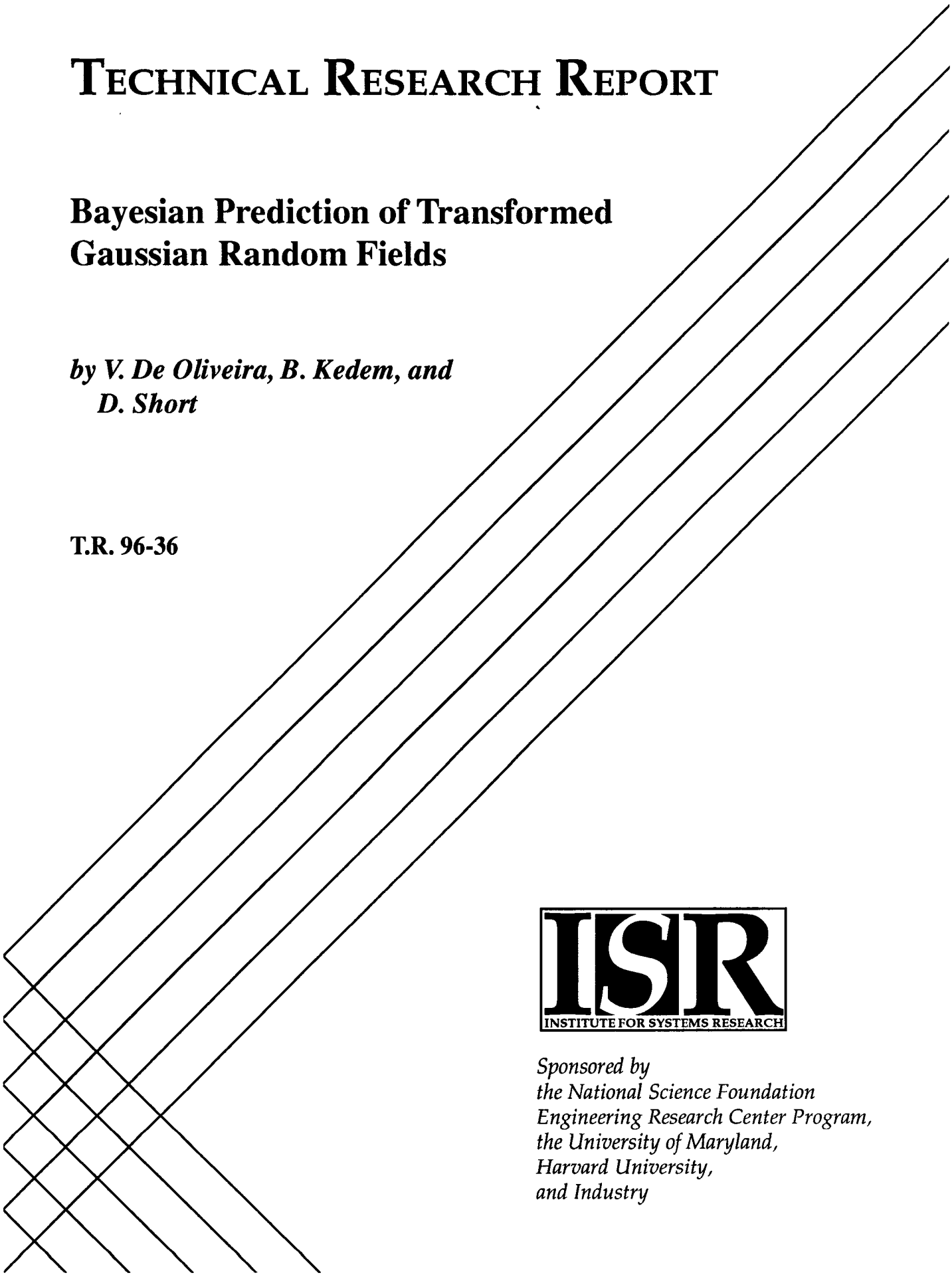
TECHNICAL RESEARCH REPORT

Bayesian Prediction of Transformed Gaussian Random Fields

by V. De Oliveira, B. Kedem, and D. Short

T.R. 96-36



Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Bayesian Prediction of Transformed Gaussian Random Fields

Victor De Oliveira[a], Benjamin Kedem[a], and Dave Short[b]

[a]Department of Mathematics and Institute for Systems Research
University of Maryland
College Park, Maryland 20742
U.S.A.

[b]Laboratory for Atmospheres
NASA/GSFC
Greenbelt, Maryland 20771
U.S.A.

April, 1996[1]

[1] Work supported by NASA grant NAG52783 and NSF grant EEC-940-2384

Abstract

The purpose of this work is to extend the methodology presented in Handcock and Stein (1993) for prediction in Gaussian random fields to the case of transformed Gaussian random fields when the transformation is only known to belong to a parametric family. As the optimal predictor, the median of the Bayesian predictive distribution is used since the mean of this distribution does not exist for many commonly used nonlinear transformations. Monte Carlo integration is used for the approximation of the predictive density function, which is easy to implement in this framework. An application to spatial prediction of weekly rainfall amounts in Darwin Australia is presented.

Key words:
Box-Cox transformation, cross-validation, Monte Carlo integration, rainfall, spatial prediction, trans-gaussian kriging.

1 Introduction

Optimal statistical prediction (interpolation) in random fields is a very important problem for the study and analysis of spatial data coming from natural sciences such as epidemiology, geology, hydrology, and meteorology, to mention a few. There, some quantity of interest, say Z, varies over a certain domain in space (and/or time) in a complex way, far from being understood. The quantity Z is measured at a finite set of locations, but inference about Z (or some other closely related quantity) is required for many other ungauged -and often ungaugeable- locations in the domain of interest. The observed data are viewed as part of a realization of a *random field*, and the unobserved part of the same realization is predicted using location-dependent covariates, while exploiting the existing dependence structure in the random field. Inference is summarized by the pair $(\hat{Z}(\mathbf{s}_0), \hat{\sigma}(\mathbf{s}_0))$, where $\hat{Z}(\mathbf{s}_0)$ is the predictor for $Z(\mathbf{s}_0)$, the value of Z at location $\mathbf{s}_0$, and $\hat{\sigma}(\mathbf{s}_0)$ is a measure of prediction uncertainty associated with $\hat{Z}(\mathbf{s}_0)$. An example of this is the commonly used technique of kriging (Cressie, 1993).

In most of the theoretical and applied work it is assumed, explicitly or implicitly, that the observations form a sample from a single realization of a Gaussian or nearly Gaussian random field. In the kriging literature, the Gaussian assumption justifies the use of linear predictors. In the Bayesian approach to spatial prediction, the Gaussian assumption is also prevalent as we can see from the recent works of Kitanidis (1986), Handcock and Stein (1993), Handcock and Wallis (1994), Brown et. al. (1994) and Gaudard et. al. (1995). However, many data sets from the natural sciences display markedly *non-Gaussian* behaviors of various kinds -asymmetric distributions often with heavy right tails, supported either on the positive real numbers or on a finite interval as in the case of proportions- making the Gaussian assumption unsatisfactory in many cases.

A natural way to model moderate departures from Gaussianity is to assume that up to a reasonable approximation, the field of interest was obtained as a result of applying an unknown nonlinear transformation from a parametric family to a Gaussian random field. The selection of the family of transformations could be based on theoretical grounds, but more often this family is used just as a modeling device aimed at mimicking some of the non-Gaussian features displayed by the type of data under study. The resulting model is flexible enough for describing different types of departures from Gaussianity and its statistical analysis is, as we will show, no more complex

than in the Gaussian case. Another alternative to prediction in non-Gaussian random fields is the recent work by Diggle et. al. (1995), which is based on generalized linear models.

The standard kriging approach to prediction in transformed Gaussian random fields, known as trans-gaussian kriging in the geostatistical literature, can be summarized as follows: find a transformation for which the transformed data are (approximately) Gaussian, compute the optimal (Best Linear Unbiased) predictor on that scale, and finally back-transform the predictor to the original scale making a bias correction to achieve unbiasedness (Cressie, 1993). This approach has some drawbacks. First, little is said about how to identify the 'normalizing transformation' (as a rule, the logarithmic transformation is chosen by default), and on how to transfer the uncertainty about the 'normalizing transformation' to the final prediction. Second, exact expressions for the unbiased predictor and the mean square prediction error (MSPE) are available only for the logarithmic transformation and a handful of others. Approximate expressions for the unbiased predictor and MSPE are available for general smooth transformations (computed via the delta method), but no indication is given about how good these approximations are, except for the requirement that the variance of the Gaussian field must be small. Finally, the predictor obtained in this way lacks some natural optimality property in the *original scale*, that is, the scale at which the actual data were measured, and the one that is generally of interest to the scientist -a meteorologist is mainly interested in efficient prediction of rainfall rather than log-rainfall.

The purpose of this work is to extend the methodology presented in Handcock and Stein (1993) for prediction in Gaussian random fields to the case of transformed Gaussian random fields where the transformation is only known to belong to a certain parametric family. This extension, following the Bayesian paradigm, provides an alternative to trans-gaussian kriging and *it mitigates* the drawbacks mentioned above. It takes into account some major sources of uncertainty, including uncertainty about the 'normalizing transformation', in the computation of the predictive density function in the original scale upon which the predictors and prediction intervals are computed (Kitanidis, 1986).

A peculiar feature in the framework we propose is that for many commonly used transformations, the predictive density function based on our model has no finite mean, requiring the use of some other functional as the predictor. We will use the *median* of the predictive density function as our

predictor, which is the *optimal predictor* corresponding to absolute error loss function.

The paper is divided as follows. Section 2 describes the transformed Gaussian random field model and the computation of the Bayesian predictive density function as well as summaries of it. Section 3 describes a numerical integration algorithm based on Monte Carlo integration. In section 4 we apply this model for the spatial prediction of weekly rainfall amounts collected in Darwin Australia, and some model checking techniques are adapted to this spatial setting. Summary and conclusions are given in section 5.

2 The Model

2.1 Model Description

Let $\{Z(\mathbf{s}),\ \mathbf{s} \in D\}$, $D \subseteq R^2$ be the random field of interest taking real values, and suppose we have n observations $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$ from a single realization of this field, where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are known distinct locations in D. Based on $\mathbf{Z}$ and on our prior knowledge about the random field, we want to predict the unobserved random vector $\mathbf{Z}_0 = (Z(\mathbf{s}_{01}), \ldots, Z(\mathbf{s}_{0k}))'$, where $\mathbf{s}_{01}, \ldots, \mathbf{s}_{0k}$ are known distinct locations in D. It is assumed that $\mathbf{Z}_0$ comes from the same realization as the data vector $\mathbf{Z}$.

Let $G = \{g_\lambda(.) : \lambda \in \Lambda\}$ be a parametric family of transformations where each $g_\lambda(.) \in G$ is a nonlinear monotone transformation, $g'_\lambda(x)$ exists and is continuous in $\Lambda \times R$. Our *main modeling assumption* is that for some unknown 'transformation parameter' λ, to a sufficient approximation we have

$$\{Y(\mathbf{s}) = g_\lambda(Z(\mathbf{s})),\ \mathbf{s} \in D\}$$

is a Gaussian random field with the following properties:

$$E\{Y(\mathbf{s})\} = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) = \boldsymbol{\beta}' \underline{f}(\mathbf{s}),\ \mathbf{s} \in D$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in R^p$ are unknown regression parameters, $\underline{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_p(\mathbf{s}))'$ is a set of known location-dependent covariates, and

$$Cov\{Y(\mathbf{s}), Y(\mathbf{u})\} = \frac{1}{\tau} K_\vartheta(\|\mathbf{s} - \mathbf{u}\|);\ \mathbf{s}, \mathbf{u} \in D$$

Here τ is the precision of the random field, $\tau^{-1} = Var\{Y(\mathbf{s})\}$, and $\boldsymbol{\vartheta} = (\theta_1, \ldots, \theta_q)' \in \Theta \subset R^q$ is a structural parameter controlling the range of correlation and/or the smoothness of the random field, where for every $\boldsymbol{\vartheta} \in \Theta$, $K_{\vartheta}(.)$ is an isotropic correlation function ($\|.\|$ denotes Euclidean distance).

The cases where either the random field $\{Z(\mathbf{s}),\ \mathbf{s} \in D\}$ or some known transformation of it is Gaussian can be considered as special cases of this model framework where the family contains only one member.

An example of a family of transformations that will be studied later, and one which is frequently used for 'normalizing' positive data, is the Box-Cox family of power transformations (Box and Cox, 1964)

$$g_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

The parametric family of probability distributions resulting from (1) is fairly rich. In particular, embedded in it are the two models most frequently used in practice, the Gaussian ($\lambda = 1$) and the log-Gaussian ($\lambda = 0$) distributions, so their fit can be evaluated as well as contrasted with that of other members of the family.

By the stated assumptions we have that

$$(\underline{g}_\lambda(\boldsymbol{Z}_0), \underline{g}_\lambda(\boldsymbol{Z}) | \boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda)' \sim N_{k+n}\left(\begin{pmatrix} X_0\boldsymbol{\beta} \\ X\boldsymbol{\beta} \end{pmatrix}, \frac{1}{\tau} \begin{pmatrix} E_{\vartheta} & B_{\vartheta} \\ B'_{\vartheta} & \Sigma_{\vartheta} \end{pmatrix} \right) \tag{2}$$

for some $\lambda \in \Lambda$ and $(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta})' \in R^p \times (0, \infty) \times \Theta$, where for any vector $\mathbf{a} = (a_1, \ldots, a_n)$ we define $\underline{g}_\lambda(\mathbf{a}) \equiv (g_\lambda(a_1), \ldots, g_\lambda(a_n))$, X and X_0 are known $n \times p$ and $k \times p$ design matrices, respectively defined by $X_{ij} = f_j(\mathbf{s}_i)$, $X_{0,ij} = f_j(\mathbf{s}_{0i})$, and E_{ϑ}, B_{ϑ} and Σ_{ϑ} are respectively $k \times k$, $k \times n$, and $n \times n$, correlation matrices defined as : $E_{\vartheta,ij} = K_{\vartheta}(\|\mathbf{s}_{0i} - \mathbf{s}_{0j}\|)$, $B_{\vartheta,ij} = K_{\vartheta}(\|\mathbf{s}_{0i} - \mathbf{s}_j\|)$, and $\Sigma_{\vartheta,ij} = K_{\vartheta}(\|\mathbf{s}_i - \mathbf{s}_j\|)$. In the sequel, it will be assumed that X has full rank and $\forall\ \boldsymbol{\vartheta} \in \Theta$, the matrix Σ_{ϑ} is nonsingular.

A random field $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ as described above will be called a g_λ-Gaussian random field, in analogy with a log-Gaussian random field.

We will denote all densities by p(.), where the argument(s) identify the respective distribution.

Under this framework we have from (2) that the likelihood of the model parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda)'$, based on the *original data* $\mathbf{z} = (z_1, \ldots, z_n)'$, $L(\boldsymbol{\eta}; \mathbf{z}) = p(\mathbf{z}|\boldsymbol{\eta})$, is given by

$$L(\boldsymbol{\eta};\mathbf{z}) = (\frac{\tau}{2\pi})^{n/2}|\Sigma_\vartheta|^{-1/2}\exp\{-\frac{\tau}{2}(\underline{g}_\lambda(\mathbf{z}) - X\boldsymbol{\beta})'\Sigma_\vartheta^{-1}(\underline{g}_\lambda(\mathbf{z}) - X\boldsymbol{\beta})\}J_\lambda$$

for $z_i \in g_\lambda^{-1}(R)$, and is 0 otherwise, where $J_\lambda = \prod_{i=1}^n |g'_\lambda(z_i)|$ is the Jacobian of the transformation.

The choice of prior distribution for the parameters requires, for this model, some care because the meaning of $\boldsymbol{\beta}, \tau$ and $\boldsymbol{\vartheta}$ depend on the realized value of λ. More specifically, each transformation (i.e. each λ) will change the location and scale of the transformed data, as well as the correlation structure, to a lesser extent though, so assuming them independent a priori of λ would give nonsensical results (Box and Cox, 1964 ; Hinkley and Runger, 1984). To determine the prior distribution for the model parameters in our present framework we use the argument originally given by Box and Cox (1964) and developed further in Sweeting (1985) for smooth families of transformations. Assume $p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}|\lambda) \propto p(\boldsymbol{\vartheta})h(\lambda)/\tau$ for some function h(.) to be determined, and let λ_1 be any reference value of λ for which the likelihood is appreciable. By the assumed smoothness of the family G, for all λ in some neighborhood of λ_1, $g_\lambda(Z(\mathbf{s}))$ will be approximately linearly related to $g_{\lambda_1}(Z(\mathbf{s}))$, that is

$$g_\lambda(Z(\mathbf{s})) \simeq a_\lambda + l_\lambda g_{\lambda_1}(Z(\mathbf{s})) \tag{3}$$

for some constants a_λ and l_λ. Now $h(\lambda)$ is chosen to make the prior distributions involving λ and λ_1 consistent with (3) which requires $h(\lambda) = l_\lambda^{-p}$. Box and Cox argued that a pragmatic choice would be to take $l_\lambda = J_\lambda^{1/n}$, the geometric mean of the Jacobian, so

$$p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}|\lambda) \propto \frac{p(\boldsymbol{\vartheta})}{\tau J_\lambda^{p/n}}$$

This is an improper distribution with the unusual feature of being dependent on the data. On the other hand, following Pericchi (1981) we have that if we start with a conditional prior of the form $p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}|\lambda) \propto p(\boldsymbol{\vartheta})h(\lambda)\tau^{\frac{p}{2}-1}$, which does not assume independence a priori between $\boldsymbol{\beta}$ and τ, then by the same consistency argument used before we have that $h(\lambda)$ must be constant, and the resulting conditional prior is no longer dependent on the data, although it has some drawbacks discussed in Sweeting (1985). In this work we will use the Box-Cox's alternative as our *reference prior*. The results of the data

analysis we perform in section 4 suggest that essentially identical predictive inferences are obtained by using Pericchi's alternative.

Our full prior specification is then given by

$$p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda) \propto \frac{p(\boldsymbol{\vartheta})p(\lambda)}{\tau J_{\lambda}^{p/n}} \tag{4}$$

where $p(\boldsymbol{\vartheta})$ and $p(\lambda)$ are the prior marginals of $\boldsymbol{\vartheta}$ and λ respectively.

2.2 Posterior of Model Parameters

We wish now to obtain the joint posterior distribution of the model parameters, which can be factored as $p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda|\mathbf{z}) = p(\boldsymbol{\beta}, \tau|\boldsymbol{\vartheta}, \lambda, \mathbf{z})p(\boldsymbol{\vartheta}, \lambda|\mathbf{z})$. The *key point* for all the statistical analysis that follow is to note that conditional on $\boldsymbol{\vartheta}$ and λ, ours is a general linear model for the transformed data $\underline{g}_{\lambda}(\mathbf{z})$, and so standard Bayesian theory for these models apply (Zellner, 1971a; Broemeling, 1985). Based on this and following Kitanidis (1986) we have that the conditional posterior $p(\boldsymbol{\beta}, \tau|\boldsymbol{\vartheta}, \lambda, \mathbf{z}) = p(\boldsymbol{\beta}|\tau, \boldsymbol{\vartheta}, \lambda, \mathbf{z})p(\tau|\boldsymbol{\vartheta}, \lambda, \mathbf{z})$ is Normal-Gamma, since

$$\begin{aligned}(\boldsymbol{\beta}|\tau, \boldsymbol{\vartheta}, \lambda, \mathbf{z}) &\sim N(\hat{\boldsymbol{\beta}}_{\vartheta,\lambda}, \frac{1}{\tau}(X'\Sigma_{\vartheta}^{-1}X)^{-1}) \\ (\tau|\boldsymbol{\vartheta}, \lambda, \mathbf{z}) &\sim Ga(\frac{n-p}{2}, \frac{2}{\tilde{q}_{\vartheta,\lambda}})\end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{\vartheta,\lambda} = (X'\Sigma_{\vartheta}^{-1}X)^{-1}X'\Sigma_{\vartheta}^{-1}\underline{g}_{\lambda}(\mathbf{z})$ is the generalized least squares estimate of $\boldsymbol{\beta}$ based on the transformed data when $\boldsymbol{\vartheta}$ and λ are known, and $\tilde{q}_{\vartheta,\lambda} = (\underline{g}_{\lambda}(\mathbf{z}) - X\hat{\boldsymbol{\beta}}_{\vartheta,\lambda})'\Sigma_{\vartheta}^{-1}(\underline{g}_{\lambda}(\mathbf{z}) - X\hat{\boldsymbol{\beta}}_{\vartheta,\lambda})$.

To compute the second factor in the joint posterior we note that $p(\boldsymbol{\vartheta}, \lambda|\mathbf{z}) = p(\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda|\mathbf{z})/p(\boldsymbol{\beta}, \tau|\boldsymbol{\vartheta}, \lambda, \mathbf{z})$, and applying Bayes theorem in the numerator we get

$$p(\boldsymbol{\vartheta}, \lambda|\mathbf{z}) \propto |\Sigma_{\vartheta}|^{-1/2}|X'\Sigma_{\vartheta}^{-1}X|^{-1/2}\tilde{q}_{\vartheta,\lambda}^{-\frac{n-p}{2}}J_{\lambda}^{1-\frac{p}{n}}p(\boldsymbol{\vartheta})p(\lambda) \tag{6}$$

The proportionality constant that makes the above a pdf can only be determined numerically. This will be accomplished in section 3 by using Monte Carlo integration.

Remark 1. By (numerically) integrating (6) with respect to $\boldsymbol{\vartheta}$, we get the marginal posterior $p(\lambda|\mathbf{z})$ upon which an estimate of λ can be obtained.

2.3 Prediction of $\mathbf{Z}_0$

The purpose of our analysis is to make conditional inference about $\mathbf{Z}_0$, an unobserved part of a realization of the random field, letting the parameter vector $\boldsymbol{\eta}$ play only an instrumental role, although its uncertainty should be recognized and taken into account. That is, to base the prediction of $\mathbf{Z}_0$ on the *Bayesian predictive density function* in the original scale (Aitchison and Dunsmore, 1975)

$$\begin{aligned} p(\mathbf{z}_o|\mathbf{z}) &= \int_\Omega p(\mathbf{z}_o, \boldsymbol{\eta}|\mathbf{z}) d\boldsymbol{\eta} \\ &= \int_\Omega p(\mathbf{z}_o|\boldsymbol{\eta}, \mathbf{z}) p(\boldsymbol{\eta}|\mathbf{z}) d\boldsymbol{\eta} \end{aligned} \tag{7}$$

where $\mathbf{z}_o = (z_{o1}, \ldots, z_{ok})'$, and $\Omega = R^p \times (0, \infty) \times \Theta \times \Lambda$. Evidently, the Bayesian predictive density function is obtained from both the subjective and data based information available about $\mathbf{Z}_0$. In this approach to prediction, the transformation parameter λ is just another of our uncertain parameters. Instead of choosing a single λ, pretending it is known, and predicting with the resulting model, all the entertained models are used in the predictive inference, and the uncertainty about λ is naturally transferred to the final prediction. We call this the *full Bayesian approach.*

The joint posterior distribution $p(\boldsymbol{\eta}|\mathbf{z})$ is obtained as the product of the densities in (5) and (6), while from (2) we have $(\underline{g}_\lambda(Z_0)|\boldsymbol{\beta}, \tau, \boldsymbol{\vartheta}, \lambda, \mathbf{z}) \sim N_k(M_{\vartheta,\lambda}, \frac{1}{\tau} D_\vartheta)$, where

$$\begin{aligned} M_{\vartheta,\lambda} &= B_\vartheta \Sigma_\vartheta^{-1} \underline{g}_\lambda(\mathbf{z}) + H_\vartheta \boldsymbol{\beta} \\ H_\vartheta = X_0 - B_\vartheta \Sigma_\vartheta^{-1} X \quad &, \quad D_\vartheta = E_\vartheta - B_\vartheta \Sigma_\vartheta^{-1} B_\vartheta' \end{aligned}$$

Therefore

$$\begin{aligned} p(\mathbf{z}_o|\boldsymbol{\eta}, \mathbf{z}) &= (\frac{\tau}{2\pi})^{k/2} |D_\vartheta|^{-1/2} \prod_{j=1}^{k} |g_\lambda'(z_{oj})| \\ &\quad \times \exp\{-\frac{\tau}{2} (\underline{g}_\lambda(\mathbf{z}_o) - M_{\vartheta,\lambda})' D_\vartheta^{-1} (\underline{g}_\lambda(\mathbf{z}) - M_{\vartheta,\lambda})\} \end{aligned} \tag{8}$$

Analogously, as in the general linear model (Kitanidis, 1986), after integrating out analytically $\boldsymbol{\beta}$ and τ in (7) we obtain

$$
\begin{aligned}
p(\mathbf{z}_o|\mathbf{z}) &= \int_\Lambda \int_\Theta p(\mathbf{z}_o|\boldsymbol{\vartheta},\lambda,\mathbf{z}) p(\boldsymbol{\vartheta},\lambda|\mathbf{z}) d\boldsymbol{\vartheta} d\lambda \\
&= \frac{\int_\Lambda \int_\Theta p(\mathbf{z}_o|\boldsymbol{\vartheta},\lambda,\mathbf{z}) p(\mathbf{z}|\boldsymbol{\vartheta},\lambda) p(\boldsymbol{\vartheta}) p(\lambda) d\boldsymbol{\vartheta} d\lambda}{\int_\Lambda \int_\Theta p(\mathbf{z}|\boldsymbol{\vartheta},\lambda) p(\boldsymbol{\vartheta}) p(\lambda) d\boldsymbol{\vartheta} d\lambda} \qquad (9)
\end{aligned}
$$

noting that $(\underline{g}_\lambda(Z_0)|\boldsymbol{\vartheta},\lambda,\mathbf{z}) \sim T_k(n-p, m_{\vartheta,\lambda}, (\tilde{q}_{\vartheta,\lambda} C_\vartheta)^{-1})$, a k-variate Student t-distribution with $n-p$ degrees of freedom, location parameter $m_{\vartheta,\lambda}$, and scale matrix $\tilde{q}_{\vartheta,\lambda} C_\vartheta$, where

$$
m_{\vartheta,\lambda} = B_\vartheta \Sigma_\vartheta^{-1} \underline{g}_\lambda(\mathbf{z}) + H_\vartheta \hat{\boldsymbol{\beta}}_{\vartheta,\lambda}
$$

$$
C_\vartheta = D_\vartheta + H_\vartheta (X' \Sigma_\vartheta^{-1} X)^{-1} H_\vartheta'
$$

and therefore

$$
\begin{aligned}
p(\mathbf{z}_o|\boldsymbol{\vartheta},\lambda,\mathbf{z}) &= \frac{\Gamma(\frac{n-p+k}{2}) \prod_{j=1}^k |g_\lambda'(z_{oj})|}{\Gamma(\frac{n-p}{2}) \pi^{k/2} |\tilde{q}_{\vartheta,\lambda} C_\vartheta|^{1/2}} \times \qquad (10) \\
&\quad [1 + (\underline{g}_\lambda(\mathbf{z}_o) - m_{\vartheta,\lambda})' (\tilde{q}_{\vartheta,\lambda} C_\vartheta)^{-1} (\underline{g}_\lambda(\mathbf{z}_o) - m_{\vartheta,\lambda})]^{-\frac{n-p+k}{2}}
\end{aligned}
$$

By integrating $p(\mathbf{z},\boldsymbol{\beta},\tau|\boldsymbol{\vartheta},\lambda)$ with respect to $\boldsymbol{\beta}$ and τ we get

$$
p(\mathbf{z}|\boldsymbol{\vartheta},\lambda) \propto |\Sigma_\vartheta|^{-1/2} |X' \Sigma_\vartheta^{-1} X|^{-1/2} \tilde{q}_{\vartheta,\lambda}^{-\frac{n-p}{2}} J_\lambda^{1-\frac{p}{n}} \qquad (11)
$$

where the proportionality constant is independent of $\boldsymbol{\vartheta}$ and λ, so its value is irrelevant for the computation of $p(\mathbf{z}_o|\mathbf{z})$ in (9). We see that the predictive density function in (9) is a mixture of transformed noncentral t-distributions with mixing distribution $p(\boldsymbol{\vartheta},\lambda|\mathbf{z})$.

Up to this point all the computations were performed analytically. However to proceed, the integration of $\boldsymbol{\vartheta}$ and λ in (9) must be performed numerically due to the intractable form of the function in the integrand. In the next section we discuss a numerical algorithm to obtain an accurate and precise approximation for $p(\mathbf{z}_o|\mathbf{z})$ as well as summaries of it.

From now on, we specialize our analysis to the case of prediction at a single location, i.e., $k = 1$.

Once the predictive density function has been computed, the next step is to use appropriate functionals of it as predictive summaries. The most common practice is to use $E\{Z_0|\mathbf{Z}\}$ as the predictor for Z_0, which is optimal

under squared error loss, and to use $E\{Z_0|\mathbf{Z}\} \pm 2(Var\{Z_0|\mathbf{Z}\})^{1/2}$ as an approximate 95% prediction interval. In general this can not be the case for the present model. For many commonly used transformations of a Gaussian field the mean of $(Z_0|\mathbf{Z})$ does not exist. As an example suppose that we are entertaining the Box-Cox family of transformations given in (1) and $\lambda = 0$, i.e., $g_0(.) = \log(.)$. Then from (10), $(Z_0|\boldsymbol{\vartheta}, \lambda, \mathbf{Z})$ has a log-Student t-distribution. But the mean of this distribution does not exist (Zellner, 1971b), therefore neither does the mean of $(Z_0|\mathbf{Z})$. The same situation occurs for $\lambda \leq (n-p)^{-1}$.

To circumvent this, we will use as our predictor for Z_0

$$\hat{Z}_0 = \text{Median of } (Z_0|\mathbf{Z})$$

This predictor has some attractive properties. It is the *optimal predictor* corresponding to the absolute error loss function (Aitchison and Dunsmore, 1975). For the model considered here, the predictive distribution is often skewed, so even for transformations for which the mean does exist, the median seems a more sensible measure of location than the mean.

As our measure of prediction uncertainty we use the 95% (or some other level) prediction interval symmetric about $\hat{Z}_0$, which is readily obtained from $p(z_o|\mathbf{z})$.

3 Numerical Integration Algorithm

In this section we describe a numerical integration *Monte Carlo* algorithm as studied in Geweke (1989), in order to obtain approximations for $p(z_o|\mathbf{z})$ as well as summaries of it. This method compares favorably with the more traditional numerical quadrature methods since it can be easily implemented and has good convergence properties.

Let us consider the case when the prior distributions $p(\boldsymbol{\vartheta})$ and $p(\lambda)$ are proper, i.e., they integrate finitely. In this case using the second equation in (9) the algorithm to approximate the predictive distribution $p(z_o|\mathbf{z})$ goes as follows:

- Discretize the effective range of Z_0, obtaining the set $S = \{z_o^{(j)} : j = 1, \ldots, r\}$ where the approximation is sought.
- Generate independently $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_m$ i.i.d. $\sim p(\boldsymbol{\vartheta})$ and $\lambda_1, \ldots, \lambda_m$ i.i.d. $\sim p(\lambda)$.

- For all $z_o \in S$, the approximation to $p(z_o|\mathbf{z})$ is given by

$$\hat{p}_m(z_o|\mathbf{z}) = \frac{\sum_{i=1}^{m} p(z_o|\boldsymbol{\vartheta}_i, \lambda_i, \mathbf{z}) p(\mathbf{z}|\boldsymbol{\vartheta}_i, \lambda_i)}{\sum_{i=1}^{m} p(\mathbf{z}|\boldsymbol{\vartheta}_i, \lambda_i)} \tag{12}$$

where $p(z_o|\boldsymbol{\vartheta}, \lambda, \mathbf{z})$ and $p(\mathbf{z}|\boldsymbol{\vartheta}, \lambda)$ were respectively given in (10) and (11). The same algorithm was used by Gaudard et. al. (1995).

As shown in Geweke (1989), under mild regularity conditions the above estimator enjoys the following properties:

1. $\hat{p}_m(z_o|\mathbf{z}) \xrightarrow{a.s.} p(z_o|\mathbf{z})$ as $m \to \infty$. This assures us that we get an accurate approximation provided m is chosen large enough.

2. $m^{1/2}(\hat{p}_m(z_o|\mathbf{z}) - p(z_o|\mathbf{z})) \xrightarrow{d} N(0, \sigma^2)$ as $m \to \infty$, where σ^2 can be estimated consistently. This second result provides a measure of the precision of the estimator in (12). For any given z_o, a consistent estimator of the standard error in the approximation, $\frac{\sigma}{m^{1/2}}$, is $\frac{\hat{\sigma}_m}{m^{1/2}}$ where

$$\hat{\sigma}_m^2 = \frac{\sum_{i=1}^{m} (p(z_o|\boldsymbol{\vartheta}_i, \lambda_i, \mathbf{z}) p(\mathbf{z}|\boldsymbol{\vartheta}_i, \lambda_i) - \hat{p}_m(z_o|\mathbf{z}))^2}{m^2} \tag{13}$$

and $m\hat{\sigma}_m^2 \xrightarrow{a.s} \sigma^2$.

Remark 2. In the case that $p(\boldsymbol{\vartheta})$ and/or $p(\lambda)$ are improper, the Bayesian predictive density can be approximated using Monte Carlo integration by *importance sampling.* See Geweke (1989) for details.

4 Application to Rainfall Prediction

The data set that will be analyzed in this section is formed by rainfall totals, in mm, accumulated over a period of 7 days, from the 76th to the 82th day of 1991 in Darwin Australia, which is part of the rainy season there. The rainfall was measured using tipping buckets in $n = 24$ stations located in a region -called the D-scale- of about 12 km$\times$12 km. See Figure 1(a). This region is located on the coastal plain of the Adelaide river where the terrain is very flat and no orographic patterns are expected. A schematic description of the D-scale region, location of the stations and rainfall amounts is shown in Figure 1(b).

The family of transformations entertained in this application is the Box-Cox parametric family given in (1), which offers a great deal of flexibility in 'normalizing' positive data. By performing some exploratory data analysis, no significant relation between rainfall totals and the spatial coordinates was found. In the absence of additional covariate information we assume a model with constant mean, so $p = 1$ and $E\{Y(\mathbf{s})\} = \beta_1$. A histogram of the 24 observations appears in Figure 1(c), suggesting that our data set was generated by a skewed distribution.

As our working family of isotropic correlation functions we use the general exponential correlation function (Yaglom, 1987, page 364 ; Diggle et. al., 1995)

$$\begin{aligned} K_\vartheta(l) &= \exp\{-\nu l^{\theta_2}\} \\ &= \theta_1^{l^{\theta_2}} \end{aligned} \tag{14}$$

where l represents Euclidean distance, and $\nu > 0$, $\theta_1 = e^{-\nu} \in (0,1)$ and $\theta_2 \in (0,2]$ are unknown parameters. In what follows we will work with the second parameterization in (14) since it eases the interpretation and the required numerical integrations.

This family, which contains the exponential ($\theta_2 = 1$) and the squared exponential ($\theta_2 = 2$) correlation functions as two of its members, is easy to compute and is parameterized by physically interpretable quantities. θ_1 controls the *range of correlation* and is viewed as the correlation between two observations 1 km apart; for any fixed θ_2, the correlation between observations decays with distance faster for small values of θ_1 when compared to large values. θ_2 controls the *smoothness* of the random field; it is mean square continuous for $\theta_2 \in (0,1]$ while for $\theta_2 \in (1,2]$ is mean square differentiable. This is a flexible family covering an ample spectrum of behaviors regarding range of correlation and smoothness, two aspects of a random field that strongly influence the shape of the predictive density function, specially its spread. Another large, but (computationally) more complicated family of correlation functions was used in Handcock and Stein (1993) and Handcock and Wallis (1994).

We assume that θ_1 and θ_2 are independent a priori and assign them non-informative prior distributions; $\theta_1 \sim$ Unif(0,1) and $\theta_2 \sim$ Unif(0,2]. Likewise, since little or no prior information is available about the 'normalizing transformation', we assume $\lambda \sim$ Unif(-2,2).

Remark 3. Strictly speaking $\boldsymbol{\vartheta} = (\theta_1, \theta_2)'$ defines the correlation structure of the transformed data, which depends on the unknown λ, making elicitation of informative priors quite troublesome in this case, and the use of noninformative priors the natural choice. Nevertheless, transforming the original data does not change dramatically the correlation structure due to the smooth nature of this family of transformations; by setting $\lambda_1 = 1$ in (3) follows that $Corr\{g_\lambda(Z(\mathbf{s})), g_\lambda(Z(\mathbf{u}))\} \simeq Corr\{Z(\mathbf{s}), Z(\mathbf{u})\}$, at least for λ close to 1.

Applying the algorithm described in section 3 with $m = 200$, in formula (12), we computed the predictive density function for $Z(\mathbf{s}_0)$ as well as its median and 95% prediction interval for the locations $\mathbf{s}_0$ = (6,5), (9,7), (3,4), (7,9), (8,3), and (4,9), (they are marked with an 'x' in Figure 3(a)) covering different sections of the region of interest. These are plotted in Figure 2. We see in this case that the predictive densities are close to being symmetric and have different location and spread characteristics depending on the relative positions of $\mathbf{s}_0, \mathbf{s}_i$, $i = 1, \ldots, 24$, and on the data vector z. Note also the large prediction uncertainty mainly due to the small sample size and lack of covariate information. Plots (not shown here) indicate that the standard errors given in (13) are all on the order of 10^{-5}, providing an adequate precision for our purposes. If needed, greater precision can be achieved by increasing m.

4.1 Checking model adequacy

In order to check the adequacy of our model for prediction purposes, we use a *cross-validation* approach based on single-point-deletion predictive distributions as described in Gelfand et. al. (1992).

Let $Z_i = Z(\mathbf{s}_i)$ be the random variable, $z_{i,obs}$ the observed value of Z_i and $\mathbf{z}_{(i)} = (z_{1,obs}, \ldots, z_{i-1,obs}, z_{i+1,obs}, \ldots, z_{n,obs})'$, the data vector with the i-th observation deleted, $i = 1, \ldots, 24$. (the observations were ordered in an arbitrary way, shown in Figure 3(a), having no influence on the analysis). The model checking is based on the predictive distributions $p(z_i|\mathbf{z}_{(i)}), i = 1, \ldots, 24$. The idea is that if the model is adequate for prediction, then based on $\mathbf{z}_{(i)}$ we expect to be able to predict Z_i reasonably well on the average. It should be noted that this predictive approach to model checking is the most natural in the present situation since prediction is the intended use of the model.

Many proposals have been suggested to measure closeness between the

predicted and observed values. One of them, adapted to our situation where predictive distributions have no finite moments, is to use the following standardized residuals: $r_i = (z_{i,obs} - \hat{z}_i)/u_i$, where $\hat{z}_i$ = Median of $(Z_i|\mathbf{z}_{(i)})$ and $u_i = L_i/4$, L_i = length of the 95% prediction interval for Z_i based on $p(z_i|\mathbf{z}_{(i)})$, $i = 1, \ldots, 24$. If the model is adequate, the average of these residuals is expected to be close to zero, with not many large values. These residuals are plotted for each location in Figure 3(b). We have $\overline{r} = 0.17$, and only two out of 24 have absolute values slightly larger than 2, the ones corresponding to stations numbered 16 and 17. Note that both stations are on the border of the convex hull determined by all the stations where prediction becomes harder. By inspecting the signs of these residuals, no over or underprediction tendency is noted, and the model seems to perform adequately.

On the other hand, we also plotted all the predictive distributions $p(z_i|\mathbf{z}_{(i)})$, $i = 1, \ldots, 24$ in Figure 4, where a vertical line was placed at each $z_{i,obs}$. Following Geisser and Eddy (1979), the idea now is to consider each of these densities as a quasi-likelihood; the larger $p(z_{i,obs}|\mathbf{z}_{(i)})$ is, the better the model is predicting at the i-th location. In our case, in about half of the cases $z_{i,obs}$ is very close to the mode of $p(z_i|\mathbf{z}_{(i)})$ and in just two of the cases $z_{i,obs}$ fails to be inside the corresponding 95% prediction intervals (again stations numbered 16 and 17). No over or underprediction tendency is noted, and the model seems also to perform adequately under this criterion. In fact, these cross-validation predictive densities can be inspected in different ways to extract useful information. For example, the predictive densities corresponding to the neighboring stations 9 and 13 have significantly larger uncertainty than the rest of the stations, suggesting that prediction is harder in this section of the region. In contrast, the predictive densities corresponding to the neighboring stations 15 and 19 display smaller uncertainty than the rest, suggesting that prediction is easier in this section of the region.

It must be noted that none of these techniques pretend to prove the correctness of the model, but just detect the presence of blunders or locations with troublesome prediction. They are fully exploratory, and in particular because of the strong dependence among the statistics $\{r_i\}_{i=1}^n$, no formal inference has been contemplated.

We also investigated the sensitivity of the predictive distributions to the choice of the prior conditional distribution $p(\boldsymbol{\beta}, \tau, \vartheta|\lambda)$ given in section 2. By comparing the results we obtained in Figure 2 using Box-Cox's alternative with the predictive distributions corresponding to the same locations

obtained using Pericchi's alternative, also given in section 2, we found that for each one of these locations, the two predictive densities are visually indistinguishable, and for all practical purposes, they provide identical inferences.

Finally, we computed the marginal posterior distribution of λ, which is plotted in Figure 5. Note that the Gaussian model, corresponding to $\lambda = 1$, is not a good approximation for our data set since this posterior gives negligible density to this value. The log-Gaussian model, corresponding to $\lambda = 0$, is a frequently used alternative but is not the best single choice for this data set. The maximum a posteriori (MAP) estimate of λ is $\tilde{\lambda} = -0.5$.

5 Summary and Conclusions

In this work we have presented a Bayesian methodology for prediction in transformed Gaussian random fields where the transformation is only known to belong to a parametric family. It provides an alternative to trans-gaussian kriging for prediction in non-Gaussian random fields. This approach mitigates some of the drawbacks of trans-gaussian kriging mentioned in the introduction. The predictors and prediction intervals are based upon the Bayesian predictive density, accounting for major sources of uncertainty and therefore producing a more realistic inference. An application of this methodology was given for the spatial prediction of rainfall amounts.

References

[1] Aitchison, J. and Dunnsmore, I.R. (1975). Statistical Prediction Analysis. Cambridge University Press.

[2] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). J. Roy. Stat. Soc. B, 26, 211-252.

[3] Broemeling L.D. (1985). Bayesian Analysis of Linear Models. New York: Marcel Dekker.

[4] Brown, P.J., Le, N.D. and Zidek, J.V. (1994). Multivariate spatial interpolation and exposure to air pollutants. Can. J. Stat., 4, 489-509.

[5] Cressie, N.A.C. (1993). Statistics for Spatial Data. Rev. ed. New York: John Wiley.

[6] Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1995). Non-Gaussian geostatistics. Technical report MA95/103. Department of Mathematics and Statistics. Lancaster University.

[7] Gaudard, M., Karson, M., Linder, E. and Sinha, D. (1995). Bayesian spatial prediction. Submitted.

[8] Geisser, E. and Eddy, W.F. (1979). A predictive approach to model selection. J. Amer. Stat. Assoc. 66, 153-160.

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

[10] Geweke, J. (1989). Bayesian inference in Econometric models using Monte Carlo integration. Econometrica 24, 1317-1339.

[11] Handcock, M.S. and Stein, M.L. (1993). A Bayesian analysis of kriging. Technometrics 35, 4, 403-410.

[12] Handcock, M.S. and Wallis, J.R. (1994). An approach to statistical spatio-temporal modeling of meteorological fields. J. Amer. Stat. Assoc. 89, 368-378.

[13] Hinkley, D.V. and Runger, G. (1984). The analysis of transformed data (with discussion). J. Amer. Stat. Assoc., 79, 302-320.

[14] Kitanidis, P.K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. Water Resour. Res. 22, 449-507.

[15] Pericchi, L.R. (1981). A Bayesian approach to transformations to normality. Biometrika, 68, 35-43.

[16] Sweeting, T.J. (1985). Consistent prior distributions for transformed models. Bayesian Statistics 2, Valencia University Press, 755-762.

[17] Yaglom, A.M. (1987). Correlation Theory of Stationary and Related Random Functions I. Basic Results. Springer-Verlag.

[18] Zellner, A. (1971a). An Introduction to Bayesian Inference in Econometrics. New York: John Wiley.

[19] Zellner, A. (1971b). Bayesian and Non-Bayesian analysis of the log-normal distribution and log-normal regression. J. Amer. Stat. Assoc. 66, 327-330.

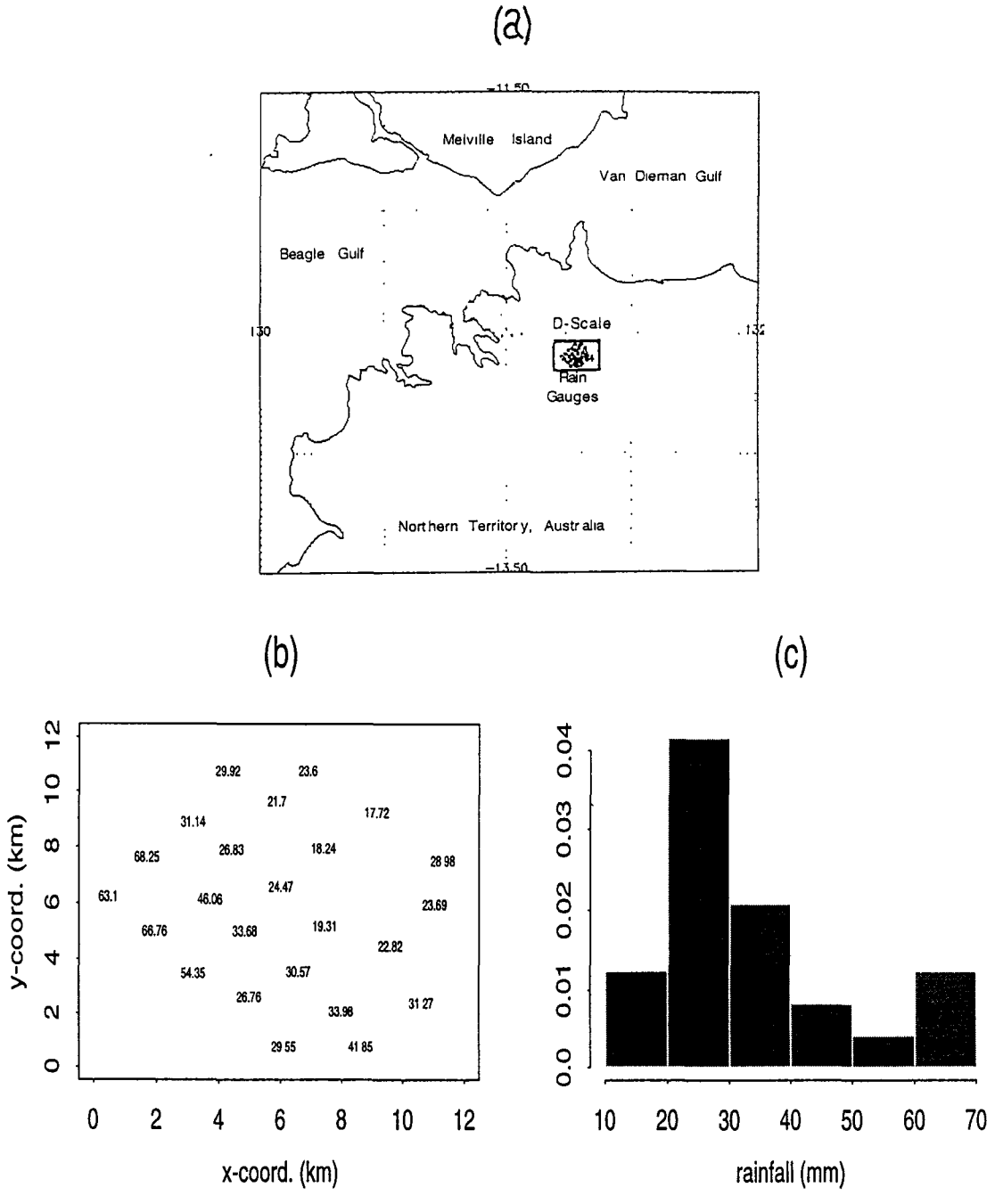


Figure 1: (a) Geographical location of the D-scale region in Darwin Australia. (b) Position of the 24 stations and their respective weekly rainfall amounts. (c) Histogram of the weekly rainfall amounts.

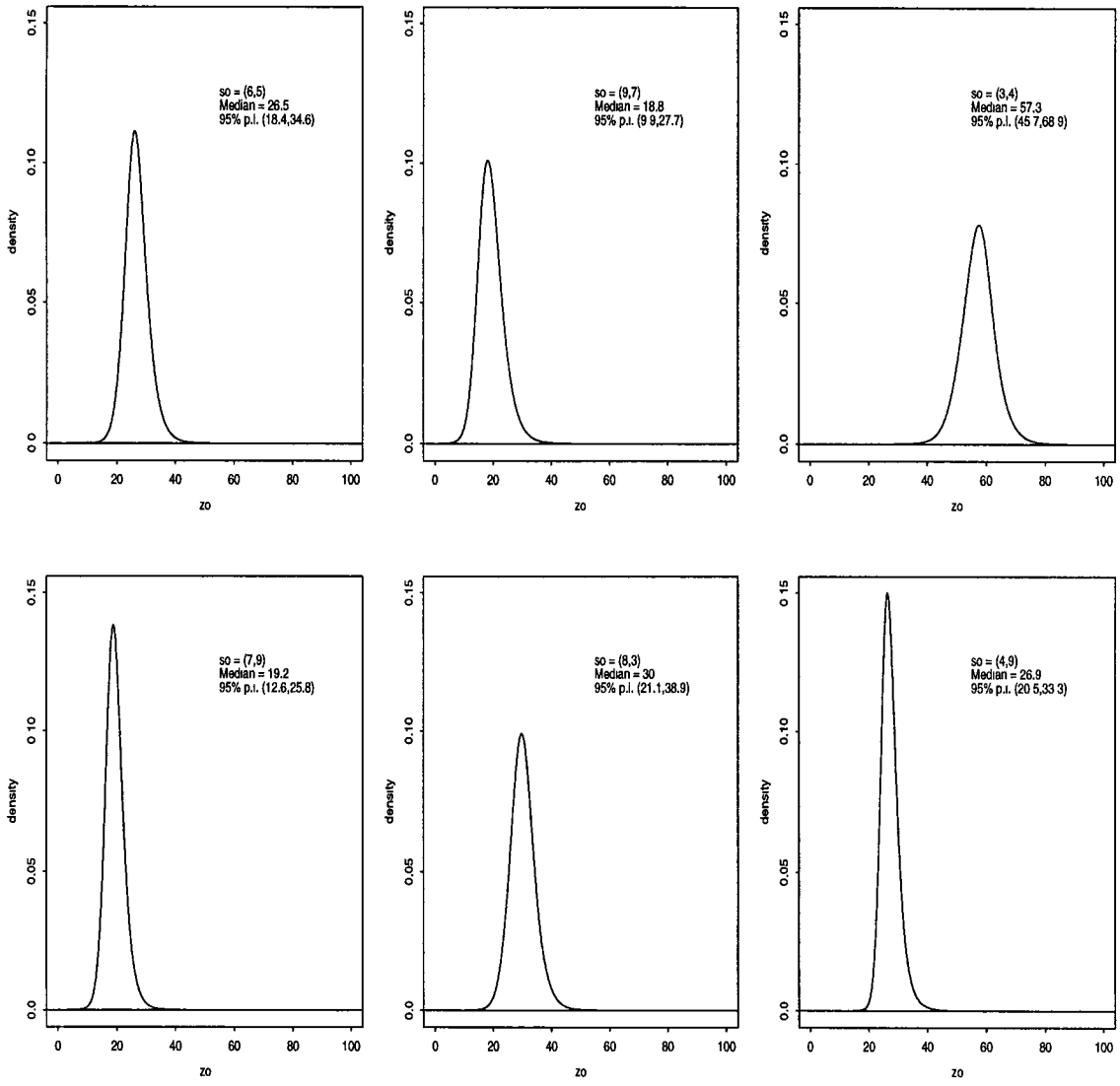


Figure 2: Bayesian predictive density functions, medians and 95% prediction intervals for 6 locations.

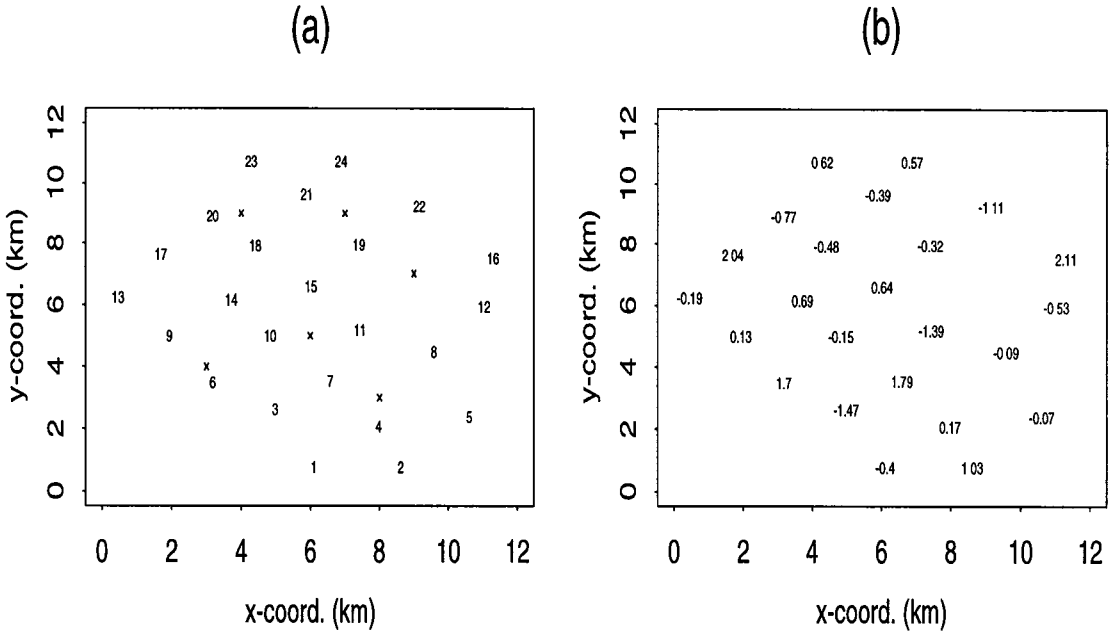


Figure 3: (a) Enumeration of the stations. An 'x' corresponds to a location where prediction was made in Figure 2. (b) Cross-validation standardized residuals $\{r_i\}_{i=1}^{24}$, $\overline{r} = 0.17$.

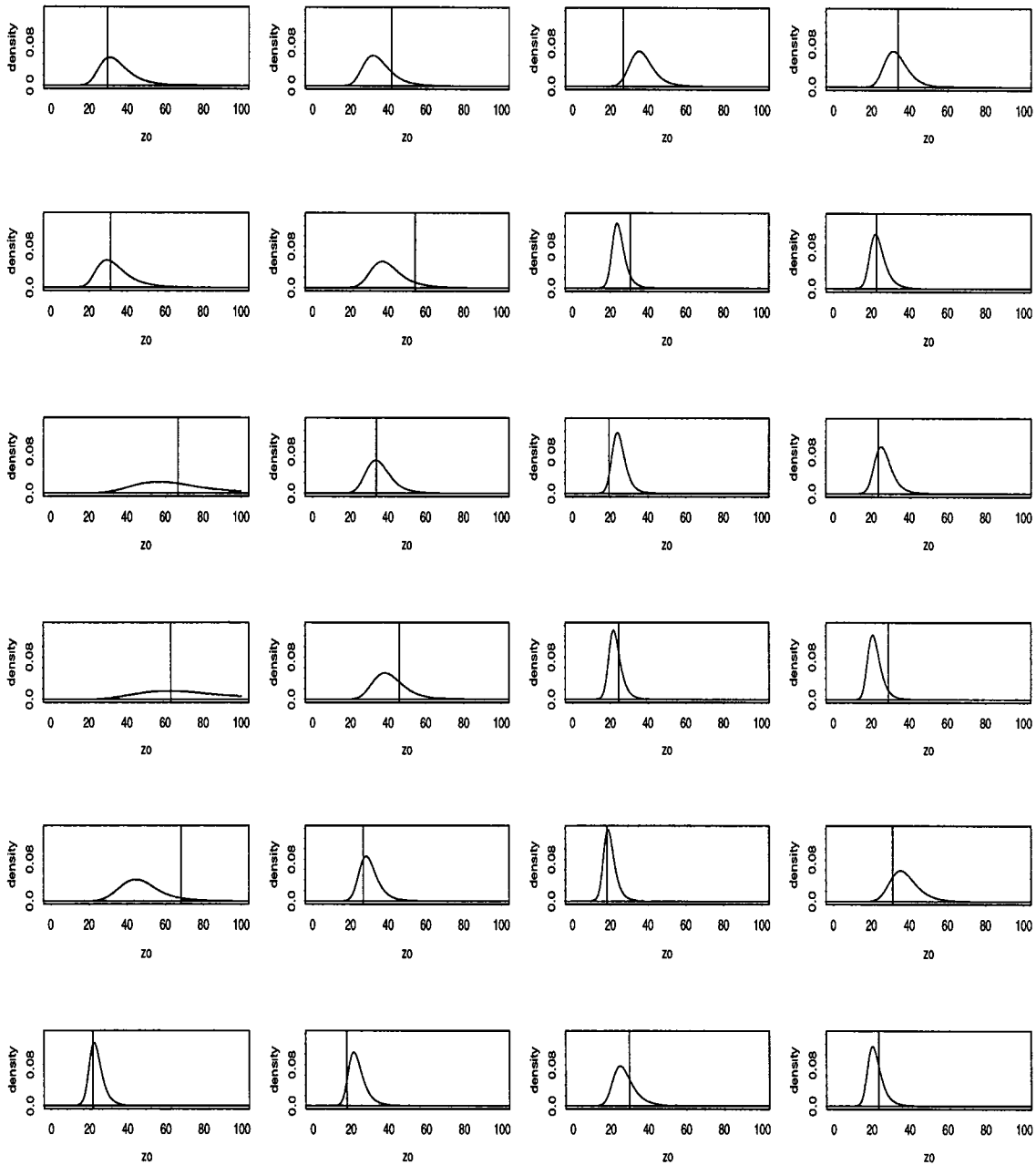


Figure 4: Cross-validation Bayesian predictive density functions for locations $1, \ldots, 24$ corresponding to Figure 3(a). Vertical lines are placed at $z_{i,obs}$, $i = 1, \ldots, 24$.

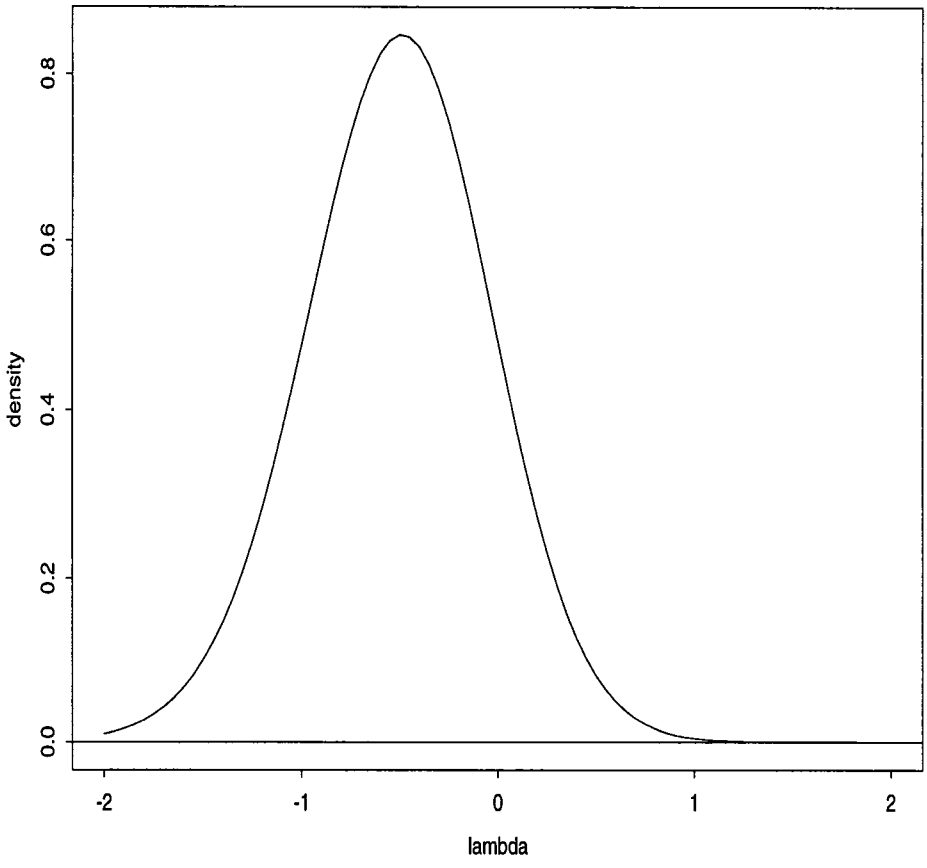


Figure 5: Marginal posterior distribution of λ.