value is a flat prior, which is a prior distribution that assigns equal likelihood to all possible values of the parameter $f(m) \sim 1$, while the prior distribution of an unknown standard deviation is a flat prior for the logarithm of the standard deviation $f(\sigma) \sim \frac{1}{\sigma}$.

We use two criteria to compare the estimated and true data distributions. The first criterion is Kullback–Leibler divergence (Kullback and Leibler, 1951). Kullback–Leibler divergence calculates a penalty when using one predicted distribution when, in fact, the data is distributed according to another one (so that the perfect Kullback–Leibler divergence value is zero). The second criterion is the integral of squared differences between the antecedent and the estimated cumulative distributions. We use this criterion because it coincides with a common kriging model diagnostic, the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007),

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} (F(y) - G(y))^2 \mathrm{d}y = \mathrm{E}_{g(x)} \left[\int_{-\infty}^{+\infty} \left(F(y) - \mathbb{1}_{x<y} + \mathbb{1}_{x<y} - G(y) \right)^2 \mathrm{d}y \right] = \\
&= \mathrm{E}_{g(x)} \left[\int_{-\infty}^{+\infty} \left(F(y) - \mathbb{1}_{x<y} \right)^2 \mathrm{d}y \right] + 2\mathrm{E}_{g(x)} \left[\int_{-\infty}^{+\infty} \left(F(y) - \mathbb{1}_{x<y} \right) \left(\mathbb{1}_{x<y} - G(y) \right) \mathrm{d}y \right] + \\
&\quad + \mathrm{E}_{g(x)} \left[\int_{-\infty}^{+\infty} \left(\mathbb{1}_{x<y} - G(y) \right)^2 \mathrm{d}y \right] = \\
&= \mathrm{E}_{g(x)} \left[\int_{-\infty}^{+\infty} \left(F(y) - \mathbb{1}_{x<y} \right)^2 \mathrm{d}y \right] - \int_{-\infty}^{+\infty} G(y) \times (1 - G(y))\, \mathrm{d}y = \\
&= \overline{\mathrm{CRPS}}(F) - \int_{-\infty}^{+\infty} G(y) \times (1 - G(y))\, \mathrm{d}y = \overline{\mathrm{CRPS}}(F) - \mathit{constant}, \qquad (2)
\end{aligned}
$$

where $F(y)$ is the estimated cumulative distribution function, $G(y)$ is the true cumulative distribution function, $g(x)$ is the probability density function of G, $\overline{\mathrm{CRPS}}$ is the average CRPS, and *constant* is some nonnegative value.

The following numerical algorithm was used to compare the four methods of model fitting:

1. Generate random samples from the standard normal distribution.
2. Estimate model parameters using four approaches.
3. Reconstruct the original distribution using estimated model parameters.
4. Compare reconstructed distribution with the standard normal distribution by two criteria.
5. Repeat steps 1 to 4 to obtain an accurate estimation of two criteria. In our experiment, the number of repetitions ranged from 1,000,000 to 4,000,000.

The expected value for the average of the Kullback–Leibler criterion can be found analytically for several cases as shown in the second columns of Tables 1 and 2. The calculations were made

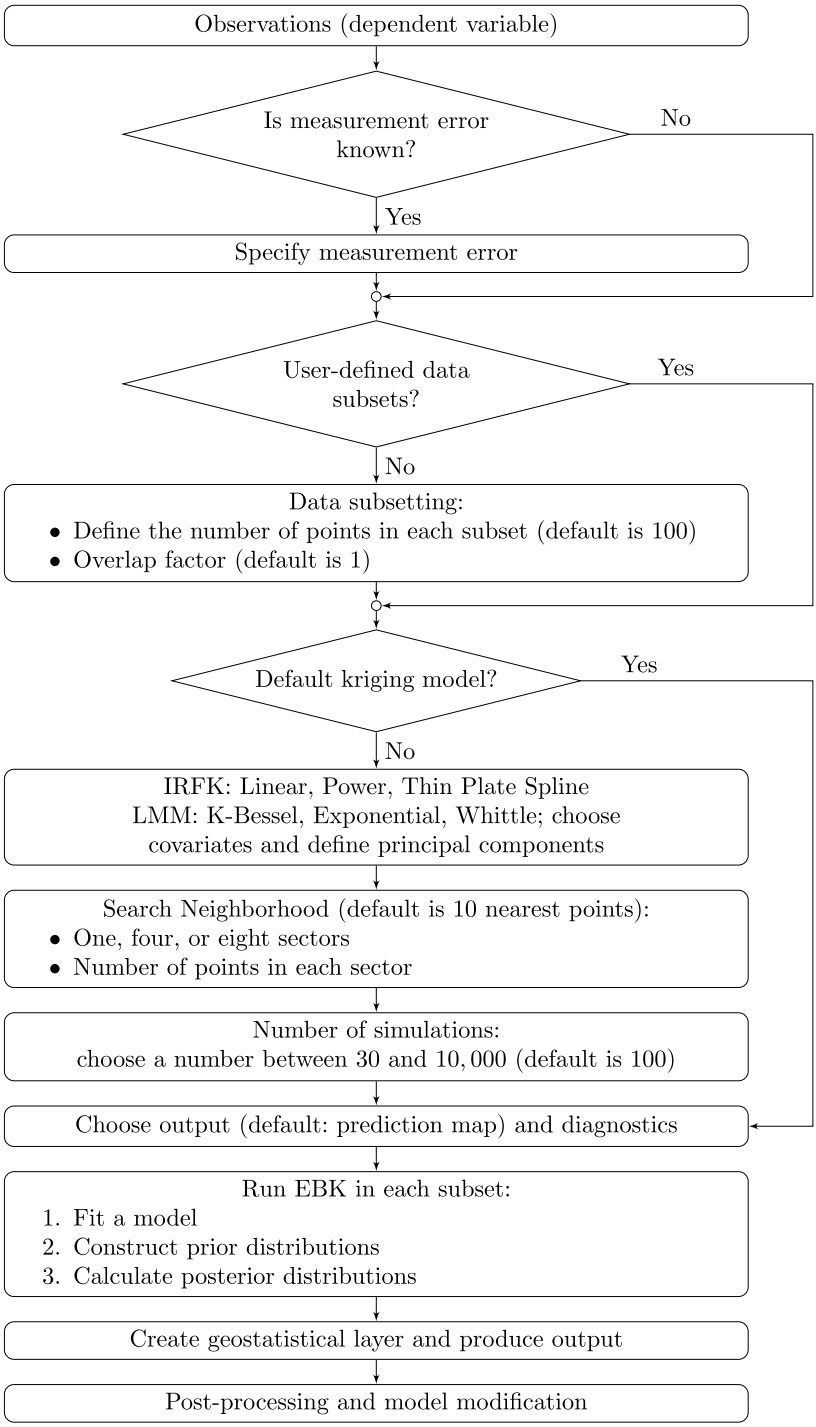

Fig. 1. EBK usage.

Table 1
Average Kullback–Leibler divergence and average of the integral of squared differences for six samples from the standard normal distribution when the mean value is estimated by four methods.

Method	Average Kullback–Leibler divergence formula for n samples	Average Kullback–Leibler divergence for 6 samples	Average of integral of squared differences for 6 samples
Empirical Bayesian	$\frac{1+(1+2n)\log\left(1+\frac{1}{2n}\right)}{2+4n}$	0.0785	0.0454
Empirical posterior distribution	$\frac{1}{2}\log\left(1+\frac{1}{n}\right)$	0.0771	0.0452
Posterior distribution obtained from noninformative prior distribution	$\frac{1}{2}\log\left(1+\frac{1}{n}\right)$	0.0771	0.0452
Maximum likelihood	$\frac{1}{2n}$	0.0833	0.0461

Table 2
Average Kullback–Leibler divergence and average of the integral of squared differences for six samples from the standard normal distribution when the standard deviation value is estimated by four methods.

Method	Average Kullback–Leibler divergence formula for n samples	Average Kullback–Leibler divergence for 6 samples	Average of integral of squared differences for 6 samples
Empirical Bayesian	Not available	0.222	0.0119
Empirical posterior distribution	$2\frac{n-1}{(n-2)^2}+\log\left(\frac{2}{n}\right)+\Psi\left(\frac{n}{2}\right)$	0.449	0.0125
Posterior distribution obtained from noninformative prior distribution	$\frac{1}{n-2}$	0.250	0.0129
Maximum likelihood	$\frac{\frac{2}{n-2}+\log\left(\frac{2}{n}\right)+\Psi\left(\frac{n}{2}\right)}{2}$	0.162	0.0121

Where $\Psi(x)=(\log(\Gamma(x)))'$ is the digamma function.

using analytical formulas (when available), Mathematica (Wolfram Research, Inc., 2016), and Monte Carlo simulations (in other cases). Mathematica and Monte Carlo simulations require about 4 and 12 hours, respectively, for each calculated value on two Intel Xeon CPU E5-2670 processors.

For n simulated samples, let $m^\star$ be their average value. The empirical prior distribution is generated as an average value for n random normal variables with mean $m^\star$ and a variance of one. Fig. 2 shows the posterior distributions assuming noninformative and empirical prior distributions for $n = 6$. Note that in Fig. 2, the empirical posterior distribution is the same as the posterior distribution when a noninformative prior distribution is used. Our experiment shows that the posterior distribution given an empirical prior distribution is narrower than the posterior distribution assuming a noninformative prior distribution. This effect can be stronger in higher dimensions.

It is well known that Bayesian calculations with a noninformative prior distribution produce correct prediction intervals for the unknown true mean value and, consequently, for the random sample from the standard normal distribution. In turn, the average estimated quantile of true p quantile by EBK and maximum likelihood methods are equal to $\Phi\left(\sqrt{\frac{2n+1}{2n+2}}\Phi^{-1}(p)\right)$ and

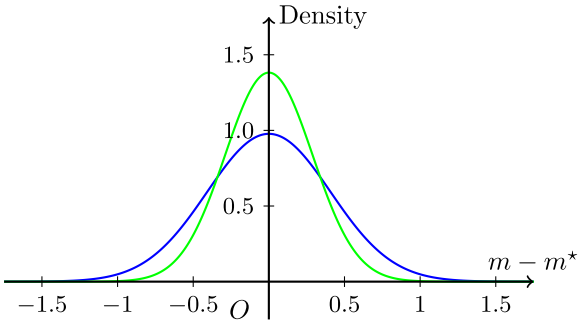

Fig. 2. Posterior mean value distribution estimated by EBK (green) and noninformative prior distribution (blue) methods for six samples from a standard normal distribution. Both distributions are shifted by the average of the simulated values.

$\Phi\left(\sqrt{\frac{n}{n+1}}\Phi^{-1}(p)\right)$ respectively; where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \cdot dy$ is the cumulative standard normal distribution and $\Phi^{-1}(x)$ is the inverse of the cumulative standard normal distribution.

Fig. 3 compares average estimated quantiles for the three approaches with true quantiles. The largest bias is observed for the maximum likelihood method because it does not include the uncertainty of the estimated model parameters (in this case, the mean value). In Fig. 3, the line for the EBK method lies between the maximum likelihood and the true quantile lines. Since the relationship $\frac{n}{n+1} < \frac{2n+1}{2n+2}$ is satisfied for any positive n; the average estimated quantile obtained by the EBK method is always estimated more accurately than the maximum likelihood method estimation. When the number of samples is greater than 5, the difference between the true and the estimated value by EBK quantiles is less than one percent. It can be concluded that the EBK method slightly underestimates the uncertainty of the estimated mean value when the number of samples is very small. However, in kriging applications, the effective number of data points for mean value estimation is usually sufficiently large, and small bias in the quantiles of the mean value distribution is not a problem in practice.

Figs. 4 and 5 show a comparison of the average Kullback–Leibler divergences and the average of the integral of squared differences between the estimated and true distributions when the mean value was estimated by all four methods. In both cases, the maximum likelihood method produces the largest bias, while the EBK method is slightly less accurate than other methods based on a noninformative prior distribution. The difference between Kullback–Leibler divergence and the average of the integral of squared differences is due to the additional penalty for the underestimated tails of the distribution. Note that the average of the integral of squared differences coincides with CRPS diagnostics, and it seems that CRPS does not pay much attention to unreliable estimations.

Table 1 shows analytical formulas for the Kullback–Leibler divergence and numerical values for two criteria for the case of six samples. Because inequality

$$\frac{1}{2}\log\left(1+\frac{1}{n}\right) < \frac{1+(1+2n)\log\left(1+\frac{1}{2n}\right)}{2+4n} < \frac{1}{2n}$$

is true for any positive n, the ranking according to the average Kullback–Leibler divergences remains the same. According to our numerical calculations, this is also true for the integral of squared differences. We conclude that in this simulation experiment, the EBK method performance is only slightly worse than empirical posterior and noninformative prior distributions.

We turn to the comparison of the standard deviation estimation by four methods.

Figs. 6, 7, and 8 show posterior distributions of the standard deviation value divided by the maximum likelihood estimation of the noninformative prior distribution, EBK method, and empirical posterior distribution. The normalization of standard deviation makes posterior distributions independent from simulated values, allowing the comparison of different methods. In the case of

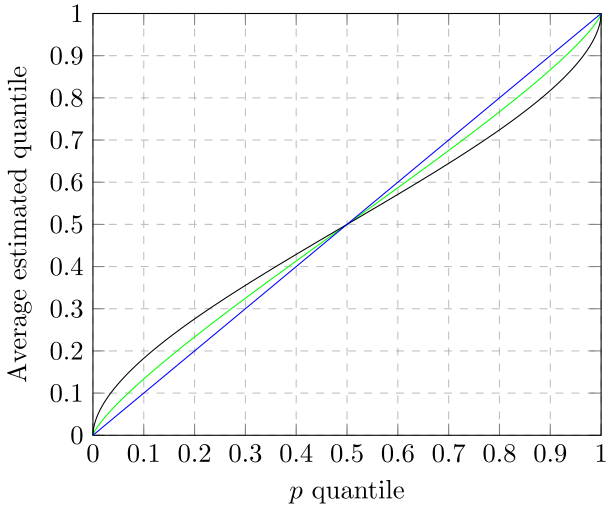

Fig. 3. Average estimated quantile versus true quantile when the mean value was estimated from one sample ($n = 1$) by the EBK method (green), maximum likelihood method (black), and empirical posterior distribution and posterior distribution assuming noninformative prior distribution (blue).

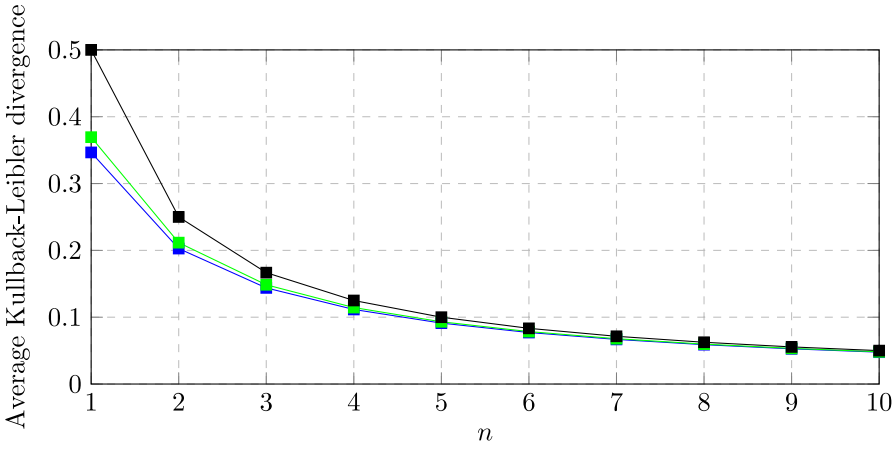

Fig. 4. Average Kullback–Leibler divergence for a varying number of samples from a standard normal distribution when the mean value is estimated by an empirical posterior distribution and a posterior distribution assuming a noninformative prior distribution (blue), empirical Bayesian method (green), and maximum likelihood method (black).

one sample, a posterior distribution assuming a noninformative prior distribution does not have a finite mean and variance. It also does not have a finite variance when the number of samples is two. This is an undesirable feature for the estimation of the data variance in kriging. Most likely, the lighter tail estimated by EBK is more realistic. An empirical posterior distribution gives a nonzero density for the standard deviation of zero when the number of samples is one. This is a serious problem, because a standard deviation of zero is not possible when a nonzero sample is generated from a standard normal distribution. The probability density close to zero standard deviation is still large when the number of samples is increased. However, any method that gives a larger uncertainty than a noninformative prior one should be questioned.

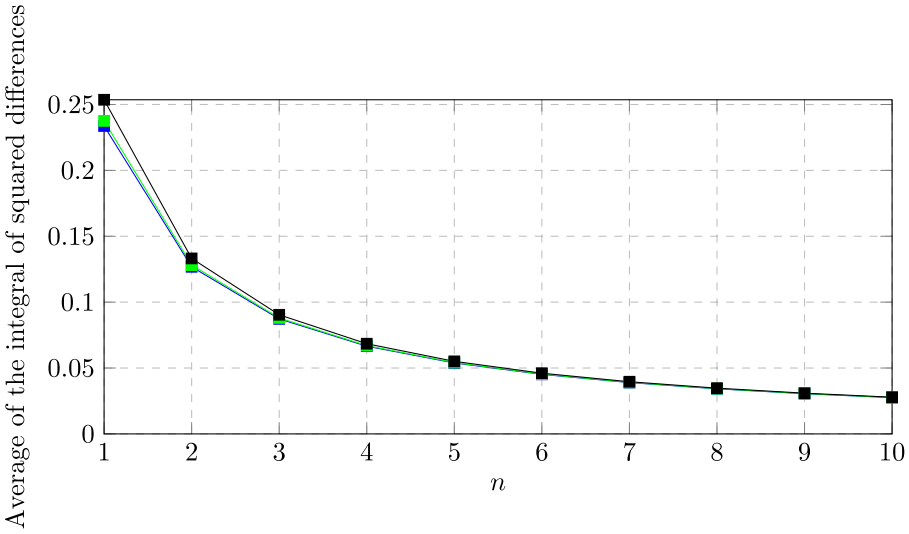

Fig. 5. Average of the integral of squared differences for a varying number of samples from a standard normal distribution when the mean value is estimated by an empirical posterior distribution and a posterior distribution assuming a noninformative prior distribution (blue), empirical Bayesian method (green), and maximum likelihood method (black).

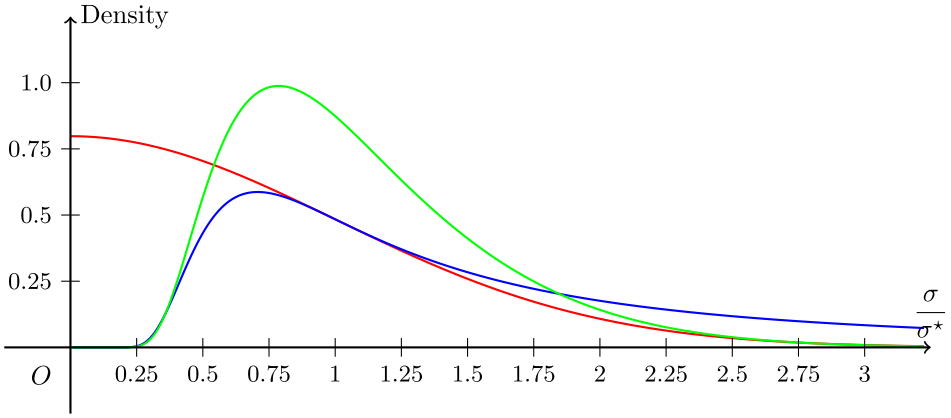

Fig. 6. Posterior distribution of the standard deviation value for one sample from a standard normal distribution. The blue line is the posterior distribution assuming a noninformative prior distribution, the green line is the EBK method, and the red line is the empirical posterior distribution. The standard deviation value is divided by the maximum likelihood estimation of the standard deviation value.

Figs. 9 and 10 show a comparison of the average Kullback–Leibler divergences and the average of the integral of squared differences between estimated and true distributions when the standard deviation value is estimated by all four methods. The comparison based on a Kullback–Leibler divergence indicates serious problems with an empirical posterior method, while EBK and noninformative prior distribution show very similar performance. According to the average of the integral of squared differences criterion, a noninformative prior distribution performed poorly when the number of samples was small. This is because the density assigned for large standard deviation values is too large.

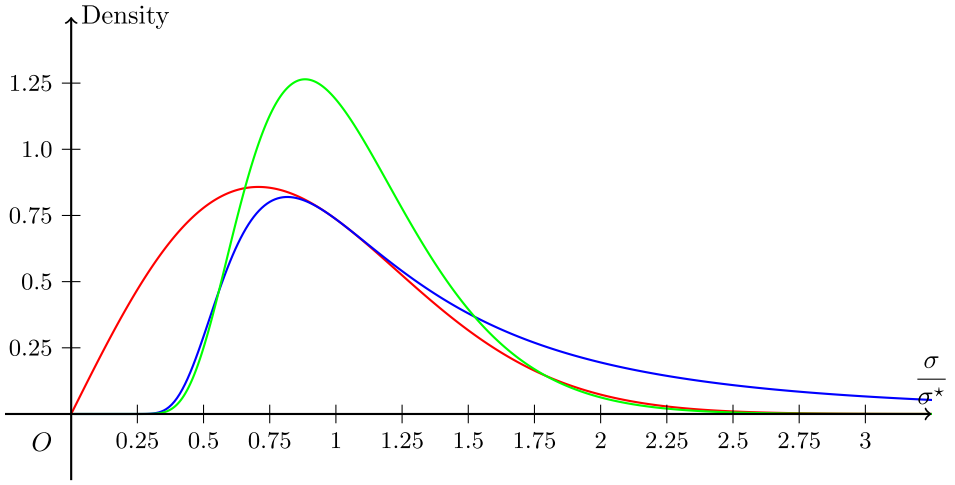

Fig. 7. Posterior distribution of the standard deviation value for two samples from a standard normal distribution. The blue line is the posterior distribution assuming a noninformative prior distribution, the green line is the EBK method, and the red line is the empirical posterior distribution. The standard deviation value is divided by the maximum likelihood estimation of the standard deviation value.

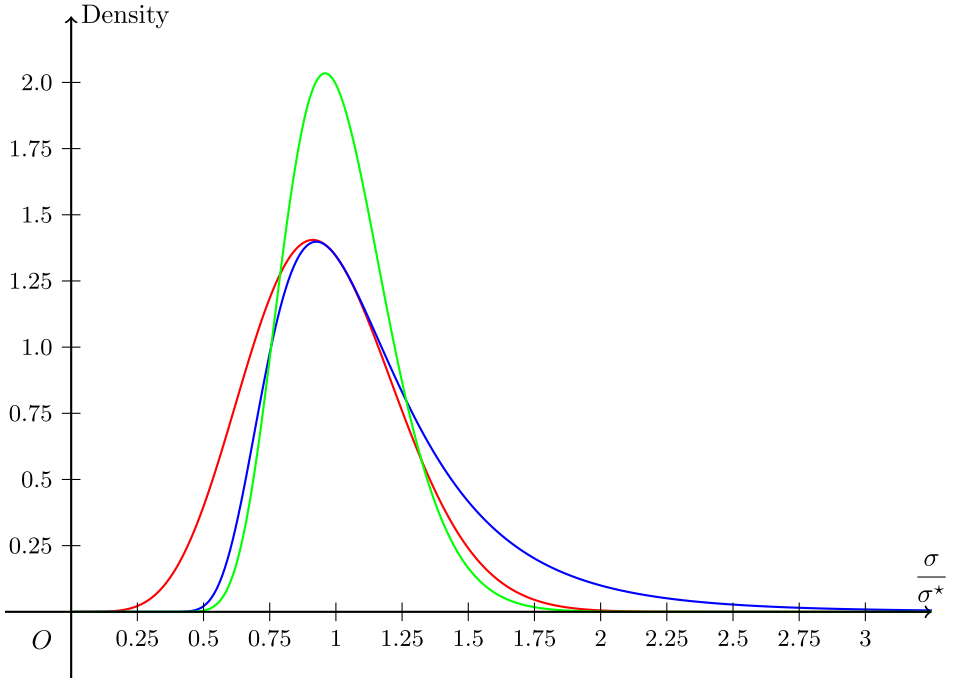

Fig. 8. Posterior distribution of the standard deviation value for six samples from a standard normal distribution. The blue line is the posterior distribution assuming a noninformative prior distribution, the green line is the EBK method, and the red line is the empirical posterior distribution. The standard deviation value is divided by the maximum likelihood estimation of the standard deviation value.

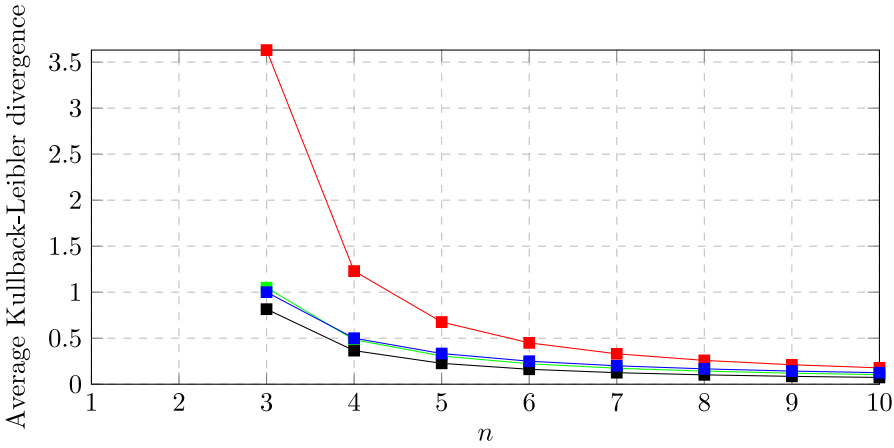

Fig. 9. Average Kullback–Leibler divergence for a varying number of samples from a standard normal distribution when the standard deviation value is estimated by the empirical Bayesian method (green), maximum likelihood method (black), empirical posterior distribution (red), and posterior distribution assuming noninformative prior distribution (blue).

Table 2 shows analytical formulas for the Kullback–Leibler divergence with the exception of the EBK method and numerical values for the two criteria for the case of six samples. We found that, numerically, the ranking according to the average Kullback–Leibler divergence is the same for the number of samples between 4 and 100 for both criteria. The Kullback–Leibler divergence is several times larger for the empirical posterior distribution than for the three other methods when the number of samples is small. This is because the empirical posterior distribution assigns a probability for a standard deviation near zero that is too large.

The problem found with the empirical posterior method can lead to a problem with semivariogram model estimation in kriging. For instance, misspecification of the estimated measurement error (nugget effect parameter) usually leads to unstable predictions when several measurements are collected near each other because the reliable estimation of the spatial variation at zero distance between points in kriging depends on the number of observations separated by very short distances. To illustrate this problem, we use simulated data with two coincident measurements at the location with coordinates $x = 0$ and $y = 0$, shown in Table 3. The model with a zero nugget parameter cannot describe this data properly. However, when simulations are generated from an empirical prior distribution, there are many models with zero or a very small nugget parameter value, as shown in Figs. 11 and 12 (left). Bayesian updating fixes this problem. Figs. 11 and 12 (right) show a dramatic change in the posterior distribution of the nugget parameter. Fig. 13 shows an even greater difference between the empirical posterior and the EBK methods: before Bayesian updating, the estimated model corresponds to rather independently distributed values, while after Bayesian updating, the EBK model describes strongly dependent data.

Based on the results of the simulation experiments presented in this section, we conclude that EBK typically produces better results than EBB and the method based on the noninformative prior distribution when the data values or the data locations are not ideal for kriging predictions. EBK works as well as other methods used in the comparison when the data is nearly perfect.

3. Four EBK models' comparison using simulated 2D Gaussian data

In this section, we compare EBK, EBB, and EML methods introduced in Section 1.

We used simulated data that is very similar to that described in section 4.1 of Zammit-Mangion and Cressie (2017). The data was simulated within the unit square on the grid 1000 by 1000, assuming a Gaussian process with an exponential model with an effective range of 0.15 or 0.9; signal-to-noise ratios (SNR) of 0.2, 1, 5, and 100, with or without a deterministic linear trend in

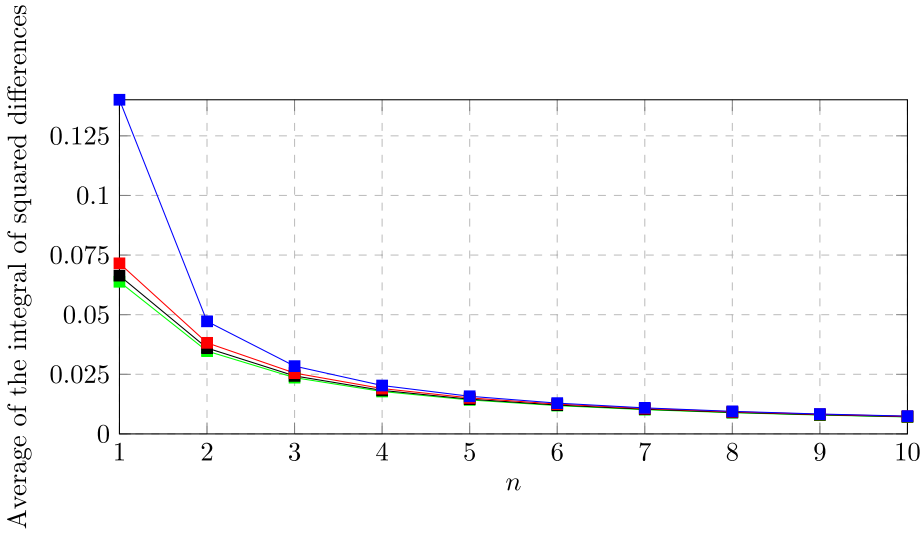

Fig. 10. Average of the integral of squared differences for a varying number of samples from a standard normal distribution when the standard deviation value is estimated by the empirical Bayesian method (green), maximum likelihood method (black), empirical posterior distribution (red), and posterior distribution assuming noninformative prior distribution (blue).

Table 3
Data generated from IRFK with a linear model having a nugget parameter of 1 and slope of 2.

#	X	Y	Z	#	X	Y	Z	#	X	Y	Z
1	0	0	4.15223	10	3	1	4.47932	19	2	3	0.61548
2	0	0	3.03133	11	4	1	6.58857	20	3	3	4.39039
3	1	0	−0.46877	12	0	2	1.25062	21	4	3	3.32127
4	2	0	6.59000	13	1	2	4.02829	22	0	4	−4.51207
5	3	0	11.10625	14	2	2	1.71498	23	1	4	−3.92381
6	4	0	7.30192	15	3	2	3.05332	24	2	4	1.14864
7	0	1	−5.98125	16	4	2	6.37620	25	3	4	−0.40270
8	1	1	2.08528	17	0	3	−3.07545	26	4	4	0.06026
9	2	1	3.59865	18	1	3	0.00798				

a horizontal direction. The number of simulated data points for fitting the models were 1000 or 100,000, 95 percent of which were randomly sampled in the left half (LH) of the square and 5 percent in the right half (RH). The validation dataset consists of randomly chosen simulated data values, which do not coincide with input data locations (10,000 points for both left- and right-hand sides). We simulated 100 datasets for each combination of factors in the simulation experiment. The locations of the simulated data points are shown in Fig. 14. The simulated data are available from the authors upon request.

In the simulation experiment, we used three (EBK, EBB, and EML) models with the same parameters:

- A power semivariogram model
- Nonoverlapping data subsets with approximately 100 data points in each one
- 100 unconditional simulations
- A search neighborhood with a maximum of 64 nearest neighbors

The only difference between this model and the default EBK model is the number of neighbors used for prediction (the default model uses up to 15 neighbors, which is not sufficient for predicting very noisy data with SNR $= 0.2$). Note that the power semivariogram model is normally used for

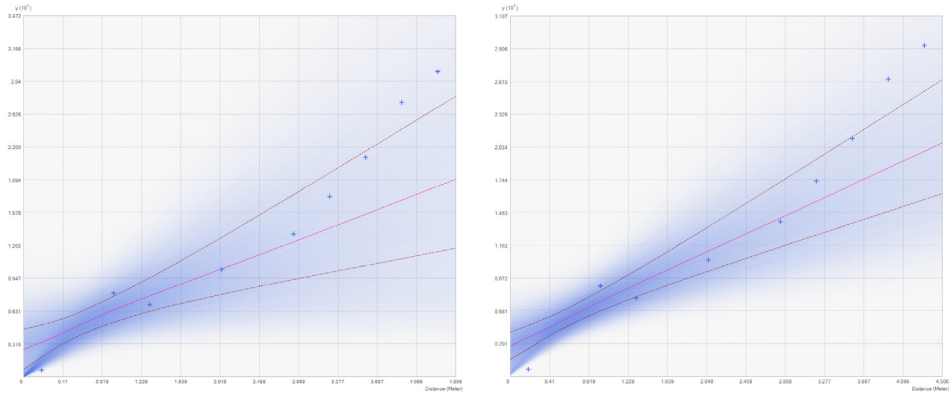

Fig. 11. Semivariogram model estimation comparison for the empirical prior distribution (left) and EBK posterior distribution (right) using data from Table 3.

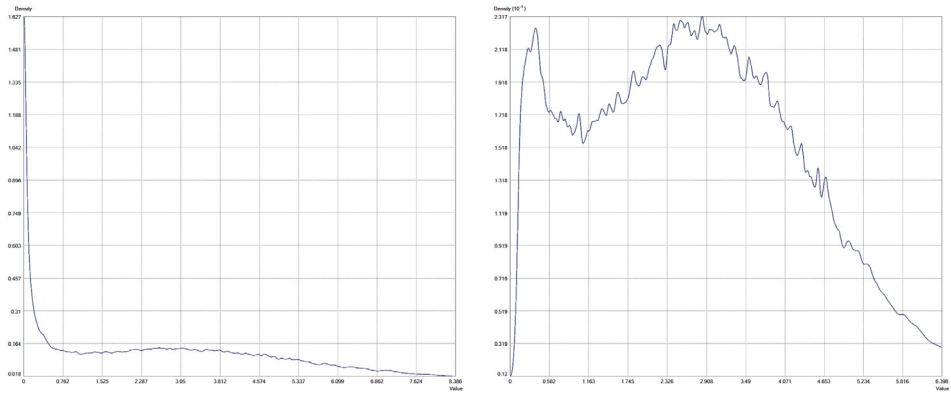

Fig. 12. Nugget parameter estimation comparison for the empirical prior distribution (left) and EBK posterior distribution (right) using data from Table 3.

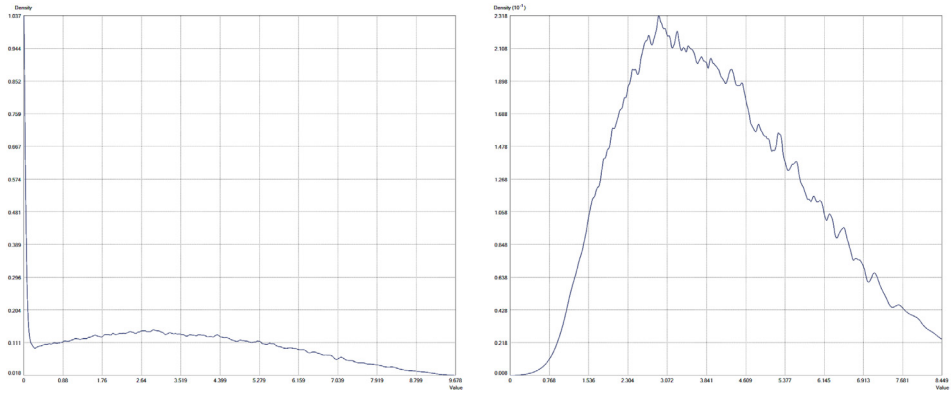

Fig. 13. Linear semivariogram slope estimation comparison for the empirical prior distribution (left) and EBK posterior distribution (right) using data from Table 3.

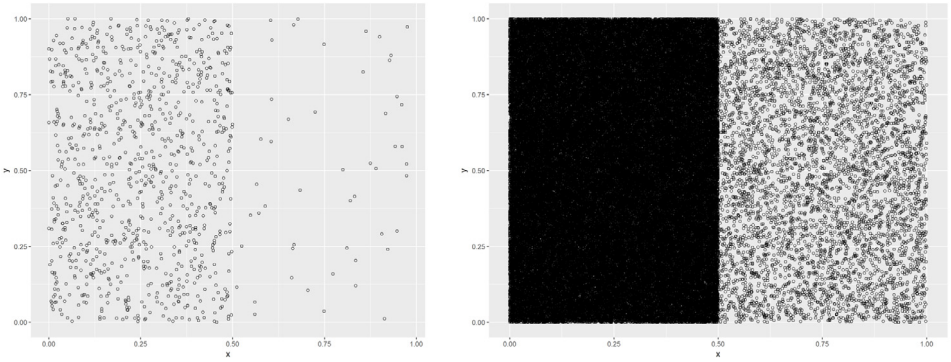

Fig. 14. The locations of the simulated data points used for fitting the models. Left: 1000 points. Right: 100,000 points. 95 percent of points fall in the left half of the square, and the remaining 5 percent in the right half of the square.

the data with a locally varying mean value. Such data can be generated by the spatially varying fractional diffusion process, which is different from the stationary Gaussian process used to simulate data for this computer experiment. Therefore, we additionally used another EBK model that better describes the simulated process: a model with K-Bessel covariance, x- and y-coordinates as the covariates, data subsets with 1000 observations, 500 unconditional simulations, and the same search neighborhood as in the EBK model with a power semivariogram. We used the abbreviation EBK*lm* for this model.

In addition to the four models based on EBK method, we used FRK model and simple kriging (SK) with true covariance parameters as a reference model. For the FRK, we used the same fine-tuned model parameters as those described in Zammit-Mangion and Cressie (2017) except that we used x- and y- coordinates as covariates. The covariates should improve predictions when the deterministic trend is added to the data. The FRK model is conceptually similar to the EBK*lm* model because both models are variants of the linear mixed model for large datasets.

All EBK models and SK use up to 64 nearest neighbors for prediction, while FRK uses all the data. The maximum number of nearest neighbors for prediction in all Geostatistical Analyst models is limited because, in the case of real data interpolation, the greater the number of nearest observations that are used for prediction, the more the prediction is corrupted due to the data and model uncertainties, see Gandin (1963, chapter 3, section 3.4). However, in the case of simulated stationary Gaussian data with fixed measurement error, the greater the number of neighbors, the more accurate the prediction. Therefore, the SK diagnostic is not ideal but only close to it. We expected that SK should have the best performance results, followed by FRK and EBK*lm*.

In Figs. 15 and 16, to compare six kriging models, we used CRPS prediction diagnostics (see Section 2) and the number of validation data values inside the 90 percent prediction interval (PI90). Ideally, the CRPS values should be as small as possible, and the PI90 values should be close to 90 percent. Each figure shows graphs for four SNR values at the unobserved locations, and varying factors of the data-generating processes (size of the input dataset, range of data correlation, and presence of trend). The beanplots show the actual density curve. Beanplots are more informative than the mean values or boxplots, which are most useful for normal and unimodal data.

Fig. 15 shows a diagnostic comparison of small simulation datasets at the right-side unobserved locations (areas with the lowest density of the observed values). All five fitted models assume a trend in the data, and their performance is better when the trend is present (graphs on the right). We see that there is a small difference in performance between EBK and EBB models, while EBK*lm* is usually slightly better than EBK. The EML model is consistently less accurate than all other models because, typically, EML cannot accurately reconstruct the generating process, and the misfit to the true model varies greatly. Therefore, some fine-tuning to this model is required. We see that the empirical Bayesian method works particularly well in this simulation experiment.

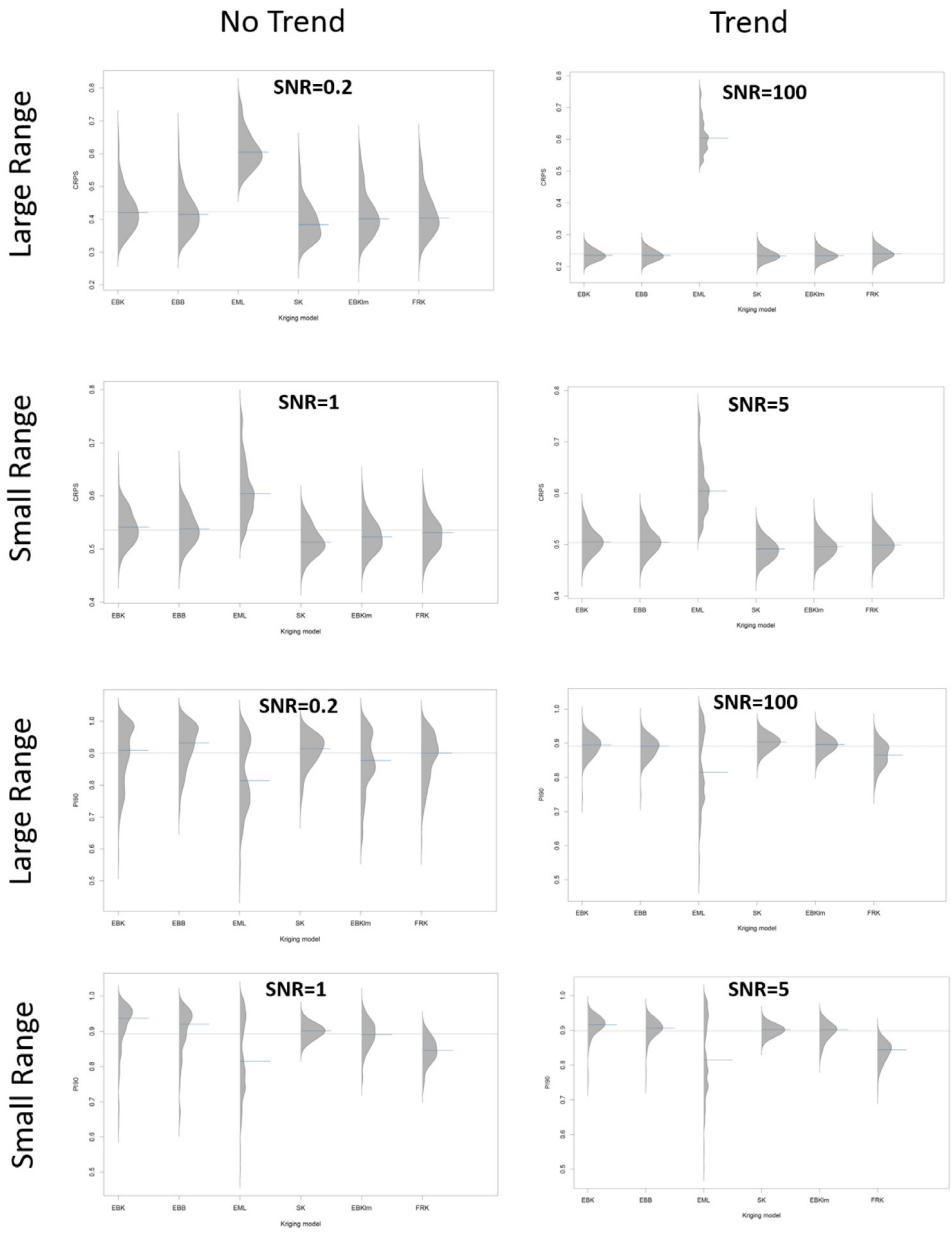

Fig. 15. Diagnostic comparison for unobserved locations on the right side of the data extent (small datasets). Left: No trend. Right: Linear trend in a horizontal direction. Top four graphs: CRPS diagnostic. Bottom four graphs: PI90 diagnostic.

Fig. 16 shows a diagnostic comparison for large simulation datasets at the left side of the square with simulated data (areas with the highest density of observed values). Similar to predictions with small simulated datasets, the performance of all models is better for simulated datasets with added trend. For kriging models with a power semivariogram, simulated data with a smaller range of data correlation are predicted more accurately than simulated data with a larger range. This is because

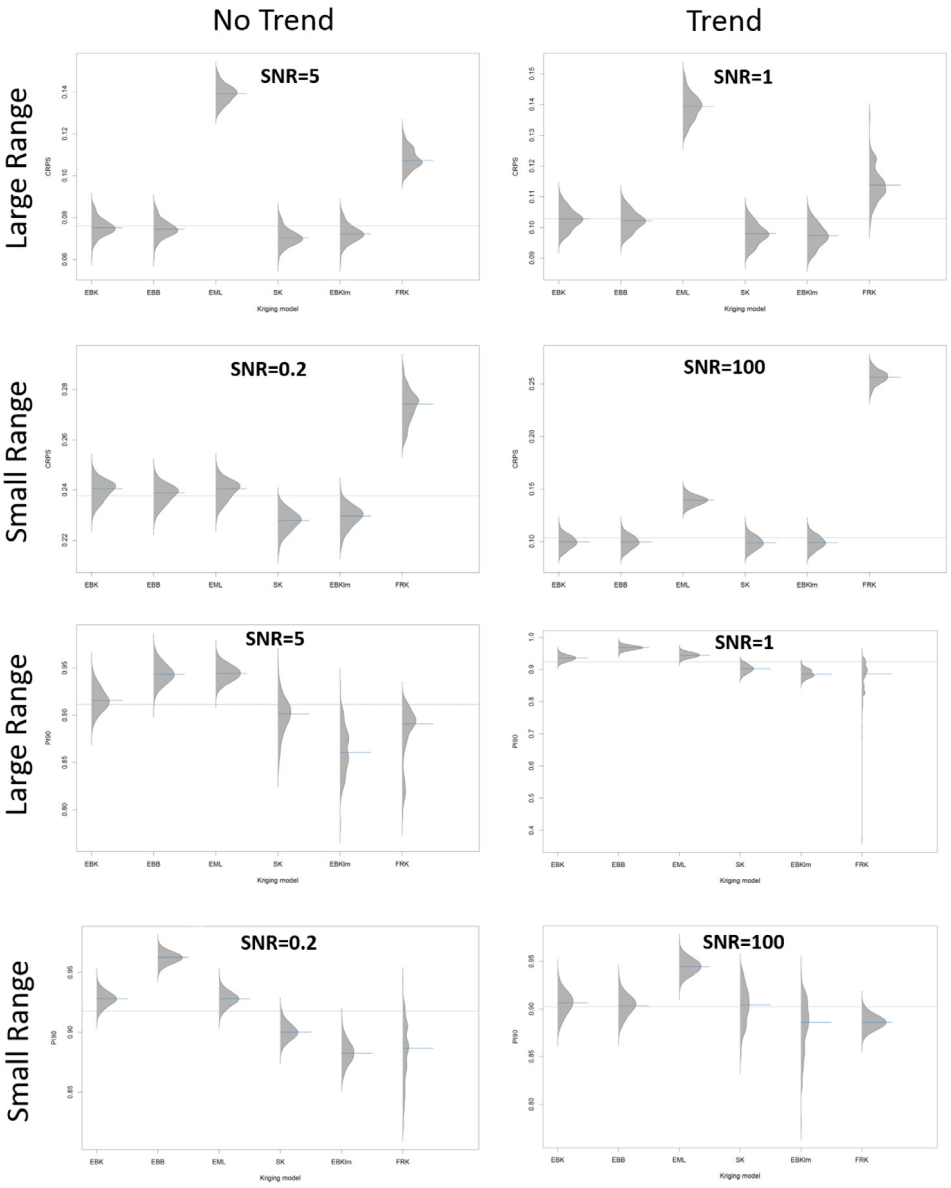

Fig. 16. Diagnostic comparison for unobserved locations on the left side of the data extent (large datasets). Left: No trend. Right: Linear trend in a horizontal direction. Top four graphs: CRPS diagnostic. Bottom four graphs: PI90 diagnostic.

true semivariogram model reconstruction requires more data in both the data subsets and the search neighborhood. Additionally, prediction performance of the EBK and EBB models is improved by increasing the SNR value, while the EBK*lm* model performance is good in all situations. We also see that the EBK and EBB models overestimate prediction uncertainty, while EBK*lm* underestimates it. This is primarily because a relatively large number of the fitted semivariagram models Θ_i approximate a true exponential covariance model less accurately than the spectrum of the K-Bessel covariance models, which include the exponential model as a particular case. Finally, prediction

intervals produced by the EBB model are shifted more from the desired 90 percent value than prediction intervals produced by the EBK model.

Summarizing the results of the simulation experiments presented in this section, we conclude that EBK performance is usually more accurate than the model with empirical posterior distribution even when the data is nearly perfect, in particular, when observations do not include coincidental data. The difference between the EBK and EBB models is usually in the accuracy of estimating the 90 percent prediction interval. We have also observed that the EBK*lm* model with large data subsets performs almost perfectly in the ideal situation when the data values are Gaussian, stationary, and randomly sampled. Finally, we have seen that merging the semivariogram models estimated in the data subsets with maximum likelihood algorithm (EML model) is not a good idea even for perfect data.

4. Case study: a comparison of Bayesian, empirical Bayesian, and frequentist kriging models

In this section, we compare EBK with four spatial prediction methods recently developed for interpolation of medium-sized spatial datasets, where the measurement numbers are between several thousand and a few hundred thousand and the model parameters may vary in space. Specifically, we compare EBK with MPP, SPDE, LTK, FRK, and LPI.

Only models accessible with little time investment were chosen for comparison. As such, a further requirement was that selected models be functional without the need for complex parameterization. In particular, we used the default EBK model (EBKdef), although it is easy to improve the model based on the results of basic exploratory spatial data analysis. However, other models' performance could also be improved with substantial knowledge about those models' features.

The performance of each spatial interpolator was evaluated empirically using a randomly selected 20 percent of average annual rainfall values in South Africa (Lynch, 2004). There were 8397 measurements covering an area of roughly 1 million square kilometers, which consists of bushveld, grasslands, forests, deserts, mountains, and a coastline of more than 3000 kilometers (Lynch, 2004). The data is provided with (Krivoruchko, 2011).

For the model performance comparison, we used the following prediction diagnostics: maximum predicted value, root mean square prediction error (RMSPE), root mean square standardized prediction error (RMSSPE), and the number of validation data values inside the 90 percent prediction interval (PI90).

The RMSPE and RMSSPE statistics are computed using the following formulas:

$$\sqrt{\frac{\sum_{i=1}^{N}\left(z_i-\hat{Z}(s_i)\right)^2}{N}} \text{ and } \sqrt{\frac{\sum_{i=1}^{N}\left(\frac{z_i-\hat{Z}(s_i)}{\sigma(s_i)}\right)^2}{N}},$$

where z_i and $\hat{Z}(s_i)$ are observed and predicted values at locations s_i, $\sigma(s_i)$ is the prediction standard error, and N is the number of validation locations. Ideally, the RMSPE value should be as small as possible (unless the model is not overfitted). If the average of kriging standard errors $\sigma(s_i)$ is close to the RMSPE, then the variability in prediction is correctly assessed because two different estimates of the prediction error give similar results. Therefore, the RMSSPE value should be close to 1.

Fig. 17 shows 1679 validation locations where predictions are required after fitting a kriging model using 6718 training precipitation measurements. The predictions are required inside the South African border. The South African relief is complex, and accurate data interpolation is challenging even in the areas with dense observations such as the Cape Town vicinity shown at bottom right of Fig. 17. Note that an accurate prediction of large rainfall values is important because a profitable forestry and agriculture business in South Africa requires greater than 800 mm of rain annually.

Before presenting the validation results, we qualitatively compared dense predictions inside South Africa's border made by different kriging models using the entire dataset with 8397 measurements. Fig. 18 (at left) shows the density of data locations, and Fig. 18 (at right) shows the annual

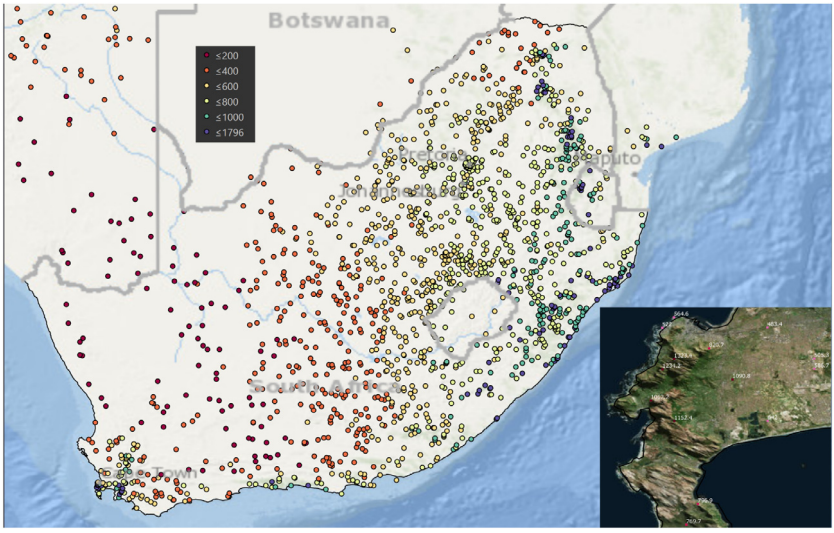

Fig. 17. 1679 validation locations with average annual precipitation measurements. Inset, bottom right: Monitoring locations near Cape Town. The predictions are required inside the South African border.

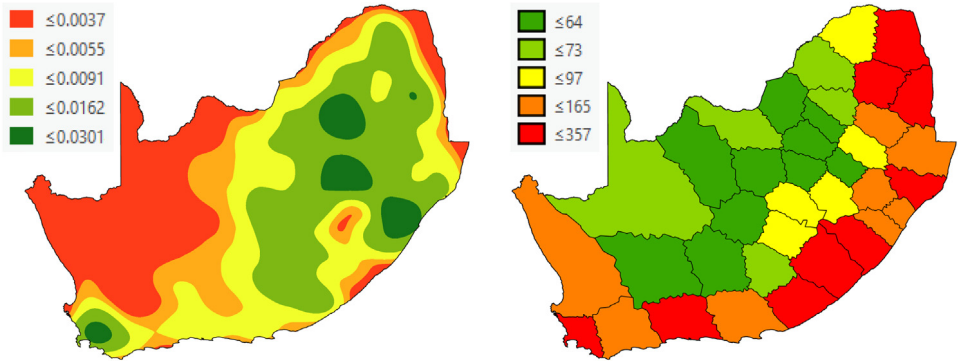

Fig. 18. Left: Density of data locations. Right: Annual rainfall standard deviation in areas with approximately 200 observations shows how the prediction standard error map should look like qualitatively.

rainfall standard deviation in the areas with approximately 200 observations. Note that standard deviation values in the red areas are about eight times larger than the values in the green polygons. We expect that the estimated prediction uncertainty should be larger in the areas with greater data variation. When the data variation is about the same, the prediction error should be larger in the areas with a lower density of observations.

Fig. 19 shows the average annual precipitation predictions using SPDE (top left), MPP (top center), FRK (top right), EBKdef (bottom left), LTK (bottom center), and LPI (bottom right) models. Predictions produced by the MPP model are very smooth; they largely show a trend in the data rather than the variable at the fine scale as in the case of the default EBK model, while predictions made by the other four models are somewhere in between predictions produced by the EBK and MPP models.

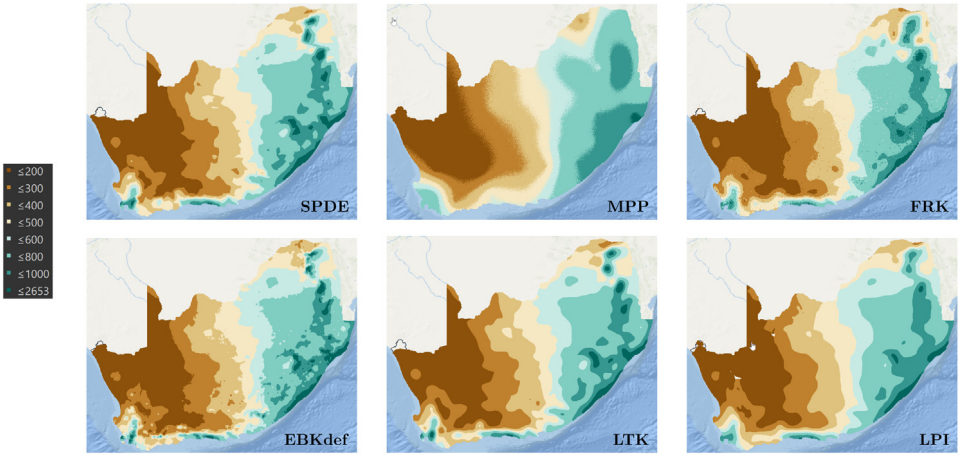

Fig. 19. Average annual precipitation predictions using SPDE (top left), MPP (top center), FRK (top right), EBKdef (bottom left), LTK (bottom center), and LPI (bottom right) models.

Fig. 20 shows the average annual precipitation prediction standard errors produced with SPDE (top left), MPP (top center), FRK (top right), EBKdef (bottom left), LTK (bottom center), and LPI (bottom right) models. We use different data intervals because the range of prediction standard errors is very different for different methods. The same data intervals would produce less informative maps. There is a big difference in the spatial structures of the prediction errors: SPDE, FRK, LTK, and LPI prediction uncertainty largely reflects the density of the data locations, MPP combines the density of the points with the distances to the knots' locations, while the EBKdef prediction standard errors pattern is qualitatively similar to the precipitation standard deviation map in Fig. 18 (right). Additionally, the EBK prediction error is a function of density of data locations. Note that the FRK model also incorporates information about observed values locations to the prediction standard error map, but in a different way: the predicted uncertainty values decrease near the observation locations. The prediction standard errors are classified into different ranges for each model. The smallest and largest prediction standard errors are produced by the EBKdef model, forming the largest range of the error values. The most uniform prediction standard errors are observed for MPP, with the lowest value being very large (larger than the largest prediction standard error produced by FRK) and a small difference between the lowest and highest values. The range of the prediction standard errors is also small for FRK.

In meteorological literature, precipitation values are typically modeled using gamma distribution (Husak et al., 2007), and the mean and variance of that distribution are related. Therefore, precipitation prediction standard error should be a function of the prediction. Accordingly, large prediction standard errors in the northwestern part of South Africa near the Kalahari Desert (with low annual rainfall) and small prediction standard errors in the tropical part of the country (with high annual rainfall) do not appear realistic.

To simplify the comparison of the prediction standard error maps with the data standard deviation map, Fig. 21 shows the average prediction standard errors produced by six models in the same areas as in Fig. 18 at right. We see that at a lower resolution, five maps look very similar, although their average prediction standard error intervals are clearly different: they are small for MPP and FRK and relatively large for SPDE, LTK, and LPI. The average prediction standard error map produced by EBKdef is very different, both qualitatively and quantitatively, from five other maps.

We think that prediction standard error maps production that mimics the observed values density is a serious problem because such maps ignore complex spatial data variation, which is a common feature of the data collected over large territories.

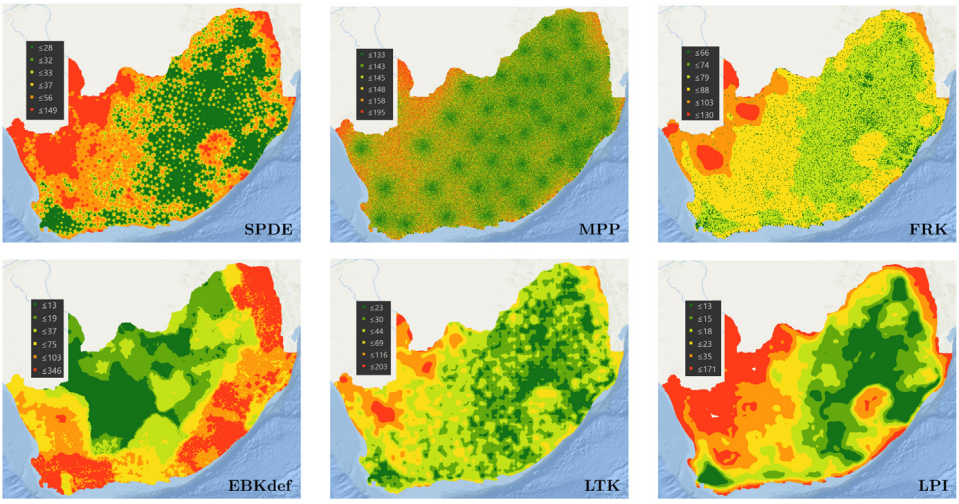

Fig. 20. Average annual precipitation prediction standard errors produced by SPDE (top left), MPP (top center), FRK (top right), EBKdef (bottom left), LTK (bottom center), and LPI (bottom right) models. Note that color intervals are different for each map.

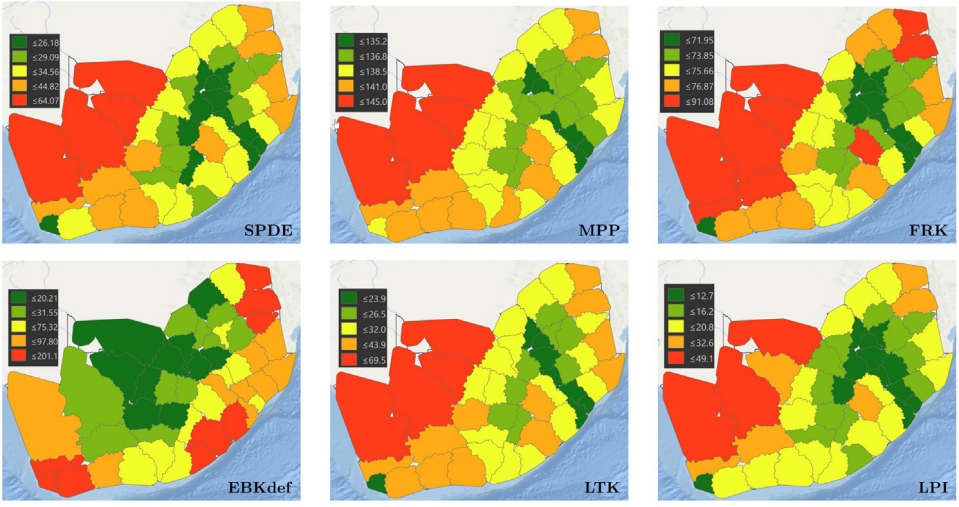

Fig. 21. Average prediction standard errors in 48 areas with the same number of observations produced by SPDE (top left), MPP (top center), FRK (top right), EBKdef (bottom left), LTK (bottom center), and LPI (bottom right) models. Note that color intervals are different for each map.

Table 4 shows the prediction validation diagnostics summary at 1679 locations whose values were not used in the model fitting. The maximum observed values in the training and validation datasets were 3199 and 1796 millimeters respectively. In Table 4, the abbreviations *EBB* and *EML* denote empirical Bayesian bootstrap and maximum likelihood models estimated in the same data subsets and with the same model parameters as the EBKdef model (see more about EBB and EML models in Section 1). The order of the models in Table 4 is based on the increasing RMSPE values. All results were produced on an Intel Core i7-4712HQ processor.

Table 4
Average annual precipitation data validation diagnostics.

Model	Max prediction	RMSPE	RMSSPE	Inside 90% prediction interval (%)	Model fitting and validation (seconds)
EBKdef	1487	83.66	1.04	89.9	7
EML	1489	83.67	1.86	72.3	4
EBB	1480	83.77	1.74	73.7	5
LTK	1354	94.13	3.40	62.7	48
SPDE	1254	97.77	3.58	62.4	240
LPI	1331	103.52	1.08	90.4	2
FRK	1236	103.91	1.35	85.5	194
MPP	1056	141.23	1.02	89.4	1196

The validation diagnostic comparison can be summarized as the following:

- EBK prediction standard errors correctly describe the precipitation data variability, while global kriging models with spatially varying covariance function just show the observed values density distribution.
- Predictions produced by local models (EBKdef, EBB, and EML) are more accurate in the areas with large values.
- The RMSPE values were found to be much lower for models estimated in the data subsets compared to other predictors.
- EBKdef prediction errors were consistently estimated, with the RMSSPE being close to one and the 90 percent prediction interval covered almost ideally. The only other models with consistently estimated prediction standard errors are LPI and MPP, but their prediction accuracies are among the least precise.
- The main difference between EBK, EBB, and EML models is in the prediction standard error accuracy: The EBB and EML methods heavily underestimate the prediction uncertainty.
- EBKdef calculations are much faster than LTK, SPDE, FRK, and MPP models.

Conclusions based on the validation exercise using just one dataset can be biased, and we conducted another simulation experiment with highly variable data as in the South African rainfall use case discussed in this section. The result of the models' comparison was qualitatively similar to the validation summary above, see Gribov and Krivoruchko (2019).

5. Satellite data prediction comparison

In this section, we demonstrate that EBK models do a good job of providing both self-tuned and hand-tuned outcomes in comparison with 12 other hand-tuned kriging model outcomes shown in the very recently published satellite data prediction comparison experiment (Heaton et al., 2018). The dataset for prediction comparison consisted of preprocessed observations of daytime land surface temperatures on the 500×300 grid, ranging from longitude values of approximately $-95.9°$ to $-91.3°$ and latitude values of $34.3°$ to $37.1°$, as measured by the Terra instrument on board the MODIS satellite on August 4, 2016. The missing data pattern on August 6, 2016, from the same MODIS satellite data was used to separate each dataset into training and test sets. After the split, the training set for the MODIS data consisted of 105,569 observations. The experiment designers instructed the competitors that the data can be reasonably modeled using the Gaussian process with just one exponential covariance model with a linear effect for latitude and longitude and a nugget variance.

Each method was compared in terms of mean absolute error (MAE), root-mean-squared error (RMSE), continuous rank probability score (CRPS; see Gneiting and Raftery (2007)), interval score (INT; see Gneiting and Raftery (2007)), and prediction interval coverage (CVG; the percent of containing the observed value in the estimated 95 percent intervals). To calculate the CRPS and INT, it was assumed that the associated predictive distribution was well approximated by a Gaussian distribution with the mean centered at the predicted value and standard deviation equal to the

Table 5
Numerical scoring for five competing methods on the satellite data.

Method	MAE	RMSE	CRPS	INT	CVG	Run time (min)	Cores used
NNGP	1.21	1.64	0.85	7.57	0.95	2.06	10
SPDE	1.10	1.53	0.83	8.85	0.97	120.33	2
EBK Default	1.22	1.80	0.91	9.51	0.94	0.33	32
EBK Exponential	1.10	1.49	0.75	7.54	0.94	19.58	32
EBK Linear	1.06	1.45	0.76	8.03	0.95	0.58	32

prediction standard error. However, the EBK predictive distribution is always more or less non-Gaussian, even in the case when simulated Gaussian data is used as input to the models. Therefore, we calculate the CRPS without assuming the prediction normality. For each prediction location, EBK calculates a distribution for the predicted signal. From that distribution, a large number of samples are generated, and the cumulative distribution is calculated using the simulated values. Then the CRPS is calculated as an integral of squared differences between the cumulative distribution and the indicators; see formula (20) in Gneiting and Raftery (2007).

It should be noted that Heaton et al. (2018) describes all 12 kriging models used in the competition and provides scripts with fine-tuning models used in the experiment. It is possible that outcomes from the LatK, SPDE, FRK, and MPP models in Section 4 can be improved using those scripts. However, in Section 4 we used the default EBK model for a fair comparison. The default model can be easily improved as shown in this and in the companion papers (Gribov and Krivoruchko, 2019).

Fitting a model to gridded, nearly stationary Gaussian data is an easy task for any reasonable geostatistical model when both training and testing datasets are chosen randomly. But here the predicted locations coincide with large clouds observed over the data extent, and the distance from the prediction location to the nearest observation can be as large as the range of the data correlation.

Three EBK models selected for the comparison run on the virtual machine hosted on a DELL R640 server with 64 GB of RAM and two Intel Xeon Gold 6142 2.60 GHz CPUs with 16 cores each—for a total of 32 threads. Table 5 shows the numerical scoring for the two best models in Heaton et al. (2018, table 3), namely, SPDE (Rue et al., 2009, 2017) and NNGP (Datta et al., 2016a,b), as well as scorings for three EBK models. First, we used the default EBK model. Its RMSE, CRPS, and CVG scoring would be, correspondingly, the fifth, the fourth, and the second best among 12 models that were compared in the competition experiment. Notice also that this model runs faster than any model used in Heaton et al. (2018). The default EBK model performs relatively poorly because it uses only 10 nearest observations when predicting to the sampled or unsampled locations. However, in the large areas without observations, the default EBK predictions are typically produced using observations located in a small area far from the prediction location. This may lead to inaccurate predictions. Therefore, for the hand-tuned models, based on our knowledge about EBK features, one should use the search neighborhood with the largest available number of sectors (eight), and the overlap factor should be increased. Our second model, LMM with exponential covariance and a linear effect for x- and y-coordinates, uses the following input parameters: subset size of 125 points, 500 simulations, overlap factor of 5, eight sectors search neighborhood with up to 8 observations in each sector, and empirical data transformation. Numerical scoring for that model would be the best among 12 competitors for the RMSE and CRPS diagnostics and the second best for the CVG. Finally, we used the EBK model with the smallest number of parameters, IRFK with the linear semivariogram model with the following parameters: subset size of 100 points, 30 simulations, overlap factor of 3, and eight sectors search neighborhood with up to 8 observations in each. The third EBK model has better scoring values than all competing models in Heaton et al. (2018) and, additionally, it runs faster than all models from the original prediction competition experiment. Numerical scorings for EBK models can be further improved by using more sophisticated data subsets created around unsampled areas under the clouds.

It should be noted that we used projected data instead of the original coordinates in latitude and longitude because EBK implementation has more options for projected data, including sectors in the search neighborhood, see also (Gribov and Krivoruchko, 2019; Krivoruchko and Gribov, 2019).

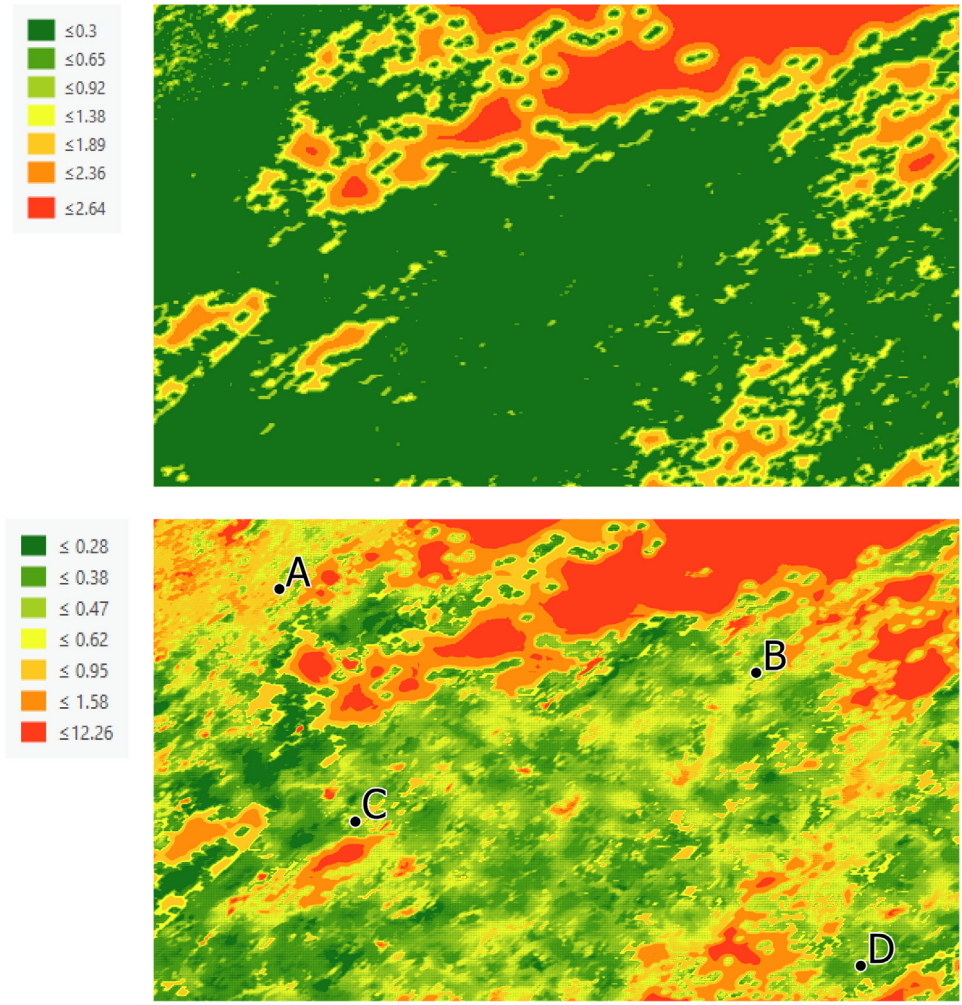

Fig. 22. Prediction standard error maps. Top: SPDE. Bottom: EBK Exponential.

The public GitHub repository (Lindgren, 2019) provides scripts used for 12 models fitting and the numerical scores calculations in Heaton et al. (2018) and we used one R script to reproduce the prediction and prediction standard error maps produced by SPDE, the winner of the competition conducted in Heaton et al. (2018). Fig. 22 shows two prediction standard error maps created by SPDE and EBK Exponential models for the entire satellite data extent and Fig. 24 shows histograms of the estimated prediction standard errors. One can see that SPDE assigns the same prediction standard error over the sampled areas, while EBK Exponential has complex structure in the areas with observed data. Inside the large unsampled areas, the EBK Exponential prediction standard error is increasing with increasing distance between the prediction location and the nearest observed location while SPDE produces the same prediction standard errors starting from the distance about the estimated range of the data correlation.

Large difference between two prediction standard error maps can be explained by the following considerations:

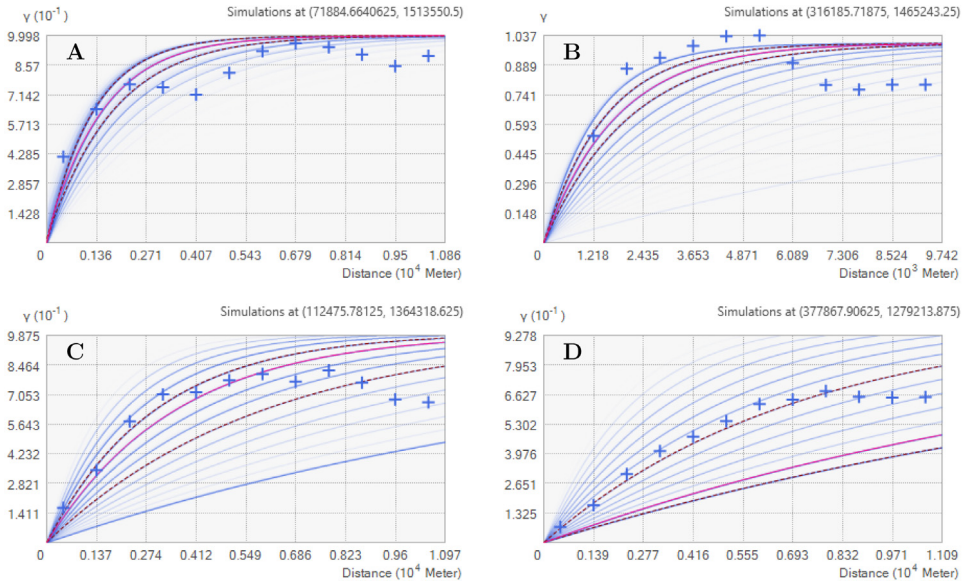

Fig. 23. EBK Exponential semivariogram spectra at four prediction locations indicated in Fig. 22 on bottom.

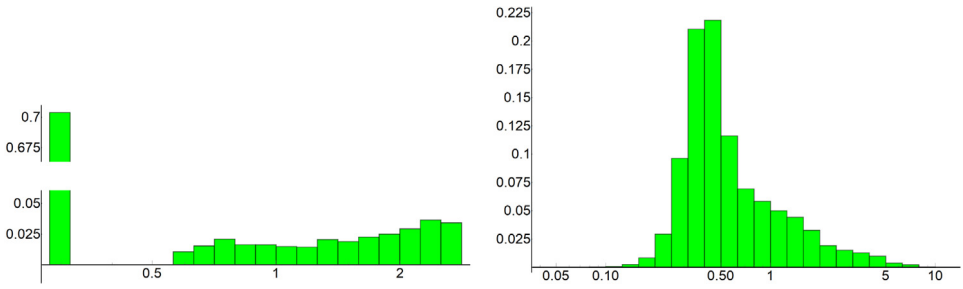

Fig. 24. Histograms of prediction standard error map values shown in Fig. 22. Left: SPDE. Right: EBK Exponential.

- EBK estimates the semivariogram locally. Fig. 23 shows exponential semivariogram models spectra at four locations shown in Fig. 22 on bottom with letters A, B, C, and D. We see rather high variability of the semivariogram models resulting in substantial variability of the prediction standard errors while SPDE uses practically the same exponential semivariogram parameters everywhere as shown in Table 6 produced using R script taken from (Lindgren, 2019).
- EBK Exponential uses non-linear data transformation, which introduces dependence between predictions and prediction standard errors (Gribov and Krivoruchko, 2012), (Krivoruchko, 2011, chapter 8), and (Cressie, 1993, section Trans-Gaussian Kriging).

It is possible that EBK overestimates the predicted temperature uncertainty, especially in the middle of the areas were observations are not available. On the other hand, it is unlikely that rapidly changing temperature over the territory with highly variable vegetation can be accurately described by one and the same simple model.

Table 6
Estimated exponential semivariogram parameters by SPDE.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
nugget.sd	0.2187114	0.0008017660	0.2168905	0.2188260	0.2199711	0.2191782
field.range	0.1559963	0.0003488976	0.1553658	0.1559717	0.1567454	0.1559000
field.sd	2.6112336	0.0048917002	2.6004106	2.6117397	2.6195795	2.6132662

6. Conclusion

We conducted two large simulation experiments and two case studies using highly variable observed precipitation data and nearly Gaussian satellite data with approximately linear trend to justify the EBK key assumption (the data-dependent prior distribution) and to investigate the EBK predicting quality. The data was simulated from standard normal distribution and from stationary spatial Gaussian processes.

In the first simulation experiment, we compared empirical Bayesian method with three other techniques that do not require specifying prior distribution. We found that our method works like other techniques used in the comparison when the data is nearly ideal. The empirical Bayesian method produces better results than the empirical posterior distribution method and the method based on the noninformative prior distribution when the data values or the data locations are not ideal for kriging predictions.

In the second experiment and in the case study, we compared prediction diagnostics for four EBK models, one benchmark model, and several models developed for moderately large data interpolation. Results of the models' comparison show that EBK outperforms other predictors, and increasingly so with data complexity.

The data used for the models comparison and instructions for their usage are provided in the supplementary materials.

EBK prior distribution can be made less data dependent by

1. Adding one more simulation from the fitted models Θ_i (step 4 of the EBK modeling algorithm in Section 1), then fitting models to the simulated data and using new fitted model parameters $\hat{\Theta}_i$ as the new prior distribution.
2. Adjusting the Bayesian update (step 5) to $w_i \propto \sqrt{f\left(\mathbf{z}|\hat{\Theta}_i\right)}$.

Our preliminary research shows that the LMM model with additional simulation and model fitting steps produces confidence intervals more closely centered to selected confidence level.

Acknowledgments

We thank two anonymous reviewers for critically reading the manuscript and suggesting substantial improvements. We also thank Noel Cressie and Andrew Zammit-Mangion, University of Wollongong, for advice on how to set up the comparison experiment of Section 3, and Steve Lynch, Esri, for help with automation of the EBK output in that simulation experiment.

Appendix. Data and scripts for the computer experiments reproduction

Supplementary materials related to this article can be found online at http://dx.doi.org/10.1016/j.spasta.2019.100368. We provide data and scripts to reproduce the simulation experiment and two case studies.

References

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2003. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC, p. 454. http://dx.doi.org/10.1201/9780203487808.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, Inc., p. 695.
Cressie, N.A.C., 1993. Statistics for Spatial Data, Revised Edition. John Wiley & Sons, Inc., p. 900.
Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 209–226. http://dx.doi.org/10.1111/j.1467-9868.2007.00633.x.
Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016a. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. J. Amer. Statist. Assoc. 111 (514), 800–812. http://dx.doi.org/10.1080/01621459.2015.1044091.
Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016b. On nearest-neighbor Gaussian process models for massive spatial data. Wiley Interdiscip. Rev. Comput. Stat. 8 (5), 162–171. http://dx.doi.org/10.1002/wics.1383.
Diggle, P.J., Ribeiro, Jr., P.J., 2007. Model-based Geostatistics. In: Springer Series in Statistics, Springer, New York, http://dx.doi.org/10.1007/978-0-387-48536-2.
Esri, 2019a. ArcGIS geostatistical analyst. Retrieved May, 2019. URL https://www.esri.com/en-us/arcgis/products/geostatistical-analyst/overview.
Esri, 2019b. Get started with geostatistical analyst in arcGIS pro. Retrieved May, 2019. URL https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/get-started-with-geostatistical-analyst-in-arcgis-pro.htm.
Esri, 2019c. What is EBK regression prediction? Retrieved May, 2019. URL https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/what-is-ebk-regression-prediction-.htm.
Esri, 2019d. How kernel interpolation with barriers works. Retrieved May, 2019. URL https://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/how-kernel-interpolation-with-barriers-works.htm.
Finley, A., Banerjee, S., 2017. SpBayes: Univariate and multivariate spatial-temporal modeling. Retrieved May, 2017. URL https://CRAN.R-project.org/package=spBayes.
Finley, A.O., Banerjee, S., Gelfand, A.E., 2015. SpBayes for large univariate and multivariate point-referenced spatio-temporal data models. J. Stat. Softw. 63 (13), 1–28, http://www.jstatsoft.org/v63/i13/.
Gandin, L., 1963. Objective Analysis of Meteorological Fields. Gidrometeorologicheskoe Izdatel'stvo (GIMIZ), Leningrad, Translated by Israel Program for Scientific Translations, Jerusalem, 1965.
Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437.
Gribov, A., Krivoruchko, K., 2011. Local polynomials for data detrending and interpolation in the presence of barriers. Stoch. Environ. Res. Risk Assess. 25 (8), 1057–1063. http://dx.doi.org/10.1007/s00477-011-0488-2.
Gribov, A., Krivoruchko, K., 2012. New flexible non-parametric data transformation for trans-Gaussian kriging. In: Abrahamsen, P., Hauge, R., Kolbjørnsen, O. (Eds.), Geostatistics Oslo 2012. In: Quantitative Geology and Geostatistics, vol. 17, Springer Netherlands, pp. 51–65. http://dx.doi.org/10.1007/978-94-007-4153-9_5.
Gribov, A., Krivoruchko, K., 2019. Empirical Bayesian kriging implementation and usage (submitted for publication).
Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D.W., Sun, F., Zammit-Mangion, A., 2018. A case study competition among methods for analyzing large spatial data. JABES 1–28. http://dx.doi.org/10.1007/s13253-018-00348-w.
Husak, G.J., Michaelsen, J., Funk, C., 2007. Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. Int. J. Climatol. 27 (7), 935–944. http://dx.doi.org/10.1002/joc.1441.
Johnston, K., Ver Hoef, J.M., Krivoruchko, K., Lucas, N., 2001. Using ArcGIS Geostatistical Analyst. Esri Press, p. 300.
Kitanidis, P.K., 1993. Generalized covariance functions in estimation. Math. Geol. 25 (5), 525–540. http://dx.doi.org/10.1007/BF00890244.
Kitanidis, P.K., 1997. Introduction to Geostatistics: Applications in Hydrogeology. In: Stanford-Cambridge Program, Cambridge University Press, p. 249.
Kitanidis, P.K., 1999. Generalized covariance functions associated with the Laplace equation and their use in interpolation and inverse problems. Water Resour. Res. 35 (5), 1361–1367. http://dx.doi.org/10.1029/1999WR900026.
Krivoruchko, K., 2011. Spatial Statistical Data Analysis for GIS Users. Esri Press, p. 928, https://community.esri.com/thread/201550-spatial-statistical-data-analysis-for-gis-users-available-free-for-download.
Krivoruchko, K., 2012. (a) Empirical Bayesian Kriging and (b) Modeling Contamination Using Empirical Bayesian Kriging. Fall 2012. ArcUser, URL https://www.esri.com/news/arcuser/1012/fall2012.html.
Krivoruchko, K., Frączek, W., 2015. Interpolation of Data Collected along Lines. Esri, URL https://apl.maps.arcgis.com/apps/MapJournal/index.html?appid=e7bd9a788b584f21ae738363b9b55d41.
Krivoruchko, K., Gribov, A., 2019. Distance metrics for data interpolation over large areas on Earth's surface. (submitted for publication).
Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86. http://dx.doi.org/10.1214/aoms/1177729694.
Lindgren, F., 2019. Code for Heaton et al (in JABES, 2018). Retrieved May, 2019. URL https://github.com/finnlindgren/heatoncomparison.
Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498. http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.
Lynch, S.D., 2004. Development of a Raster Database of Annual, Monthly and Daily Rainfall for Southern Africa. WRC Report 1156/1/04, Water Research Commission, University of KwaZulu-Natal, Pietermaritzburg, South Africa, p. 78, URL http://www.wrc.org.za/Knowledge%20Hub%20Documents/Research%20Reports/1156-1-04.pdf.
Nychka, D., Hammerling, D., Sain, S., Lenssen, N., 2016. LatticeKrig: Multiresolution Kriging Based on Markov Random Fields. R Package Version 6.4. University Corporation for Atmospheric Research, Boulder, CO, USA, http://dx.doi.org/10.5065/D6HD7T1R.

Nychka, D., Sain, D.H.S., Lenssen, N., Smirniotis, C., 2017. Latticekrig: Multiresolution kriging based on Markov random fields. Retrieved May, 2017. URL https://CRAN.R-project.org/package=LatticeKrig.

Pilz, J., Spöck, G., 2008. Why do we need and how should we implement Bayesian kriging methods. Stoch. Environ. Res. Risk Assess. 22 (5), 621–632. http://dx.doi.org/10.1007/s00477-007-0165-7.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x, with discussion.

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., Krainski, E.T., Fuglstad, G.-A., 2017. INLA: Bayesian Analysis of Latent Gaussian Models using Integrated Nested Laplace Approximations. R package, version 17.06.20.

Rue, H., et al., 2012. The R-INLA project. Retrieved May, 2017. URL http://www.r-inla.org/.

Schabenberger, O., Gotway, C.A., 2004. Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC Texts in Statistical Science, Chapman and Hall/CRC, p. 562, URL http://www.crcpress.com/product/isbn/9781584883227.

Wolfram Research, Inc., 2016. Mathematica 10.4. URL https://www.wolfram.com.

Yaglom, A.M., 1955. Correlation theory of processes with random stationary n-th increments. Mat. Sb. 37, 141–196 (in Russian), English translation in American Mathematical Society Translations, Series 2, Vol. 8, American Mathematical Society (1958) 87–141.

Zammit-Mangion, A., Cressie, N., 2017. FRK: An R package for spatial and spatio-temporal prediction with large datasets. pp. 1–41, ArXiv e-prints http://arxiv.org/abs/1705.08105.