Using Random Effects to Build Impact Models When the Available Historical Record Is Short

FILIPE AIRES*

Estellus, S.A.S., Paris, France

(Manuscript received 22 June 2011, in final form 7 June 2012)

ABSTRACT

The analysis of the affect of weather and climate on human activities requires the construction of impact models that are able to describe the complex links between weather and socioeconomic data. In practice, one of the biggest challenges is the lack of data, because it is generally difficult to obtain time series that are long enough. As a consequence, derived impact models predict well the historical record but are unable to perform well on real forecasts. To avoid this data-limitation problem, it is possible to train the impact model over a large spatial domain by ''pooling'' data from multiple locations. This general impact model needs to be spatially corrected to take local conditions into account, however. This is particularly true, for example, in agriculture: it is not efficient to pool all of the spatial data into a single very general impact model, but it is also not efficient to develop one impact model for each spatial location. To solve these aggregation problems, mixed-effects (ME) models have been developed. They are based on the idea that each datum belongs to a particular group, and the ME model takes into account the particularities of each group. In this paper, ME models and, in particular, random-effects (RE) models are tested and are compared with more-traditional methods using a real-world application: the sales of salt for winter road deicing by public service vehicles. It is shown that the performance of RE models is higher than that of more-traditional regression models. The development of impact models should strongly benefit from the use of RE and ME models.

1. Introduction

The impact of weather on human activities is very important in many domains, including energy, agriculture (Adams et al. 1998; Kaylen and Koroma 1991), logistics in sales, insurance against catastrophes, tourism, and so on (Jewson and Brix 2005). Studying the impact of weather on socioeconomic activity has gained interest in two different contexts.

First, long-term impact models can be used to attenuate the impacts of climate change (Leckebusch et al. 2002; Rauber et al. 2005; Parry et al. 2007). Two aspects of climate evolution can be considered: 1) the impacts of the trend in climate variables such as the global temperature on the mean-state weather (e.g., desertification in North Africa) and 2) changes in the probability of extreme events that also can strongly affect human activities (e.g., intensity and frequency of heavy rainfall events). These two types of climate modifications lead to many important alterations in human response: changes in agriculture practice, awareness of higher risks from hurricanes, or the modification of energy consumption and production. Therefore, it is important to obtain models that establish the link between the climate and the underlying human activity. ''Climate or weather indices'' have been developed for this purpose (depending on the time scale under consideration). These indices are based on weather or climate information and are designed to be as correlated as possible with the human activity under study. For example, Yu et al. (2009) define a weather index that assesses the impact of climate change on tourism activity. A weather index can also be built for agriculture applications to describe the sensitivity of local yield to weather conditions (Bryla and Syroka 2007; Sultan et al. 2009). In this paper, we will use the term weather ''impact model'' instead of ''weather index'' since the latter term does not introduce the idea of impact on human

[*] Additional affiliation: Laboratoire de l'Etude du Rayonnement et de la Matière en Astrophysique, Centre National de la Recherche Scientifique/Observatoire de Paris, Paris, France.

Corresponding author address: F. Aires, Estellus, LERMA, Observatoire de Paris, 61, Ave. de l'Observatoire, 74014 Paris, France.
E-mail: filipe.aires@estellus.fr

DOI: 10.1175/JAMC-D-11-0125.1

activity, which is, however, the main point of these indices. Together with climate forecasts (from ''global climate models'' often associated with regional downscaling modules), impact models can be used to estimate the consequences of climate change for human activities (Schimmelpfennig 1996). This type of study allows us to optimize long-term investments like agriculture practice (Lewandrowski and Schimmelpfennig 1999), wind energy or hydraulic dams investments, and health policies (Greenough et al. 2001) and to define adequate mitigation and adaptation strategies.

Second, since the weather affects human activity directly and in real time, impact models can also be used to alleviate weather-related risks in the present climate. In some domains, weather can be the main variability factor affecting human activity, and using weather information in management processes can be essential (Marteau et al. 2004; Jewson and Brix 2005).

In this paper, we are more directly interested in this second application. The construction of an impact model is necessary for both types of applications, however. Studying the consequences of the future climate requires an impact model fitted to the past and present climate. Therefore, the innovations and technical solutions presented in this paper are directly applicable to climate change studies, too.

Once they are built, impact models can be used in different ways. 1) Insurance contracts can be derived using them: for practical reasons, meteorological variables can be easier and cheaper to measure than, for example, agriculture production. As a consequence, agriculture production can be insured against adverse weather conditions (Bryla and Syroka 2007; Sultan et al. 2009). 2) When enough people are interested in some weather-related anomaly protection, the financial market is said to be ''liquid'' because the transactions are rapid and numerous. In this case, financial tools referred to as derivative products can be defined and used. They are called weather derivatives, and a dedicated trading market has been developed (Jewson and Brix 2005; Barrieu 2009). 3) If the weather forecasts and the impact models are good enough, they can be used directly to better manage the weather-impacted activity. For example, weather forecasts can be used to optimize distribution issues for stores, to manage inventory, or to adapt the number of employees at tourist facilities. In this paper, these three strategies that benefit from the impact models (insurance, finance, or direct use) will not be considered; only the creation of statistically robust impact models able to perform reliable forecasts will be analyzed.

Impact models are generally based on a statistical regression (Sultan et al. 2009). The fitting of the models uses a data record with the meteorological variables (plus any available ancillary or a priori information) as inputs and the socioeconomic variable as output. A first difficulty is nonstationarity: if a trend exists for one of the considered variables, it then becomes difficult to apply past or present-day statistics to future climate. Because we will consider present-day impacts, this difficulty will not be addressed in this paper.

A second very important difficulty comes from the limited amount of available data in historical records. In agriculture, for example, an event represents 1 yr of production and it is generally difficult to obtain historical records for more than one or two decades. These data records need to be long and rich enough to accurately document the complex relationships that need to be captured by the impact model. Fitting a statistical model with the level of complexity that is required to represent multivariate and nonlinear relationships from agronomy, economics, or human behavior with such a limited amount of data is a challenge. A true dilemma appears: the complexity of the relationships requires as much information as possible in the inputs to the impact model, but not enough data are available to fit it. This dilemma results in the overfitting problem: the fitting of the model works very well for the historical record, but the model is unable to perform well when applied to new data (Geman et al. 1992; Hawkings 2004; Bishop 1995). The overfitting appears very often, for example, in agriculture applications because data records are short but weather–yield relationships are very complex and involve many parameters (Sultan et al. 2009).

To solve this problem, regularization techniques can be used (Tikhonov 1963; Vapnik 1997, 1998). Regularization can act on 1) the fitting algorithm, 2) the representation of the data, or 3) the structure of the model itself. This latter regularization is called structural stabilization and is often associated with a decrease in the number of degrees of freedom in the model to be fitted.

A particular structural stabilization strategy will be investigated in this paper: the idea is that, often, a general behavior for the impact model has to be adjusted from one group of data to another. This type of behavior is considered in mixed-effects (ME)[1] models. The idea of this type of model is that a general behavior can vary from one group to another (Pinheiro and Bates 2009). The fitting of the model uses all of the available data, but its behavior will be different from one group to another. The use of an ME model involves structural stabilization because the hierarchical nature of the problem allows one to reduce the number of parameters in the model (as

[1] The ME models are a particular subcase of Bayesian hierarchical models (Gelman et al. 2003).

compared with all of the models in which a particular regression would be used for each group). As a consequence, this structural stabilization regularizes the inversion. Such ME models have been used in weather impact models. For instance, in (Pezzulli et al. 2006), the electricity demand is forecast using an ME model. In Richardson and Schoeman (2004) the link between climate and ecology is investigated, and Chen et al. (2004) illustrate a very important application for agriculture.

The purpose of this paper is to show that ME models are particularly well suited to solving impact-model applications when only a limited historical record is available. To illustrate these ideas, a real-world application will be presented and solved using a particular set of ME models: the random-effects (RE) model. This method concerns a weather impact model for salt sales. Freezing and snowing conditions require the deicing of roads in winter by public service vehicles. The lower the temperature is, the higher the sales of salts are. This is a good example of a human socioeconomic activity that is exposed to weather risks. The available historic record is only about 6 yr, however, and each year includes only 3–4 months of data (i.e., winter months).

In section 2, the datasets used in this study are presented. The ME and RE models are introduced in section 3, considering single- and multi-input impact models. Section 4 presents the results, and conclusions are drawn in section 5.

2. The datasets

a. Salt sales

The monthly sales of salt are available for 6 yr and for over 118 so-called counties in England. These counties are not administrative counties but are ''selling sectors'' for a particular company. The sales occur mostly during autumn and winter months. In this study, results will be obtained using January data since this is a particularly important month for salt sales. Similar models could be developed for the other months, however, provided that they include salt sales. An ME model could also be used that uses the month as the grouping factor, but this approach is beyond the scope of this paper, which focuses on a first analysis of this problem. As a consequence, the number of samples for each county is limited to the six January months that are available. This can appear to be too limited to perform a fit of impact models, but this is a typical situation when solving a weather impact problem. Furthermore, the goal of this paper is to present tools that can solve these data-limitation issues.

The company that sells this salt to public road deicing services could be interested in two possible applications of a weather impact model. First, combining a seasonal weather forecast with the impact model can provide a sales forecast. This is very interesting in terms of logistics. Second, the company can buy insurance or financial products such as a ''weather derivative'' to reduce its risk to adverse conditions as represented, in this case, by a mild winter.

TABLE 1. Standard deviation and correlation matrix of the five monthly weather variables: maximum temperature, minimum temperature, number of air frost days, total precipitation, and number of sunny days. The statistics are performed on the pooled dataset that includes all of the January data for each one of the 118 counties.

		Correlation				
	STD	Tmax	Tmin	AirF	Rain	SunD
Tmax	1.2	1.00	0.75	−0.57	0.21	0.16
Tmin	1.6		1.00	−0.92	0.39	−0.02
AirF	5.0			1.00	−0.40	0.09
Rain	40.8				1.00	−0.38
SunD	15.4					1.00

b. Meteorological stations

The observations of 19 weather stations over England have been gathered for this study. Each of the 118 counties mentioned in the previous section has been associated with its closest weather station.

These weather stations provide monthly information for the averaged maximal temperature Tmax, the averaged minimal temperature Tmin, the number of days of air frost over the month AirF, the total rainfall (denoted as ''Rain''), and the number of sunshine days SunD. It is anticipated that lower Tmax and Tmin and higher AirF will increase salt usage and sales. The Rain and SunD information should be less related to salt sales because they are less related to temperature. Increases in rain associated with low temperature would be a good source of information because of the possible freezing of precipitation. Rainy days with cloud cover (lower SunD) are associated with higher temperatures for winter months, however, and therefore the Rain and SunD information should be more difficult to exploit.

Table 1 provides some statistics for the 19 stations, for January data only. The standard deviation (STD) for each of the available variables is given in the first column. Because only January data are used, the STDs for Tmax and Tmin are low but the number of air frost and sunshine days can vary significantly. The right part of Table 1 provides the correlation among these variables. Tmin and AirF are highly anticorrelated (−0.92). Tmax is related to Tmin (0.75) and, to a lesser extent, AirF (−0.57). Rain and SunD are less correlated with the three temperature-related variables, and therefore they provide additional information.

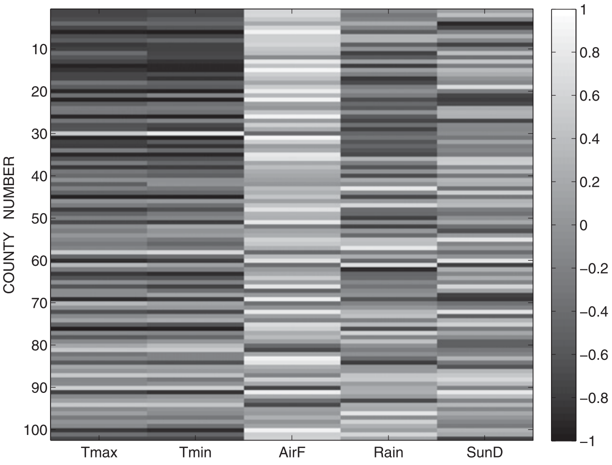

FIG. 1. Correlations between salt sales and five weather variables for the 118 available counties.

c. *Preliminary analysis of the sales–weather relationship*

A preliminary analysis to measure the relationships between the weather and salt sales can use simple correlation measures. Figure 1 provides the correlations between the five weather variables and the sales, for each of the 118 counties. It can be observed that, in general, Tmin and Tmax are negatively correlated with the sales. AirF is less (positively) correlated in amplitude. Rain and SunD are less related to sales. This general behavior varies with the county number. This can be explained by two reasons: first, only 6 months of data are available to compute the individual correlation for one county, which introduces a high level of uncertainty for the correlations. Second, the relationship linking weather and salt sales is actually dependent on the location, which encourages the use of models that consider local statistics, such as the ME models.

Table 2 provides, in the right column, the correlation between the five weather variables and salt sales when all of the January data (for all counties) are pooled in a single dataset. Tmax and Tmin have the two best correlations, equal to −0.37. AirF is also interesting, with a 0.3 correlation with salt sales. Rain and SunD appear to be less related to salt sales.

An individual county and station at 53.35°N, 2.34°W (near Pemberton) has also been chosen because its correlations are encouraging. Table 2 shows that this site has correlation statistics that are similar to those for the "pooled" dataset. Table 2 also introduces the percentage of variance explained by the five weather variables when a linear regression is used. The correlations and the percentage of variance explained by the weather variables are high for this county.

TABLE 2. Correlations (corr) between salt sales and the five monthly weather variables when the data of the 118 counties are pooled together (second column) or for a particular individual county (third column). For this county the percentage of variance explained by a linear regression (expl) is also provided. Only January data are used.

		One county	
	Pooled corr	Corr	Expl (%)
Tmax	−0.38	−0.72	0.52
Tmin	−0.37	−0.69	0.48
AirF	0.31	0.65	0.43
Rain	−0.13	−0.30	0.09
SunD	0.04	0.40	0.16

3. Mixed-effects models

a. *Motivation*

Linear ME models are an extension of traditional linear regression. Depending on the user community, they are also called hierarchical or multilevel models. ME models are designed to solve applications in which the data are "nested," meaning that the dataset under study is based on a hierarchy of different populations whose differences relate to that hierarchy. Examples of two-level hierarchy are self-similar individuals gathered into *m* groups or multiple measurements on each one of *m* entities. A special case is considered in which "repeated measurements" are performed over time on these *m* individuals—the dataset is said to be "longitudinal" (Weiss 2005). The hierarchy structure can include more than two levels; for example, statistics are performed on students that can be gathered in classes, gathered in schools, gathered in counties, and so on.

In this section, we will follow the notation introduced in Lindstrom and Bates (1988), except for some minor differences. Let us first consider the following linear model:

$$Y = \mathsf{P}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ is the vector of unobservable model parameters and ε is a random variable representing a noise term.[2] The coefficients $\boldsymbol{\beta}$ describe the linear dependency of the response variable Y (i.e., salt sales) to the predictor variables P (i.e., the weather variables). The "effects" in the model are the sensitivity of the response variable to a

[2] In this paper, vectors are represented with boldface and matrices are represented with boldface sans serif. Capitalized Latin letters denote random variables, and lowercase Latin letters represent outcomes from a random variable. Greek lowercase letters denote unobservable parameters of the model.

change in one predictor variable. The development of such a statistical model can be done for forecasting applications, but it can also be used to perform sensitivity analyses (Saltelli et al. 2000) since these effects often have an interesting (physical) meaning.

Often, the effects are supposed to be constant for the entire population of the n samples in a global dataset $\mathcal{D}$. In this case, the parameters $\boldsymbol{\beta}$ are called fixed effects because they do not change from a particular group to another. Traditional linear or nonlinear regression techniques are designed to estimate these fixed-effect parameters.

In other cases, the sample dataset can be divided into m natural groups, and the REs can be dependent on these groups. A purely RE model is given by

$$Y_j = \mathbf{P}_j^{\mathrm{T}}\mathbf{B}_j + \varepsilon_j \quad \forall \quad j = 1, \ldots, m.$$

In this case, the effects $\mathbf{B}_j$ are called random effects because their variability from one group to another is supposed to follow a random distribution (often an unbiased Gaussian distribution).

It is important to note that a particular model can have simultaneously fixed and random effects—it is then called a mixed-effects model. In this case, the predictor variable $\mathbf{P}$ can be decomposed in $\mathbf{P} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ represents the predictors related to the fixed effects and $\mathbf{Z}$ represents the predictors related to the random effects. A fixed effect should correspond to a factor that has the same weight across all groups, whereas a random effect corresponds to a factor that has different weight across groups. Fixed effects represent the general behavior, and the random effects describe the variability among the m groups.

The advantage of the ME models is that it is possible, for the resulting model, to be closer to the true nature of the data. This means that the models are able to better represent the process being analyzed, with fewer parameters than, for example, an independent linear regression for each one of the j groups.

b. Mixed-effects model

For each group j, the linear ME model is represented by

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{B}_j + \boldsymbol{\varepsilon}_j \qquad (1)$$

$$\text{with} \quad \mathbf{B}_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2\boldsymbol{\Delta}) \quad \text{and} \quad \boldsymbol{\varepsilon}_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2\boldsymbol{\Lambda}_i).$$

Here $\mathbf{Y}_j$ is a random variable representing an observation vector of dimension n_j (i.e., the number of samples in group j). Both $\mathbf{Y}_j$ and $\mathbf{Y}_k$ are independent for $j \neq k$. The random variables $\mathbf{X}_j$ and $\mathbf{Z}_j$ are observed covariates of dimension $n_j \times p$ and $n_j \times q$, respectively. Variable $\mathbf{X}_j$ is related to the p fixed effects, and $\mathbf{Z}_j$ is related to the q random effects. In addition, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the so-called fixed effects (i.e., $\boldsymbol{\beta}$ is identical for each group j), and the $\mathbf{B}_j = (B_{j1}, \ldots, B_{jq})^{\mathrm{T}}$ are $q \times 1$ vectors of the so-called random effects (the $\mathbf{B}_j$s are random realizations for each group j). The model is called mixed effects because it includes both fixed and random effects. Matrix $\sigma^2\boldsymbol{\Delta}$ represents the covariance matrix of the mixed effects. The $\boldsymbol{\Lambda}_j$ are $n_j \times n_j$ positive definite matrices, and they do not depend on j except for their size. It is often assumed that $\boldsymbol{\Lambda}_j = \mathbf{I}$, the identity matrix; we will follow this assumption here. Both $\mathbf{B}_j$ and $\boldsymbol{\varepsilon}_j$ are independent random variables.

Please note that in these notations and assumptions, the random effects $\mathbf{B}_j$ follow a centered Gaussian distribution. In the classical presentation of ME models, such as in Lindstrom and Bates (1988), the random effects are often supposed to be centered because this can be convenient for solving the model. In full rigor, random effects could have noncentered distributions. In this case, none of the covariates in $\mathbf{Z}$ (i.e., random-effect-sensitive observations) should be included in $\mathbf{X}$ (i.e., fixed-effect-sensitive observations). In this paper, Lindstrom and Bates's (1988) notation is followed using a fixed- and a random-effect term even if $\mathbf{X} = \mathbf{Z}$. It should be clear that the impact models presented in section 3e are purely random-effect models, however.

From previous equations, it follows that, marginally, the $\mathbf{Y}_j$ are independent multivariate normal vectors:

$$\mathbf{Y}_j \sim \mathcal{N}(\mathbf{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}_j),$$

where the covariance matrice $\boldsymbol{\Sigma}_j = \sigma^2(\boldsymbol{\Lambda}_j + \mathbf{Z}_j\boldsymbol{\Delta}\mathbf{Z}_j^{\mathrm{T}})$.

Considering matrices

$$\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_M)^{\mathrm{T}}, \quad \mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)^{\mathrm{T}}, \quad \text{and}$$
$$\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_M)^{\mathrm{T}}$$

and

$$\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_M),$$
$$\tilde{\boldsymbol{\Delta}} = \operatorname{diag}(\boldsymbol{\Delta}, \boldsymbol{\Delta}, \ldots, \boldsymbol{\Delta}),$$
$$\mathbf{Z} = \operatorname{diag}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_M), \quad \text{and}$$
$$\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M) \overset{\text{def}}{=} \mathbf{I},$$

the general model for the entire observation vector becomes

$$\mathbf{Y}\,|\,\mathbf{B} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{ZB}, \sigma^2\Lambda), \quad \text{where} \quad \mathbf{B} \sim \mathcal{N}(0, \sigma^2\tilde{\Delta}),$$

and the marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad \text{where} \quad \boldsymbol{\Sigma} = \sigma^2\mathbf{V}, \quad \text{with} \quad \mathbf{V} \overset{\text{def}}{=} (\Lambda + \mathbf{Z}\tilde{\Delta}\mathbf{Z}^{\mathrm{T}}). \tag{2}$$

A nonlinear extension of the linear ME model of Eq. (1) is possible (e.g., Lindstrom and Bates 1990) but will not be considered in this study.

c. Fitting of the mixed-effect model

The parameters that need to be fitted in the model of Eq. (2) are the vector $\boldsymbol{\beta}$ of the fixed-effects, the matrix $\boldsymbol{\Delta}$ related to the covariance matrices of the random-effects $\mathbf{B}_j$, and σ, the standard deviation of the residuals ε_j.

A sample dataset $\mathcal{D}$ of model inputs $(\mathbf{x}, \mathbf{z})$ and outputs y is used to estimate these parameters:

$$\mathcal{D} = \{(\mathbf{x}_{ji}, \mathbf{z}_{ji}, y_{ji}); \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, m\}.$$

Given the matrix $\boldsymbol{\Delta}$, the generalized least squares estimator in $\mathcal{D}$ provides

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\Delta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} \tag{3}$$

and the posterior mean:

$$\hat{\mathbf{B}}(\boldsymbol{\Delta}) = \tilde{\Delta}\mathbf{Z}^{\mathrm{T}}\mathbf{V}^{-1}[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\Delta})].$$

To obtain σ and $\boldsymbol{\Delta}$, a maximum likelihood (ML) estimator can be used (Lindstrom and Bates 1988). It maximizes the nonconstant log-likelihood of the marginal density of $\mathbf{Y}$ from Eq. (2):

$$l(\boldsymbol{\beta}, \sigma, \boldsymbol{\Delta}\,|\,\mathbf{Y}) = -\frac{1}{2}\log|\sigma^2\mathbf{V}| - \frac{1}{2}\sigma^2(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{4}$$

In this study, a modified estimator is actually used: the restricted maximum likelihood (RML; Harville 1974) takes into account biases introduced by the estimation of $\boldsymbol{\beta}$ in Eq. (3). Computational procedures for maximizing the RML are discussed in Lindstrom and Bates (1988) using a Newton–Raphson approach (used here[3]) or an "expectation–maximization" (EM) algorithm (van Dyk 2000).

d. Selection of predictors

As mentioned earlier, it is important to obtain a model that is able to perform well in both the fitting dataset (i.e., data used to fit the model) and the generalization dataset (i.e., new data not used during the fitting process). The ability to perform satisfactorily on the generalization dataset is called generalization. A model with too many degrees of freedom is said to be overparameterized and results in overfitting and poor generalization. A model with not enough parameters will have bad approximation results on the fitting dataset. As a consequence, the right compromise needs to be found; this is called the bias–variance dilemma (Geman et al. 1992).

An important way to reduce the number of parameters in the model is to limit the number of inputs. Some of the geophysical variables in section 3b can have a negligible information content relative to the other variables, or some can be redundant with each other. In this study, we use the Akaike information criterion: AIC = $2k - 2\log(l)$, where l is the maximized value of the likelihood function for the estimated model [Eq. (4)] and k is the number of parameters in the statistical model. This criterion describes the trade-off between bias and variance in the model construction. It will be used in the following as a means for model selection.

e. Impact models

The first impact model considered here to forecast salt sales Y is a single-input linear regression with predictor Tmin (i.e., the minimal temperature during the day). It is based on a second-order polynomial fit:

$$Y = (1, \mathrm{Tmin}, \mathrm{Tmin}^2)\boldsymbol{\beta} + \varepsilon, \quad (\text{Model 1}), \tag{5}$$

where the three parameters $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ need to be estimated during the fitting process. This model can be used on the pooled dataset or on individual counties.

The ME extension of this single-input model is given by

$$Y_j = (1, \mathrm{Tmin}, \mathrm{Tmin}^2)\boldsymbol{\beta} + (1, \mathrm{Tmin}, \mathrm{Tmin}^2)\mathbf{B}_j + \varepsilon_j, \quad \forall \quad j = 1, \ldots, m. \quad (\text{Model 2}). \tag{6}$$

In this case, the random vector $\mathbf{B}_j^{\mathrm{T}} = (B_{j1}, B_{j2}, B_{j3})$ follows a Gaussian distribution with no bias and a covariance matrix $\sigma^2\boldsymbol{\Delta}$ to be determined during the fitting.

A multi-input linear regression that includes all of the weather variables of section 2b is also tested. The AIC criterion of section 3d has been used to select the best combination of inputs. The SunD variable was suppressed from the multi-input models because it was increasing the AIC. The fixed-effect model is given by

[3] The Matlab software program was used to perform all computations in this study.

$$Y = (1, \mathrm{Tmax}, \mathrm{Tmin}, \mathrm{AirF}, \mathrm{Rain})\boldsymbol{\beta} + \varepsilon, \quad (\text{Model 3}), \tag{7}$$

where $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_5)$ need to be estimated during the fitting process. Again, this model can be used on the pooled dataset or on individual counties. The ME extension of this multi-input model is given by

$$\begin{aligned} Y_j &= (1, \mathrm{Tmax}, \mathrm{Tmin}, \mathrm{AirF}, \mathrm{Rain})\boldsymbol{\beta} \\ &\quad + (1, \mathrm{Tmax}, \mathrm{Tmin}, \mathrm{AirF}, \mathrm{Rain})\mathbf{B}_j + \varepsilon_j, \quad \forall \\ j &= 1, \ldots, m \quad (\text{Model 4}), \end{aligned} \tag{8}$$

where $\mathbf{B}_j^{\mathrm{T}} = (B_{j1}, \ldots, B_{j5})$, the random effects, follow a Gaussian distribution with no bias and a covariance matrix $\sigma^2\boldsymbol{\Delta}$ to be determined.

As was pointed out in section 3b, the notation of Lindstrom and Bates (1988) is followed here, which imposes that the random-effect random variables follow a centered distribution. This requires one to artificially introduce a fixed-effect term, but the predictor variables related to the fixed- and random-effect terms are identical ($\mathbf{X} = \mathbf{Z}$). So the models 2 and 4 in full rigor are RE models and are not really ME models.

4. Results

All of the results presented in this section are performed using January data months only. Because the size of the available datasets used in this study is limited, the method to assess the predictive ability of the RE models has been chosen to be based on a cross-validation strategy (Picard and Cook 1984). A random splitting of the global dataset is performed: 90% are used for the fitting of the model, and the remaining 10% are kept for the validation. The selection of these 10% of the data is done on the pooled dataset, which means that the validation samples can come from various counties, measuring the quality of the general model of Eq. (2) and not the behavior of the model in an individual group [Eq. (1)]. This fitting/performance assessment is conducted 50 times on randomly generated samples among the dataset $\mathcal{D}$. These 50 runs allow us to provide an average error estimate and its uncertainties (characterized here by their standard deviation in the 50 runs).

a. Single-input models using Tmin

1) The pooled version of model 1

A big constraint on this salt-sale application is the amount of data. A maximum historical record of 6 yr is available for each county. Furthermore, since the impact models need to be developed for a particular month (it would be difficult to obtain a model working for any month), a maximum of six data points only is available for each county. This is a strong limitation, and so the first strategy is to pool all of the data into a unique dataset so as to have enough samples to fit a global model. This model is called, in the following, the pooled model. Figure 2a represents the results of this complete pooling experiment. Four parameters are estimated for this model (i.e., the three fixed effects plus the noise-term STD). This model is obviously not satisfactory: it catches up the general behavior (i.e., a decreasing of the sales with increasing Tmin) but simplifies the problem too much by averaging the responses for all of the counties. The RMS errors are 2116 (STD = ±73) and 2235 (±119) for the fitting and validation datasets, respectively (Table 3). These values oscillate around the natural STD of the sales (i.e., 2143), which means that this model does not provide any valuable information.

2) The "individual" version of model 1

In Fig. 2b, the strategy is opposite: a single-input model 1 following Eq. (5) is fitted for each county (i.e., no pooling). The number of parameters that needs to be fitted for this set of models is 355 (the number of counties times 3 plus σ, the STD of the model error). The behavior is more specific to each county, but the lack of data in each one of them results in a difficult fit of the model. It can be seen that the estimation of the individual shapes is unstable, especially for the counties with a small amount of data. The convexity sign can even change from one county to another. There are simply not enough samples to constrain the three parameters of each county model. Table 3 shows that the fitting errors are very low [784 (±235)] but the RMS errors in the validation dataset are very high [15 674 (±4017)]. This proves that there is strong overparameterization and overfitting. This model is therefore not satisfactory either.

3) The RE model 2

The single-input RE model of Eq. (6) has 13 main parameters: three parameters $\boldsymbol{\beta}$ describing the general behavior, plus the 3 × 3 covariance matrix $\boldsymbol{\Delta} = \sigma^{-1}\,\mathrm{cov}(\mathbf{B}_j)$ of the Gaussian distribution describing the random effects (section 3b), plus σ, the STD of the model error. The behavior of this model is represented in Fig. 2c. It is clear here that the particular features of this model allow one to adjust the general behavior of the Tmin–Y relationship to the particularities of each county. The fitting and validation RMS errors are much better: 1291 (±67) and 1543 (±95), respectively (Table 3).

The interesting point with the RE model is that the number of parameters is a good compromise between the four parameters of the pooled model, and the 355

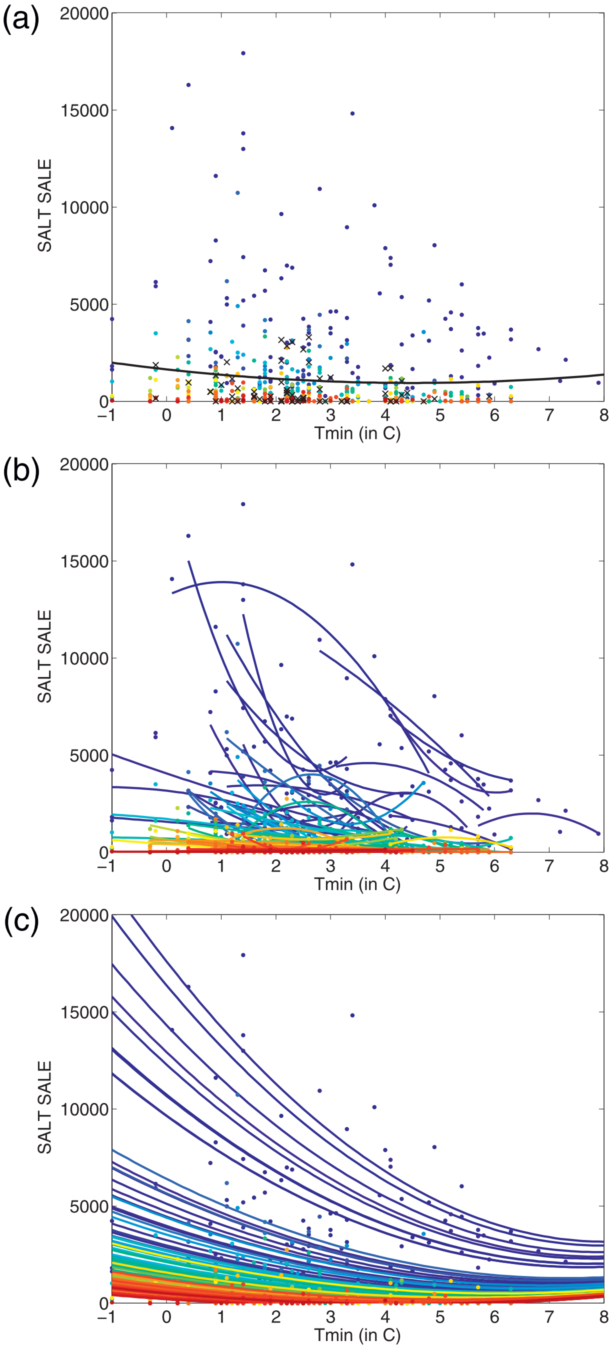

FIG. 2. Scatterplot showing Tmin and salt sales. Each color represents one of the 118 counties. (a) The black line represents the Tmin–salt sales relationship for the pooled single-input model 1. (b) The color lines represent the relationships for the 118 ''individual'' single-input runs of model 1. (c) The color lines represent the relationships for the 118 counties in the single-input RE model 2.

TABLE 3. RMS errors of the single- and multi-input models for the fitting and validation datasets. The results are provided for the pooled, individual, and RE models.

	Single input		Multiple input	
Model	Fitting	Validation	Fitting	Validation
Pooled	2116 (73)	2235 (119)	2094 (56)	2188 (87)
Individual	784 (235)	15 674 (4017)	0000 (0)	10 732 (3142)
Mixed effects	1291 (67)	1543 (95)	992 (33)	1331 (49)

parameters of the individual model. This is a structural stabilization, a very powerful regularization technique. The advantage of this approach is that all of the samples are used to fit the general behavior model (decrease in sales with increasing Tmin), but the specific data for each county are used to adjust the general behavior to each individual county.

b. *Multi-input model*

The models that will be considered in this section will use not only the Tmin information like in previous sections, but also Tmax, AirF, and Rain (see section 2b) (i.e., the SunD variable has been suppressed using the AIC criterion in section 3d). Because the models have four inputs, it is difficult to represent their behavior in graphs such as in section 3a. Similarly to the approach of the previous section, pooled and individualized standard regressions are compared with an RE model.

1) THE POOLED VERSION OF MODEL 3

The first model that is tested uses the pooled dataset: all of the counties are gathered together and a single model tries to represent a general weather–sales relationship. This is a standard linear regression model with six inputs (i.e., the five weather variables plus the intersect) and one output (i.e., salt sales), following Eq. (7). The results are close to those of the single-input pooled model: 2194 (±56) and 2188 (±87) for fitting and validation RMS errors, respectively (Table 3). Again, the results are not satisfactory: the model is able to reproduce the general weather–sales relationship, but it cannot represent the particularities of each county.

2) THE INDIVIDUAL VERSION OF MODEL 3

The other strategy is, again, to test a model for each of the 118 counties. This means that 118 × 6 parameters are determined for the set of 118 models. The RMS error for the fitting dataset is zero (Table 3): the six-parameter model is able to represent perfectly the six data values used to fit it, but this overparameterized model is unable to generalize its behavior to an independent dataset: the RMS error for the validation is equal to 10 732 (±3142). This means that this model has no value for sales forecasting.

3) THE RE MODEL 4

Last, a multi-input RE model based on Eq. (8) is trained on the entire dataset of all the counties. The

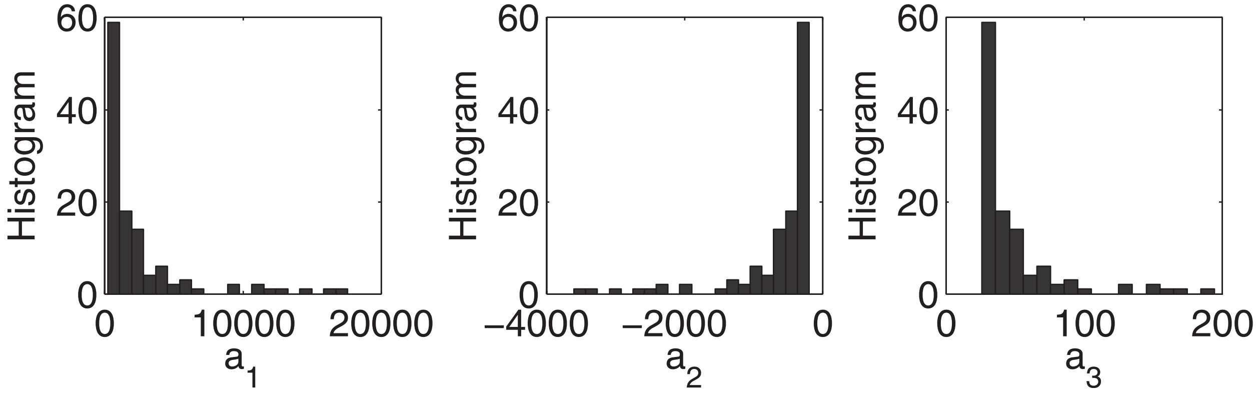

FIG. 3. Histogram of the parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ for RE model 1 using the single-input model of Eq. (5).

coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)$ of the fixed effects and the covariance matrix $\boldsymbol{\Delta}$ of the random effects are obtained using the ML fitting process. In this experiment, all six of the inputs are considered to have a random effect. Table 3 gives the RMS errors for the fitting and the validation datasets: 992 (±33) and 1331 (±49), respectively. This is the best validation score obtained in this study.

c. Discussion

The three parameters of the model of Eq. (5) are very unstable for the individual models (both single and multi-input). Uncertainties can be estimated on them; any regression package provides this type of information. Figure 3 represents the histogram of the RE model-2 parameters for each county (similar distributions would be found for the six-input RE model 4, not shown). In total, 355 parameters are actually used in Fig. 3 (i.e., 3 times the number of counties plus 1) but the regularization from the RE model allows one to reduce the degrees of freedom, and the intercounty variability appears to be reasonable. The use of an RE model allows for the shrinkage of the domain of variability of the parameters relative to the range of variability of the unstable individual models. The distribution represents the intercounty variability of the parameters. These histograms are highly non Gaussian, with strong saturation effects. The assumption that the $\mathbf{B}_j$ parameters can be represented by a Gaussian distribution appears to be a crude hypothesis. An RE model with non-Gaussian distributions would probably be more suitable for this problem. The analytical derivations for the ML fitting (section 3c) would become much more complex in this case, however. An alternative would be to preprocess the weather variables so that their probability function becomes closer to a Gaussian distribution (using, for instance, a cumulative distribution function–matching algorithm).

The covariance matrix $\boldsymbol{\Delta} = \sigma^{-1}\,\mathrm{cov}(\mathbf{B}_j)$ describing the intercounty variability of the six parameters is also obtained for the RE model 4. From this covariance matrix, a correlation matrix can be estimated (see Table 4). It can be seen that most of these correlations are significant. It is possible, in the RE fitting, to specify a priori if some of these parameters should be related. In this way, it is possible to limit even further the number of parameters in the RE models with potential regularization benefits. In the experiments of this study, however, all of the variables were set to be possibly correlated to monitor the relationships in these parameters without any a priori specifications. For example, it is interesting to note that the parameters associated with Tmax and Tmin have only a 0.18 correlation, which is surprising for these two variables that are correlated at the 0.75 level (Table 1).

Table 3 provides the RMS errors for the fitting and the validation datasets for the six considered models. It has been shown that the RE models outperform the pooled models or the individual models. Furthermore, the RE model 4 (multiple input) is better than the RE model 2 (single input): the ‘‘nonlinearity’’ in the single-input model is less important than the additional information that is provided when five weather variables are used instead of only Tmin. For this RE model 4, the fitting and validation RMS errors are still significantly different; this means that the model is still sensitive to the limited size of the fitting and validation datasets and that the random choice of the 10% data for the validation

TABLE 4. Correlation matrix from the covariance matrix $\boldsymbol{\Delta} = \sigma^{-1}\,\mathrm{cov}(\mathbf{B}_j)$ of RE model 4.

	Intercept	Tmax	Tmin	AirF	Rain	SunD
Intercept	1.00	−0.82	0.18	−0.57	0.06	
Tmax	−0.82	1.00	−0.07	0.91	0.45	
Tmin	0.18	−0.07	1.00	−0.13	−0.19	
AirF	−0.57	0.91	−0.13	1.00	0.67	
Rain	0.06	0.45	−0.19	0.67	1.00	

dataset still affects the validation results. It would be beneficial to regularize even further the fitting of the model. This has been be done by limiting the number of variables that include RE (through the use of the AIC criterion in section 3d). A preprocessing on the five input variables could also be used to reduce multicollinearities that affect any regression scheme (Aires et al. 2004). The true limiting factor here is still the size of the available dataset, however.

The RE multi-input model 4 is far from perfect; its validation RMS error is equal to 1331 (±49) when the natural variability of salt sales is 2143. This means that the model is able to explain 63.6% of the total variance of salt sales. This score is good for an impact model, however. Salt sales depend not only on weather, and, to further improve this model, more information would have to be introduced in the inputs. One very important piece of information would be, for example, the salt-stock level in each county. Time-lag effects could also be considered. Nonetheless, explaining 63.6% of the variability can already be exploited in a risk-management strategy using an insurance or financial solution, or to optimize the logistics of the salt seller. Moreover, the goal of this study is really to show the difficulties of deriving an impact model when a limited amount of data is available. It has been shown that ME models and, in particular, RE models offer a pertinent and valuable solution.

5. Conclusions and discussion

The derivation of impact models often suffers from the limited size of the datasets that are available to fit them. In many applications, a general behavior of the impact model exists; local spatial conditions modify this general behavior, however. It has been shown that ME and RE models offer a valuable solution for this kind of problem. In this paper, a real-world application has been presented. It concerns the weather impact on salt sales for road deicing. With the multi-input RE model, it is possible to explain 63.6% of the sale variability with weather information only. This is a high score for an impact model.

The perspectives for this work are numerous. First, several technical improvements can be introduced in the ME and RE models: preprocessing can be used to reduce multicollinearities in the input of the model or additional available a priori information could be introduced in the model structure. For instance, the latter can be done by choosing the random effects or by introducing particular correlation structure on them. The use of this type of ME model for other applications is very promising. In particular, agriculture would be an ideal case with limited datasets available, complex multivariate and nonlinear causal relationships, multilevel dependencies, and spatial dependency.

Acknowledgments. We thank Ralf Holz for interesting discussions linked to impact models and weather risk-management strategies.

REFERENCES

Adams, R., B. Hurd, S. Lenhart, and N. Leary, 1998: Effects of global climate change on agriculture: An interpretative review. *Climate Res.,* **11,** 19–30.

Aires, F., C. Prigent, and W. Rossow, 2004: Neural network uncertainty assessment using Bayesian statistics with application to remote sensing: 3. Network Jacobians. *J. Geophys. Res.,* **109,** D10305, doi:10.1029/2003JD004175.

Barrieu, P., 2009: Produits dérivés métérologiques et environnement (Weather derivatives and the environment). Ph.D. thesis, HEC Paris, 218 pp.

Bishop, C., 1995: Training with noise is equivalent to Tikhonov regularization. *Neural Comput.,* **7,** 108–116.

Bryla, E., and J. Syroka, 2007: Developing index-based insurance for agriculture in developing countries. Sustainable Development Innovations Briefs, No. 2, Dept. of Economic and Social Affairs, United Nations, 8 pp.

Chen, C., B. McCarl, and D. Schimmelpfennig, 2004: Yield variability as influenced by climate: A statistical investigation. *Climatic Change,* **66,** 239–261.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, 2003: *Bayesian Data Analysis.* CRC Texts in Statistical Science, Chapman and Hall, 696 pp.

Geman, S., E. Bienenstock, and R. Doursat, 1992: Neural networks and the bias–variance dilemma. *Neural Comput.,* **1** (4), 1–58.

Greenough, G., M. McGeehin, S. Bernard, J. Trtanj, J. Riad, and D. Engelberg, 2001: The potential impacts of climate variability and change on health impacts of extreme weather events in the United States. *Environ. Health Perspect.,* **109,** 191–198.

Harville, D., 1974: Bayesian inference for variance components using only error contrasts. *Biometrika,* **61,** 383–385.

Hawkings, D., 2004: The problem of overfitting. *J. Chem. Inf. Comput. Sci.,* **44,** 12–12, doi:10.1021/ci0342472.

Jewson, S., and A. Brix, 2005: *Weather Derivative Valuation—The Meteorological, Statistical, Financial and Mathematical Foundations.* Cambridge University Press, 373 pp.

Kaylen, M. S., and S. S. Koroma, 1991: Trend, weather variables, and the distribution of U.S. corn yields. *Rev. Agric. Econ.,* **13,** 249–258.

Leckebusch, G., U. Ulbrich, and P. Speth, 2002: Identification of extreme events under climate change conditions over Europe and the northwest-Atlantic region: Spatial patterns and time series characteristics. *Geophys. Res. Abstr.,* **4,** EGS02-A-01566.

Lewandrowski, J., and D. Schimmelpfennig, 1999: Economic implications of climate change for U.S. agriculture: Assessing recent evidence. *Land Econ.,* **75,** 39–57.

Lindstrom, M., and D. Bates, 1988: Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Comput. Stat.,* **83,** 1014–1022.

——, and ——, 1990: Nonlinear mixed-effects models for repeated measures data. *Biometrics,* **46,** 673–687.

Marteau, D., J. Carle, S. Fourneaux, R. Holz, and M. Moreno, 2004: *La Gestion du Risque Climatique* (*Climate Risk Management*). Gestion Collection, Economica, 211 pp.

Parry, M. L., O. F. Canziani, J. P. Palutikof, P. J. van der Linden, and C. E. Hanson, Eds., 2007: *Climate Change 2007: Impacts, Adaptation and Vulnerability*. Cambridge University Press, 976 pp.

Pezzulli, S., P. Frederic, S. Majithia, S. Sabbagh, E. Black, R. Sutton, and D. Stephenson, 2006: The seasonal forecast of electricity demand: A hierarchical Bayesian model with climatological weather generator. *Appl. Stochastic Models Data Anal.,* **22,** 113–125, doi:10.1002/asmb.622.

Picard, R., and R. D. Cook, 1984: Cross-validation of regression models. *J. Amer. Stat. Assoc.,* **79,** 575–583.

Pinheiro, J., and D. Bates, 2009: *Mixed-Effects Models in S and S-Plus.* Statistics and Computing, Springer, 530 pp.

Rauber, R., J. Walsh, and D. Charlevoix, 2005: *Severe and Hazardous Weather: An Introduction to High-Impact Meteorology.* 2nd ed. Kendall/Hunt, 558 pp.

Richardson, A., and D. Schoeman, 2004: Climate impact of plankton ecosystems in the northeast Atlantic. *Science,* **305,** 1609–1612.

Saltelli, A., K. Chan, and E. M. Scott, Eds., 2000: *Sensitivity Analysis: Gauging the Worth of Scientific Models.* John Wiley and Sons, 504 pp.

Schimmelpfennig, D., 1996: Uncertainty in economic models of climate change impacts. *Climatic Change,* **33,** 213–234.

Sultan, B., M. Bella-Medjo, A. Berg, P. Quirion, and S. Janicot, 2009: Multi-scales and multi-sites analyses of the role of rainfall in cotton yields in West Africa. *Int. J. Climatol.*

Tikhonov, A., 1963: Resolution of ill-posed problems and the regularization method. *Dokl. Akad. Nauk. SSSR,* **151,** 501–504.

van Dyk, D., 2000: Mixed-effects models using efficient EM-type algorithms. *J. Comput. Graph. Stat.,* **9,** 78–98.

Vapnik, V., 1997: *The Nature of Statistical Learning Theory.* Springer-Verlag, 188 pp.

——, 1998: *Statistical Learning Theory.* John Wiley and Sons, 736 pp.

Weiss, R., 2005: *Modeling Longitudinal Data.* Springer Texts in Statistics, Springer, 432 pp.

Yu, G., Z. Schwartz, and J. Walsh, 2009: A weather-resolving index for assessing the impact of climate change on tourism related climate resources. *Climatic Change,* **95,** D19114, doi:10.1007/s10584-009-9565-7.