Kalman filter and correction of the temperatures estimated by PRECIS model

José Ruy Porto de Carvalho a,⁎, Eduardo Delgado Assad a, Hilton Silveira Pinto b

a Senior Researchers, Embrapa Informática Agropecuária, CP: 6041, CEP: 13083-886, Campinas, SP Brazil
b CEPAGRI, Institute of Biology, Unicamp, SP Brazil

A R T I C L E   I N F O

Article history:
Received 7 April 2011
Received in revised form 8 July 2011
Accepted 11 July 2011

Keywords:
State-space model
Model systematic errors
Climate prediction

A B S T R A C T

The purpose of this study is to evaluate the accuracy of the estimation the monthly mean temperature simulated by the PRECIS model—scenarios A2 and B2 of the IPCC—for Brazilian regions and to develop a Kalman filter to correct the systematic errors of the model for the months of January to June 2010. With a regionalized model, PRECIS aims to reproduce the main features of the climate in complex terrains. The temperature estimates for January to June 2010 are based on linear regression of PRECIS simulations in each pixel of the domain for two time periods, 1961–1990 and 2070–2100. These initial estimates are adapted to 1142 observing stations by a correction using the vertical temperature gradient of the Standard Atmosphere and the difference between model and real topography. The analysis was performed using monthly observed mean temperature data from meteorological stations, along with 1142 simulated data. The PRECIS model with systematic errors was ameliorated by the application of the filter resulting in an improved mean temperature prediction of 66% above the mean square error for the dry months and above 49% for the wet months, for both scenarios under study. At the half-way point, the improvement was 68% for the A2 scenario and 69% for scenario B2.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Different statistical procedures are used in meteorology for adjusting the estimates of air temperature obtained by numeric forecasting models of climate. The linear regression methods were and are still widely used (Homleid, 2004; Glahn and Lowry, 1972). Regression techniques require a large set of data to efficiently compensate for systematic errors. The Kalman filter (Kalman, 1960; Kalman and Bucy, 1961) has the advantage of compensating systematic errors recursively in order to solve problems related to linear filtering of discrete data which does not require a data series. With the evolution of computing resources, the Kalman filter has become widely used to provide non-linear procedures (Chui and Chen, 2009) and allows its use with new techniques, such as generalized additive models (Wood, 2006; Vislocky and Fritch, 1995) or neural networks (Marzban, 2003).

The theory of the Kalman filter provides equations to recursively modify the estimates of an unknown process, combining observations related to the process and knowledge about the temporal evolution (Homleid, 1995). This means that only the state estimated from the previous time step and the current measurement is needed to calculate the estimate for the current state. The past states can be estimated, as the state provided for the present and even future states. Given some initial values one can predict and adjust the model parameters by re-measurement, obtaining the estimated error on each update.

These equations were initially developed in 1960 by Hungarian-American statistician RE Kalman in the context of the U.S. space program that would lead to the Apollo 11 moon mission in 1969. Its ability to incorporate the effects of errors and their computational structure gave the Kalman filter a wide field of application, especially with regard to trajectory analysis (Brown and Hwnag, 1997; Welch and Bishop, 2010; Faria and de Souza, 2010).
This methodology has been incorporated into other areas such as statistics, especially in demographic modeling (Cruz, 2001). Its application in the economic and financial areas (Aiube, 2005) is vast, including location robotics (Ribeiro, 2010; Habermann, 2010), tomography projections (de Souza, 2008; Cruvinel et al., 2008; Laia and Cruvinel, 2008), agriculture (Timm et al., 2000; Dourado-Neto et al., 1999), pharmaceuti- cals (Rodrigues et al., 1999), weather forecasting (Hargraves et al., 2004; Libonati et al., 2008), model rainfall- runoff (Krauskopf Neto et al., 2007a, 2007b—Parts I and II; de Araújo, 2000) and radioactive release (Monache et al., 2008). The classic references on the subject are Anderson and Moore (1979) and Jazwinski (1970).

The aim of this study is to analyze the estimated mean temperature obtained by the PRECIS model (PRECIS, 2001) in the emission scenarios A2 and B2, defined by the IPCC (Intergovernmental Panel on Climate Change) for Brazilian regions and develop a Kalman filter to correct the systematic errors of the model by comparing the observed mean temperature of weather stations that make up the Agritempo system with the values estimated by the PRECIS-Br model for the months January through June 2010.

2. Materials and methods

The PRECIS model is a system of regional climate predictions developed by the Hadley Centre in England (PRECIS, 2001), in order to generate higher resolution information on regional climate change. The original system is based on scientific formulations of the PRECIS regional model and HadAM3P, the model that provides the default lateral boundary conditions (LBCs). Both are based on the atmospheric component of the Hadley Centre's coupled climate model, HadCM3. The model was adapted for Brazilian conditions by the National Institute for Space Research (INPE) that, based on original PRECIS data, has generated two time-slices 1961–1990 and 2070–2100 of monthly georeferenced grids of 78 lines and 78 columns, with resolutions of 0.5° (what, in average, gives up to 50 km) covering the hole country, which increases the previous accuracy of surface representation (Alves and Marengo, 2009). These georeferenced grids were generated considering A2 and B2 IPCC SRES (Special Report Emissions Scenarios) (Jones et al., 2003). The scenario A2 is for the high carbon emission and the scenario B2 is for the low carbon emission. For each grid pixel, for each month across the years and for each scenario one linear regression equation it was generated to estimate monthly mean temperature for the intermediate period not generated by INPE. That gives a total of 146,016 regression equations ((78 lines × 78 columns) × 12 months × 2 scenarios × 1 temperature type). The Standard Atmosphere temperature rate of 0.65 °C to each 100 m in elevation was used to correct 2010 estimated mean temperatures based on local altitude (Romani et al., 2003; USAID, 1962). The range in height above sea level is from 0 to 1500 m. As values of the mean temperature for 2010 are estimated using linear regression, this new model was called PRECIS-Br.

The results for the monthly observed mean temperature were obtained from meteorological stations for each month that compose the Agritempo system (Agritempo, 2010) distributed in all regions of Brazil being 376 points of observations in the Midwest, 400 in the Northeast, 61 in the North, 148 in the Southeast and 157 in the South, totaling 1142 meteorological stations. The data of each meteorological station was used in correlation analysis with 1142 monthly estimated mean temperature obtained by the PRECIS-Br model. The same or the nearest grid point was associated with the geographic coordinates of the meteorological station.

The observed monthly mean temperatures in 2010 are generally higher than the PRECIS-Br estimates, according to Figs. 1 and 2 for the scenarios A2 and B2 respectively. The humid months of January, February and March have significant underestimation.

A simple Kalman filter was developed for state-space models in order to correct the errors in mean temperature prediction in degrees Celsius defined as the difference between the monthly observed mean temperature obtained with the Agritempo, and monthly mean temperature estimated by the PRECIS-Br model (tmobs − tmest) (Libonati et al., 2008; Cruvinel and Laia, 2008; Homleid, 2004, Simonsen, 1991). If the errors are normally distributed, the Kalman filter is an optimal estimator because it minimizes the mean square error of estimated parameters. The estimator is classified as excellent when the gain matrix is such that the error variance of state variables is minimal.

A model of state-space usually consists of two sets of equations, the observation or measurement equation and the equation of state or transition. Let \( y_t = \text{tmobs} - \text{tmest} \) a multivariate time series with N elements, at a given time \( t \), these variables known as observed variables are a vector \( N \times 1 \), \( y_t \in \mathbb{R}^N \). There are several ways to formulate the Kalman filter equation. The formulation and notation used was adapted from Harvey (1989) and Aiube (2005), and the theoretical treatment is described in Appendix A.

This algorithm was implemented using Interactive Matrix Language (IML) through Statistical Analysis System (SAS) (SAS, 2004). The subroutine KALCVF calculates step-by-step the expected value of the state vector \( x_t \) at time \( t+1 \) and its filtered estimate at time \( t \), as well as the covariance matrix \( P_t \). The IML procedure provides a set of optimization routines to minimize or maximize a continuous no-linear function. The Kalman filter requires a number of input values in order to work correctly, such as for the coefficients of the observation equation \( d_t = [0]_{1 \times 1} \) and \( Z_t = [1 0]_{1 \times 2} \), the coefficients of the transition equation \( c_t = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \) and \( T_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \). The initial values of the variance of \( c_t \) and \( c_t \) are 0.001. The observation error covariance \( H_t \), and the transition error covariance \( Q_t \) are estimated by means of a statistical estimation procedure. The parameters of the Kalman filter are estimated by maximum likelihood through the module LIK (SAS, 2004). The LIK module calculates the average log-likelihood function of the state-space model using the decomposition of forecast error. First, the average log-likelihood function is computed using the default initial values: \( Z_0 = 0 \) and the Variance of \( Z_0 = 10^{11} \) (I = identity matrix). Each round of the Kalman filter determines the states and observations, as well as the maximum likelihood function.

It is quite common in studies of verification using skill scores (Libonati et al., 2008), (Vernon, 1953) to summarize the quality of the forecasting system. Skill scores are statistics that express
the relative quality of a forecasting system against a forecasting system of reference.

\[ SS = \frac{QME_{\text{PrecisBr}} - QME_{\text{Kalman}}}{QME_{\text{PrecisBr}}} \times 100\% \]  

Statistic SS quantifies the relative variation of the mean square error of the Kalman filter with respect to the PRECIS-Br model. Positive values of SS indicate that the filter improved the predictions.

3. Results and discussion

The mean monthly observed temperatures and those estimated by the PRECIS-Br model \((t_{\text{mest}})\) and the Kalman filter \((t_{\text{mestkal}})\) were compared for scenarios A2 and B2 for
the months January to June 2010 respectively, as shown in Fig. 3, through the procedure SGPlot (SAS, 2008). For each month the estimated mean temperature obtained by Kalman filter was the one closest to the observed mean temperature. However, it is observed that the behavior of these estimates is different in the wet months of January through March, and for the dry months of April through June. In the period of January through March, temperatures are higher and the atmospheric stability is lower due to a higher frequency of rainfall, affecting the temperature. These figures show that this variation implies less precision in estimates of the mean temperature obtained by the Kalman filter.

Fig. 2. Graph of the monthly mean average temperature (tmobs) and monthly estimated mean temperature (tmest) by the PRECIS-Br model during the period of January to June 2010 for the IPCC B2 scenario. The dashed line (black) is the linear best fit possible.
Temperatures remain stable for long periods in the period of April through June and so the error in obtaining the estimates of mean temperature by the filter is smaller. From January through March, the Kalman filter underestimates the mean temperature. This trend will change from April through June and the Kalman filter is to overestimate the mean temperature. These findings are more accurate for the B2 scenario with low carbon emissions.

These results show that the temperature estimated by the PRECIS-Br model without the correction of the Kalman filter will have an impact on the determination of evapotranspiration of crops planted in the months of January through March lower than that estimated with the correction filter. In the two scenarios for the months of January through March, with a difference of less than 2 °C, the estimated vulnerability analysis for soybeans, corn, rice and cotton in Brazil can change, according to Assad et al. (2007) and Zullo Junior et al. (2008). This means that the PRECIS-Br model output without correction by the Kalman filter is to overestimate the mean temperature difference among the crops of the period of April through June are small, causing no significant differences in the estimated impacts for the same studies.

These results corroborate those obtained by Alves and Marengo (2009) for the analysis between the years 1960 to 1990, using the PRECIS model. They concluded that “Despite the uniformity of temperature in the equatorial region throughout the year, the model underestimates the temperature, as demonstrated in the bias maps, by around 2 °C”. The similarity of the curves (Fig. 3) of both the uncorrected (PRECIS-Br) and the corrected values (Kalman Filter) when compared with the measured data shows this upward bias is more intense than it had been estimated by the model, at least for the period studied.

The average of the error1 = tmobs - timest and error2 = tmobs - timestkal where tmobs is the mean temperature observed, timest is the mean temperature estimated by the PRECIS-Br model and timestkal the mean temperature estimated by the Kalman filter for the months of January through June and the standard deviations (SD) for the scenarios A2 and B2 are shown in Table 1.

In all cases the pattern of variation obtained by the Kalman filter (error2) is lower, showing that it is more efficient procedure in reducing systematic errors.

The histograms (SAS, 2008) of the errors distributions (observed temperatures were less than the estimated temperature) have been made for both scenarios A2 and B2, are shown in Figs. 4 and 5. In both models, the PRECIS-Br and the Kalman, the error distribution approaches the Normal, with a more precise adjustment for Kalman, centered at zero. Histograms of the Kalman filter errors are more efficient in the removal of systematic errors due to the lower tails of the distribution, as shown in Table 1. Similar results are found in Libonati et al., 2008.

The statistical Skill Score (SS) (Eq. (1)) is used to quantify, in percentage, the improvements that have occurred in estimating the mean temperature by the Kalman filter with respect to the PRECIS-Br model. In all months, in both scenarios, the Kalman filter indexes showed an improvement of over 49%, as can be seen in Table 2. That is, in all situations, the mean square error obtained by the Kalman filter (error2) is considerably smaller than the mean square error obtained by the PRECIS-Br model (error1). Similar results of improved forecasts with meteorological variables, when using the Kalman filter, can be

![Fig. 3. Monthly values for all Brazilian regions for the observed mean temperature (tmobs), mean temperature estimated by the PRECIS-Br model (timest) and mean temperature estimated by the Kalman filter (timestkal) during the period of January to June 2010 and scenarios A2 and B2 of IPCC.](image)

Table 1

<table>
<thead>
<tr>
<th>Scenario A2</th>
<th>Scenario B2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>Error 1</td>
</tr>
<tr>
<td>January</td>
<td>1.38</td>
</tr>
<tr>
<td>February</td>
<td>2.19</td>
</tr>
<tr>
<td>March</td>
<td>2.24</td>
</tr>
<tr>
<td>April</td>
<td>1.49</td>
</tr>
<tr>
<td>May</td>
<td>0.76</td>
</tr>
<tr>
<td>June</td>
<td>0.34</td>
</tr>
</tbody>
</table>
checked in Libonati et al. (2008), Anadranistakis et al. (2004), Miyagishi et al. (2010), Boi (2004) and Homleid (1995, 2004). It appears that in the rainy months, the filter obtains better estimates than in the dry months. This demonstrates that the national variability for the A2 scenario is less than for the B2 scenario, for the period studied.

4. Conclusions

- The application of the Kalman filter was effective in improving the estimated prediction of the mean temperature compared to that obtained by the PRECIS-Br model for the period studied;
The observed monthly mean temperatures in 2010 are significantly higher than the PRECIS-Br estimates, both the ones with high (A2) and the ones with low (B2) carbon emission scenarios; 

Reducing the pattern variation obtained by the Kalman filter indicates improvement in forecast predictions correcting systematic errors of the PRECIS-Br model;

The use of correction factors based upon the concept of Standard Atmosphere was not sufficient to increase the accuracy of estimates of mean temperatures obtained by the PRECIS-Br model;

The Kalman filter has proved to be a versatile technique, adapting to the different seasons, although better results were obtained in the rainy season.

Fig. 5. Histogram of the percentage of errors \((tm_{obs} - tm_{est})\) for models PRECIS-Br (error1) and Kalman (error2) during the period of January to June 2010 for the IPCC B2 scenario.
• The predictions obtained by the PRECIS-Br model were improved above 49% for the months of drought and above 66% for the wet months for both scenarios A2 and B2.

Acknowledgments

The authors wish to thank the anonymous reviewers for their valuable comments and suggestions which greatly improved the quality of the article.

Appendix A

The observed variables are related to state variables $x_t$ (unobserved variables) by measurement or observation equation:

$$y_t = d_t + Z_t x_t + e_t \quad t = 1, 2, T$$  

(A.1)

where $d_t$ is an $N \times 1$ known or unknown vector, $Z_t$ is an $N \times m$ known or unknown, constant or time varying coefficients matrix, $e_t$ is a vector of serially uncorrelated disturbances with zero mean and covariance matrix $H_t$ and $x_t$ is an $m \times 1$ vector of state variables.

The state or transition variables are generated by a first order Markov process and its equation called the equation of transition:

$$x_t = T_t x_{t-1} + c_t + \eta_t$$  

(A.2)

where $T_t$ is an $m \times m$ state transition matrix, $c_t$ is an $m \times 1$ known or unknown vector and $\eta_t$ is a $g \times 1$ vector serially uncorrelated with zero mean and covariance matrix $Q_t$. The errors $e_t$ and $\eta_t$ are not correlated and uncorrelated to the initial state and normally distributed.

The Kalman filter methodology differs from other methods through the equation of the system that allows the coefficients to vary over time. Let $\hat{x}_t$ and $\eta^m_t$ a priori state estimate at time $t$. Let $\hat{x}_t$ and $\eta^m_t$ a state estimate after the event in $t$. Previous and later measurement errors are defined as:

$$e_t^- = x_t - \hat{x}_t$$  

(A.3)

$$e_t = x_t - \hat{x}_t$$  

(A.4)

And their respective covariance of the error:

$$P_t^- = E(e_t^- e_t^T^-)$$  

(A.5)

$$P_t = E(e_t e_t^T)$$  

(A.6)

Relating the later state $\hat{x}_t$ by a linear combination of the previous state $\hat{x}_t^-$ we have:

$$\hat{x}_t = \hat{x}_t^- + K_t (y_t - Z_t \hat{x}_t^- - d_t)$$  

(A.7)

where the matrix $(m \times N)$ $K_t$ is called the Kalman gain matrix that determines the extent to which the measurement will be considered to estimate the new state vector. It is used to improve predictions of the state vector. The Kalman gain matrix is obtained by minimizing the error covariance matrix by substituting Eq. (A.7) with Eq. (A.4) and its mathematical expression given by:

$$K_t = P_t^- Z_t^T (Z_t P_t^- Z_t^T + H_t)^{-1}$$  

(A.8)

The Kalman filter has two distinct phases: prediction and correction. The prediction phase uses the estimation of the state of the previous iteration to produce an estimate of the state in the current iteration. It is responsible for advancing the state variables and covariance in time obtained, thus a priori estimates for the next instant. The correction phase is responsible for the feedback. It adds new information of the observed variable in the previous estimates to obtain an improved a posteriori estimate.

The predicted equations represent a breakthrough in t-1 time to t and are defined as:

$$\hat{x}_t^- = T_t \hat{x}_{t-1} + e_t$$  

(A.9)

where $\hat{x}_t^-$ is the optimal estimator of $x_t$ and:

$$P_t^- = T_t P_{t-1} T_t^T + R_t Q_t R_t^T$$  

(A.10)

where $P_t^-$ is an error covariance matrix of state variables. When a new observation $y_t$ is added, the estimator $\hat{x}_t$ of $x_t$ can be improved. The equations of measurement update (correction) are represented by Eqs. (A.7), (A.8) and:

$$P_t = (I - K_t Z_t) P_t^-$$  

(A.11)

The first step is to determine the Kalman gain $K_t$ given by Eq. (A.8). Then a new observation $y_t$ (Eq. A.1) is incorporated in the previous forecast $\hat{x}_t^-$ (Eq. (A.9)) with the gain matrix $K_t$ using Eq. (A.7) generating the further estimation. Finally, the error covariance matrix is obtained using Eq. (A.11). The cycle of the algorithm is repeated for the time $t+1$, where $\hat{x}_t$ and $P_t$ are input data into the Eqs. (A.9) and (A.10). These recursive natures of the model make the Kalman filter an update tool for real-time measurements and are therefore widely used in various fields of research.

References


