A Practical Approach to Sequential Estimation of Systematic Error on Near-Surface Mesoscale Grids

JOSHUA P. HACKER AND DARAN L. RIFE

National Center for Atmospheric Research,* Boulder, Colorado

(Manuscript received 27 October 2006, in final form 30 March 2007)

ABSTRACT

Statistical analysis arguments are used to construct an estimation algorithm for systematic error of near-surface temperatures on a mesoscale grid. The systematic error is defined as the observed running-mean error, and an averaging length of 7 days is shown to be acceptable. Those errors are spread over a numerical weather prediction model grid via the statistical analysis equation. Two covariance models are examined: 1) a stationary, isotropic function tuned with the observed running-mean errors and 2) dynamic estimates derived from a recent history of running-mean forecasts. Prediction of error is possible with a diurnal persistence model, where the error at one time of day can be estimated from data with lags of 24-h multiples. The approach is tested on 6 months of 6-h forecasts with the fifth-generation Pennsylvania State University–NCAR Mesoscale Model (MM5) over New Mexico. Results show that for a quantity such as 2-m temperature, the systematic component of error can be effectively predicted on the grid. The gridded estimates fit the observed running-mean errors well. Cross validation shows that predictions of systematic error result in a substantial error reduction where observations are not available. The error estimates show a diurnal evolution, and are not strictly functions of terrain elevation. Observation error covariances, localization operators, and covariance functions in the isotropic case must be tuned for a specific forecast system and observing network, but the process is straightforward. Taken together, the results suggest an effective method for systematic error estimation on near-surface mesoscale grids in the absence of a useful ensemble. Correction for those errors may provide benefits to forecast users.

1. Introduction

The maturity of limited-area, mesoscale modeling systems has occurred with an explosion in the number of users who rely on precise forecasts for regional applications. Errors in specifying the initial conditions for a forecast, and errors in model formulation, result in both random and systematic error components. The random component, often measured by error variance, can be reduced with some model improvements and improvements in data assimilation techniques. Because of the close connection with specified surface parameters such as vegetation cover, soil types, and soil temperature and moisture, the systematic component often dominates the total error in near-surface forecasts.

One type of systematic error is a bias. While bias is strictly a climate-mean error, the climate of a mesoscale model is not readily quantified because models change too frequently. But some form of systematic error can certainly be estimated from available forecast samples. Although it will be slow for samples over long time periods, that systematic error will evolve in time. Evidence suggests that this component of forecast error at observation locations can be reduced through techniques ranging from running-mean linear corrections (e.g., Stensrud and Skindlov 1996; Stensrud and Yussouf 2003) and linear Kalman filters (Roeger et al. 2003) to model output statistics (MOS) approaches (Jacks et al. 1990; Hart et al. 2004; Wilks and Hamill 2007) that benefit from longer training periods. Woodcock and Engel (2005) found that a 30-day running-mean error correction improved their operational consensus forecasts, which include MOS forecasts. These
techniques are useful for correction at points where observing platforms are located.

Other users may require corrections for the systematic error at unobserved locations. The situation may arise, for example, when spatially distributed information is needed to force a soil model. Data assimilation presents another general case because it does not intrinsically account for systematic error. A method to properly spread biases from the observing locations to unobserved grid points is needed. For this purpose, Dee and Da Silva (1998) developed a sequential estimation algorithm based on estimation theory. They stopped short of applying the full recursive filter, however, noting that it is impractical when a good error propagation model is unavailable. Rather, they chose to directly model constant background error and bias covariances, and the approach requires selection of a number of tuning parameters.

Dee (2005) relaxed the definition of bias to accommodate any kind of systematic error that can be measured by appropriate averaging techniques. He explored the possibility of using a recent history of analysis increments to define an autoregressive model for data assimilation, which is essentially the same as for a short forecase. Properties of the assimilation–forecasting system used in this work prevent application of the same approach described in Dee (2005), but both rely on the slow temporal evolution of systematic error. He also showed that in the lower atmosphere, temperature analysis errors can be correlated for several days, but also contain strong diurnal variability.

In this work we develop a simpler approach designed for estimates of systematic forecast error on a near-surface mesoscale numerical weather prediction (NWP) model grid. Mass and Kuo (1998) documented the proliferation of real-time mesoscale forecasting systems—a phenomenon that has surely continued. Many of the groups operating those systems do not have the resources necessary for advanced data assimilation systems or even the capability to handle modern observational data streams. Thus, they rely on either simple data assimilation systems such as nudging, cold starts using operational forecasts, or more advanced assimilation schemes with a severely limited set of observations. Ensemble approaches at mesoscales are currently intractable for all but the largest operational centers, with a few exceptions (e.g., Eckel and Mass 2005). Our approach is aimed at mesoscale sequential forecasting systems, though it could readily be extended to the full error estimation problem.

The primary goal is to properly spread information from simple estimates of systematic error at observed locations, such as a running mean (Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Woodcock and Engel 2005), to the surrounding grid. We consider the gridded running-mean error at a given local time of day to be the system we are trying to estimate. A suboptimal filtering algorithm is derived that uses recent analyses and forecasts to update an estimate of running-mean error on the grid. We hypothesize that a diurnal persistence model, where the error can be estimated from data with lags of 24-h multiples, is sufficient for propagating the error estimate on near-surface grids. The approach is tested with forecasts from the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (NCAR) Mesoscale Model (MM5) on a domain with \( \Delta X = 3.3 \) km centered over White Sands Missile Range in New Mexico. The domain and the observation locations are shown in Fig. 1. The complex topography makes this location a difficult test, and the area mesonet augmented with observing stations operated by the military make it tractable.

In the present work we do not consider other potentially viable approaches to systematic error estimation on a grid. Because the near-surface errors may be a strong function of elevation, vegetation, and soil characteristics, a multivariate regression approach that would include these as predictors may prove useful. Rather than seek a full understanding of these predictors, we explore a simpler approach that uses a recent history of model forecasts to provide predictors of how the error at one location relates to another.

In the next section, we describe the bias correction algorithm, the data and experiments, and the practical implementation. Section 3 presents the results of running-mean error estimate experiments on the MM5 grid. Properties of the estimates are explored more deeply in section 4, and the results from sensitivity experiments are reported. Section 5 reviews the major conclusions.

### 2. Systematic error estimation for mesoscale surface grids

The goal of this work is to estimate a gridded bias analysis from “observations” of bias, and then predict a bias correction that can be applied to a forecast. As described below, we define the “bias” to be the running-mean error at an observing station. The success of simple running means for correcting biases at observed sites (e.g., Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Woodcock and Engel 2005) motivates this approach, and these serve as our observations of
bias. Observed errors can be mapped to a grid in any number of ways, just as observations themselves are often mapped to a grid to produce an analysis of the weather. Simple approaches might resemble an objective analysis procedure such as a Cressman weighting scheme (Cressman 1959). More sophisticated approaches might use isotropic and stationary weighting functions tuned on large samples of error, such as optimal interpolation (Gandin 1965). We test such an approach and go further to devise a method that objectively includes regime dependence and the effects of lower boundary forcing (topography, land surface, etc.).

a. Data and forecasts

The forecast domain, shown in Fig. 1, is part of the real-time mesoscale data assimilation and forecast system more completely described in Davis et al. (1999) and Liu et al. (2006). The MM5 is used as the forecast engine, and an extensive set of both in situ and remotely sensed observations are used in a nudging data assimilation scheme. As with any data assimilation system, systematic error is not completely removed via the assimilation process because the climate of the model is different from that sampled with observations. In this report, we focus on temperature at 2 m ($T_2$) because, as shown, the systematic errors are large.

Forecasts with a 6-h lead time, valid between 22 February 2005 and 15 August 2005, are used to test the bias estimation. The observing network contains both National Weather Service surface stations and White Sands Missile Range surface automated mesonet stations (SAMs). As usual, temperature and moisture are observed at 2 m above ground ($T_2$ and $Q_2$), and winds are observed at 10 m. Quality control of assimilated observations is achieved by comparing innovations to expected background errors, obtained from forecast histories, for a particular type of observing platform. Forecasts at observing stations are specified with a bilinear interpolation from the grid to the observation sites.

Systematic error in temperature near the earth’s surface is a large component of the total forecast error during this experiment period. The decomposed tem-
Temperature error at 1800 UTC is shown in Fig. 2. Murphy (1988) decomposed the mean-squared error into systematic and random components, namely the mean error squared (MSE) and error variance:

$$\text{MSE} = \left[ \frac{1}{K} \sum_{k} (y_o - x_f)_k \right]^2$$

for a $K$-length time series of scalar observations $y_o$ and forecasts $x_f$. Analogously, we take the square root of each side to decompose the RMSE at each observing station. The systematic error is then defined as the absolute difference between the temporal means of the observations and the forecasts at each observing station, where the mean is taken over the entire 6-month period. The random error is defined here as the standard deviation of the error. Because the sample is not long enough and the time series may not be statistically stationary, we do not claim that this systematic error is a bias, but this decomposition gives a qualitative estimate of the contribution of bias to the total error. At nearly every station the systematic component is significant, and it dominates at the stations exhibiting the largest overall error. This characteristic suggests that the total forecast error can be meaningfully reduced by reducing the systematic error.

The small number of stations in the domain can be problematic for methods that rely on normal distributions, such as the one proposed here. Neither the total error nor the systematic error distributions are well sampled. Stations S04, S09, S17, and S30 are outliers and may dominate summary statistics that include all stations. The effects of this are examined in sections 3 and 4.

b. Running-mean error estimation

A theoretical framework for bias estimation is described in Dee and Da Silva (1998), but key components of the estimation algorithm are unavailable in reality. Namely, a dynamic model for the evolution of bias does not exist, preventing the correct spatial and temporal propagation of observed bias to unobserved locations. Observing the bias, mathematically defined as the expectation of the forecast error, is intractable in most instances. Here, we describe an approximation to rigorous bias estimation that is appropriate for surface scalar quantities and uses readily available information.

Bias is formally defined as the expectation of the forecast error. The expectation can be applied in the sense of climate, or in the sense of a conditional expectation valid only for a current dynamical regime. At observing sites, the former can be estimated as the climate-mean error, while the latter can be estimated from a sufficiently large single-model, perturbed-analysis ensemble forecast system. In the mesoscale forecast system described above, which changes frequently, neither is available.

Consider instead a running-mean error, which is another estimate of systematic error often referred to as bias, which would fit the relaxed definition of bias given in Dee (2005). This value is easily accessible in any forecast system, is appropriate for each member in a multimodel ensemble forecast system (e.g., Eckel and Mass 2005), and has been shown to be an effective predictor of some systematic forecast error (e.g., Stensrud and Skindlov 1996; Stensrud and Yussouf 2003; Woodcock and Engel 2005). For ease of discussion we also refer to the running mean error as a bias.

In the following, superscript t denotes the unobservable true state, superscript f denotes a forecast, and superscript a denotes an analysis. We can write discrete representations as vectors, and the forecast valid at time $k$ is then the truth with error, $x_k^f = x_k^t - e_k^f$, where the length of these vectors is $n$ grid points. In this work we define the $n$-dimensional vector of true forecast (predicted) bias at gridpoint locations as the running-mean forecast error at time $k$, $b_k = \tilde{x}_k^t - x_k^t$. An estimate of the predicted bias, $b_k^f$, is subject to error $e_k^b$.

$$b_k^f = b_k + e_k^b = \tilde{x}_k^t - x_k^t + e_k^b = \overline{e}_k + e_k^b$$

where the overbar represents the mean over the most recent $K$ days at the same local time of day.
Observations at time $k$ are assembled in a $p$-length observation vector, $y_k$, with observation error $e_k$. That is, the true atmospheric state at points collocated with the observations is $y_k = y_k^o - e_k$. A linear forward operator $H$ relates the gridded state to the observations, and $Hx_k = y_k$ as long as $H$ is perfect, as the bilinear interpolation represented in $H$ is assumed here. A forecast error at the observed locations can thus be written

$$Hx_k^f = Hx_k - Hx_k^o = y_k^o - e_k - Hx_k^f. \quad (3)$$

We can now define the running-mean error at observed locations, which are considered observations of bias:

$$\beta_k^o = HE_k = \left( y_k^o - Hx_k^o \right), \quad (4)$$

where we have assumed that the observations $y_k^o$ are unbiased. Because we define bias observations to be the running-mean error at the observed sites, $e_k^o = 0$; observations of bias are error free. The definition could be relaxed to allow for error, but this type of error is extremely difficult to define. The observed bias defined in Eq. (4) evolves slowly compared to the innovation $(y_k^o - Hx_k^o)$, with the difference in time scale determined by the averaging sample length $K$.

The definitions above enable a restatement of the goal: to estimate a bias analysis $b_k^o$ from $\beta_k^o$, and then to predict a bias correction for the grid with $b_k^f$. When used for bias correction at observed sites, the propagator from $\beta_k^o$ to a forecast bias $b_k^f$ is usually assumed to be the identity matrix such that the dynamics of the bias are simply persistence. Here, we make the same assumption so that the forecast bias on the grid is the bias analysis from 1 day prior, $b_k^f = b_k^o$.

The error covariance of the bias prediction is

$$P_k^b = \langle e_k^b (e_k^b)^T \rangle, \quad (5)$$

where the angle brackets denote the expectation. Given observations of the running-mean error, $\beta_k^o$, and the error covariances for $b_k^o$, the minimum-variance estimate for $b_k^o$ can be written analogously to the minimum-variance state estimate in the data assimilation problem:

$$\begin{align*}
b_k^o &= b_k^o + K_k(\beta_k^o - Hb_k^o) & \quad (6)
K_k &= P_k^b H^T [HP_k^b H^T + R_k]^{-1}, & \quad (7)
\end{align*}$$

where the same assumptions about linearity and Gaussian statistics have also been made. The matrix $R_k$ is the observed bias error covariance, which is zero here because of the definition in Eq. (4). Assuming that the bias estimate itself is unbiased such that $\langle e_k^b \rangle = 0$, the error covariance of the bias estimate can be approximated as follows:

$$\begin{align*}
P_k^b &= \langle (b_k^o - b_k^f)(b_k^o - b_k^f)^T \rangle \\
\approx & \langle (x_k^o - \langle x_k^o \rangle)(x_k^o - \langle x_k^o \rangle)^T \rangle. \quad (8)
\end{align*}$$

Equation (8) also uses the assumption that $x_k^o - \langle x_k^o \rangle \ll x_k^f - \langle x_k^f \rangle$. This assumption leads to error in $P_k^b$ that is difficult to quantify.

The primary motivation for the approximation in Eq. (8) is that $P_k^b$ is not available given a biased model for $x_k^o$ (cf. Dee and Da Silva 1998), but $x_k^o$ provides information about the regime and lower-boundary forcing (topographic and otherwise). A model for $P_k^b$ can instead be built from observed biases. Unless such a model is anisotropic and nonstationary, it generally does not contain any information about regimes or heterogeneous lower boundary forcings.

Equations (6)–(8) provide a suboptimal method to estimate $b_k^o$, exploiting linear relationships between $Hx_k^o$ and $x_k^o$. Anderson (2003) clarified the statistical linearization of the analysis equation, showing that data assimilation can be cast in terms of linear regression. Here, the regression is between $Hx_k^o$ and $x_k^o$, and those coefficients are used to relate $\beta_k^o$ and $b_k^o$. The regression coefficients are imperfect primarily because of the assumptions leading to the approximation in Eq. (8), but they can be computed with high confidence if the ridge regression technique is used.

The absence of bias observation error, giving $R_k = 0$, results in a regression involving a poorly conditioned matrix $HP_k^b H^T$ in Eq. (7). This can be handled in any number of ways, and we choose a ridge regression technique for its simplicity. The diagonal of $HP_k^b H^T$ is increased by an additive ridge factor of 0.01, leading to statistically biased, but numerically accurate, estimates of the regression coefficients. The effect is equivalent to using an observation error variance of 0.01 and assuming that observation errors are uncorrelated, and the observation increments are damped by a factor of $1/(1 + 0.01) \approx 0.99$.

The biases defined in (2)–(4) are not at all equivalent to the formal definition of bias, and the interpretation of $e_k^b$ is not unique. If one’s goal is to estimate the expected climate bias from the running-mean error, then $e_k^b$ is strictly the uncertainty in that estimate, including sampling error. If one’s goal is to estimate today’s bias from the running-mean error, then $e_k^b$ combines the effects of an imperfect bias propagation model and contributions from using possibly irrelevant flow scenarios within the $K$-length sample. Regardless of the interpretation, it enables an estimate of uncertainty that can be exploited to analyze the running-mean error on a grid.
c. Sampling lengths and error covariances

It remains to construct a sample from which to draw realizations of $\mathbf{x}_f^k$ and estimate the expectation $\mathbb{E}[\mathbf{x}_f^k]$. Without a good ensemble, we rely on a recent history of running-mean forecasts, $\overline{\mathbf{x}}_f^k$. Although this does not provide the correct distribution for the most recent running mean, it is an improvement over a recent history of individual forecasts, $\mathbf{x}_f^k$, because the time scale of the running mean is much slower.

We seek a $K$ that is as short as possible to maximize the responsiveness to regime changes, but one that is long enough to give a meaningful average and result in a useful error correction at observed sites. Without a guiding theory, we adopt a trial and error approach to find the averaging length. For the experiment period, running-mean errors are computed with averaging lengths varying from 1 to 30 days, and those error estimates are used to predict and correct the subsequent forecast. The resulting RMSE is compared against the RMSE computed from the uncorrected forecast innovations.

Within the given experiment period, longer averaging times result in a smaller sample both because the first $K$ days cannot be corrected, and also because of an increased likelihood of a missing observation. In this particular analysis, running means are not computed when missing observations create gaps in the time series, giving the most accurate measure of the potential for running-mean error corrections. To eliminate the effects of sampling error, the same sample size is used for each averaging length. Random subsamples are chosen from the total available sample of predicted error at each averaging length. Results from 100 different random subsamples of size 100 are averaged. A subsample of 100 ensures at least some independence remains between each subsample at longer averaging lengths. This exercise is carried out independently for each 3-hourly valid time of day, and the results are averaged to get one estimate.

Results from this analysis suggest that an averaging length of $K = 7$ days is appropriate for these experiments. Figure 3 shows the percent change in RMSE resulting from a running-mean error correction with averaging lengths $1 \leq K \leq 30$ days. Using the daily error to predict the error the following day, corresponding to the trivial averaging length of one, is not effective at reducing the error, presumably because of randomness in the daily innovations. Averaging lengths $1 < K < 7$ days are increasingly effective, and relatively less further benefit is evident for $K \geq 7$ days. Lengths $K \geq 26$ days show decreasing effectiveness, possibly because seasonal trends are beginning to play a role. To preserve the most potential for adaptation to regime changes, and also to realize most of the potential benefit in the gridded estimates, we choose $K = 7$ days. The resulting 1-day lagged autocorrelations are shown in Fig. 4. The high values indicate the general applicability of a diurnal persistence model to predict the error. Observational records at both stations HMN and S28 have many gaps, which may be the cause of the lower values there.

The sample size for covariance estimates $I$ is chosen to be long enough that the covariance between $\mathbf{Hx}_f^k$ and $\overline{\mathbf{x}}_f^k$ can be confidently estimated. Similar to the averaging length $K$, shorter sampling lengths are desirable because they are more relevant to the current regime.
but estimates obtained from longer periods are subject
to less sampling error. For the averaging length $K = 7$
days, $I = 30$ days displays reasonable statistical stabil-
ity. This was analyzed by regressing $\mathbf{Hx}_k$ against $\mathbf{x}_k$ for
sampling lengths $3 \leq I \leq 40$ days and noting that the
regressions typically stabilized by $I = 30$ days.

Despite high confidence in the relationship between
$\mathbf{x}_k$ and $\mathbf{Hx}_k$, early experiments suggested that using $K = 7$
and $I = 30$ was subject to some sampling error, con-
taminating $\mathbf{P}_k^b$ after several cycles. Qualitative exami-
nation of the $\mathbf{b}_k^f$ fields in regions farthest from all ob-
servations suggested this behavior. The effect is similar
to the filter divergence observed in ensemble data as-
similation systems, which has been shown to be miti-
gated via a localization operator (e.g., Hamill et al.
2001). The intent is to throw away covariance estimates
that are spatially distant from the observation sites and,
thus, subject to contamination from sampling error as
the estimates repeat. We adopt the same strategy, em-
ploying a Schür product of the estimated error covari-
ance matrix $\mathbf{P}_k^b$ and an isotropic correlation function
expressed in an $n \times n$ correlation matrix $\mathbf{C}$. The func-
tion is specified by Eq. (4.10) in Gaspari and Cohn
(1999), and is chosen because of its compact support,
mathematical properties including positive definite-
ness, and widespread use for data assimilation. The
half-width of the localization function is found by mini-
mizing the RMS difference between the correlation
matrix computed from the full sample of differences
$\mathbf{x}_k^f - \mathbf{x}^f_{k-1}$, and $\mathbf{C}$, over the entire sample period. The
result is approximately 350 km for temperature, and the
result is likely influenced by the domain size. The lo-
calization function will serve to reduce error increments
primarily near the edges of the domain. Effects of the
localization are further addressed in section 4.

d. Diagnosis of contaminating observing stations

Just as in data assimilation, where a lack of observa-
tions in specific locations can lead to large errors in
state estimates, a lack of observations of systematic er-ror can lead to large errors in gridded bias estimates.
We lack an effective propagator to temporally spread
observed systematic errors into unobserved regions. An
observing station displaying large errors that are not
well correlated with the surrounding areas also has the
potential to degrade nearby error estimates. Such a
situation could arise when presented with a station that
is in a deep canyon, with few surrounding stations.
Error can be corrected at that observing station, but a lack
of nearby observing stations prevents verification that
nearby estimates are useful. Here, we show that simple
error diagnostics at observing stations may lead one to
ignore certain observing locations in the gridded error
estimation process, or understand why the results of
estimating at uncorrelated locations can produce mis-
leading summary statistics.

A combination of large errors and weak correlation
with other stations can be diagnosed with statistics from
the entire experiment period. Temperature forecast er-
ror correlations in observation space are easily com-
puted from the innovations $(\mathbf{y}_k^f - \mathbf{Hx}_k)$, and the mag-
nitude of innovation correlation coefficients between
two stations is shown in Fig. 5. The stations are arranged
in the plots such that stations located in mountainous
regions are placed below and to the right of the thick solid lines, and valley and foothill stations are placed to the top and left. The warmer
colors, indicating high correlations, are primarily be-
tween valley and foothill stations. With the exception of
S05, temperature errors at the mountain stations are
not well correlated with those at other stations, includ-
ing other mountain stations. If those stations were to be
withheld from the estimation process, linear bias esti-
mates at those locations would not be possible from the
other observations in this domain.

Temperature errors at stations S04, S30, and S31 are
well correlated with each other, but are not generally
well correlated with errors at the remaining stations.
These three are relatively close together, but subject to
unique topographic features. S04 is adjacent to the
White Sands National Monument, while S30 and S31
are located near the mouths of zonally oriented valleys.
Their proximity, and possibly their similar location in
the lee of the San Andres Mountains, appear to result
in high error correlations. Other possible influences
such as small variations in the orography, or large varia-
tions in soil properties, appear to play a smaller role
here.

The computations summarized in Fig. 5 can be used
in two ways. One possible approach is to ignore the
uncorrelated stations in the estimation process, which
in this case would lead to the elimination of most of the
mountain stations. This approach may have the unde-
sirable affect of also eliminating any useful error esti-
mates in the mountainous regions nearby. A second
possible approach is to retain these stations in the es-
timation process, and acknowledge that estimates mak-
ing use of these stations cannot be completely verified
except at the observing sites themselves. Error esti-
mates at those locations when they are withheld from
the estimation process, but included in the verification
(such as might be observed in a cross validation), are
guaranteed to be poor. Additionally, the systematic er-
rors at S09 and S17 is large, and if errors are not accu-
rately estimated there, they may be outliers in the dis-
tributions used for verification. Summary statistics such as RMSE are not resistant to outliers (e.g., Wilks 1995). Once identified as outliers, results there can be withheld from the computation of summary statistics, leading to a fairer picture of the system performance. For completeness, we keep the stations in the estimation process, and quantify their overall contribution to the summary metrics by excluding them sometimes.

e. Practical application

This section summarizes the error estimation algorithm. In section 3 we will compare the results from two approaches, differing in the specification of \( P_b^k \). The running-mean covariance-corrected (RMCC) approach uses Eq. (8), and the isotropic covariance-corrected (ICC) approach uses a stationary covariance model tuned on the observed biases \( \beta^k \).

The same class of correlation functions used for localization in RMCC is used for \( P_b^k \) in the ICC experiments, but with a different length scale. The variance is simply the mean of the temporal variance in \( \beta^k \). The covariance in \( \beta^k \) is computed as a function of distance, and the half-width of the isotropic function is chosen by fitting it to the covariances in one dimension, given the variance at a distance of zero. This assumes that the covariance in \( \beta^k \) is related to the error covariance in \( \beta^k \). The result is a half-width of approximately 38 km for temperature. We note that the scatter in \( \beta^k \) as a function of distance is appreciable, and the fit is made with low confidence, but it is the best we can do given the data. The covariance matrix constructed in this way can also be poorly conditioned for large length scales of the isotropic function in \( P_b^k \). Although this may not be a problem in our specific case, we retain the ridge regression technique for consistency with RMCC. The basic experiments are summarized for reference in Table 1, and variations on some of the parameters are explored later in section 4.

The implementation of each approach has advantages and disadvantages. The primary advantage of ICC is the economy. The obvious disadvantages are that the covariance model is isotropic and stationary in time and space. The advantage of RMCC is that systematic information about the typical recent flow is in-

Fig. 5. Magnitude of 6-h \( T_2 \) forecast error correlations between all possible pairs of stations, valid at 1800 UTC (1100 LST). Mountain stations are below and to the right of the thick solid lines.
The disadvantage is that $\hat{P}^b_k$ contains some residual systematic error because unbiased covariances cannot be estimated from a biased model.

At least $I + K$ forecasts are needed to compute $\hat{P}^b_k$ in RMCC. Using one forecast per day, valid at the same local time, $I = 30$ running means of length $K = 7$ days are computed to produce samples of vectors $\vec{x}_k$, $\vec{y}_k$, and $\vec{y}_k$. From day $I + K$, the algorithm proceeds as follows:

1) The systematic error $\beta^o_{k-1}$ is propagated to $\beta^o_k$ with the diurnal persistence model.
2) Vectors $\vec{x}_k$ and $\vec{y}_k$ are computed from the most recent $K$ forecasts and observations, respectively.
3) Matrix $\hat{P}^b_k$ is computed from the most recent sample of $I$ running means, using Eq. (8).
4) Equation (6) is used to estimate $\beta^o_k$.

A persistence model is used because of the slow temporal variation in $\beta^o_k$. Alternately, uncertainty in the evolution could be included by fitting an autoregressive (AR) model to the time series of $\beta^o_k$ and using it to predict $\beta^o_k$. The slow variation in $\beta^o_k$ means that the parameter specifying the random component of an AR(1) model is small in this case. Preliminary experimentation showed that the results are largely insensitive to the random uncertainty, which we set to zero. The next two sections explore the results and sensitivities.

f. An example with a single observation

Here, we present an example using the two covariance models, ICC and RMCC, and a single observation $\vec{y}_k$ at S32. Localization is applied to the RMCC covariances. The resulting increment $\beta^o_k - \beta^o_k$ is shown in Fig. 6, which is valid at 1800 UTC 30 March 2005.

The difference between approaches ICC (Fig. 6a) and RMCC (Fig. 6b) is clear in the increments. In the isotropic case, the correlation length scale tuned on errors at all observing stations perhaps reflects the scale of the topography, but this instance suggests the possibility of imposing large increments that cross topographic barriers. The increment imposed by RMCC is not perfectly tied to the topography, but in this case appears to respect the barriers. The possibility of nonzero increments at longer distances is also apparent. The correlation length scale is longer in RMCC because the gridded model fields are more highly correlated in space than are the innovations.

3. Gridded error estimates

This section presents a verification of $\beta^o_k$ estimates from approaches ICC and RMCC, both where observations are available and where they are not. All results are presented for 6-h forecasts of $T_2$ valid at 1800 UTC,
except where noted. The agreement between $\mathbf{Hb}_k$ and $\beta_k$ demonstrates the fit to the observed running-mean error estimates, and the total error reduction shows the benefit of a filtered running mean in predicting the errors. Because the goal is to estimate and predict the systematic error where observations do not exist, cross-validation experiments are constructed. Each station is withheld from the estimation process in separate experiments, and the error at the withheld station is measured. Statistics are then compiled at the withheld stations, measuring the skill in the bias estimation where observations are not available.

The estimated (predicted) systematic error ($\beta_k^*$) fit to the observed systematic error ($\beta_k$) is computed as a function of time $t_k$:

$$
\text{RMSE}_{\beta_k^*} = \left[ \frac{1}{p} (\beta_k^* - \mathbf{Hb}_k^*)^T (\beta_k^* - \mathbf{Hb}_k^*) \right]^{(1/2)},
$$

The results are shown in Fig. 7 for ICC (dashed) and RMCC (solid). For both covariance estimates, the predicted bias closely fits the observed bias within the first few days. Differences are small because the fit is largely determined by the ridge regression factor. Both results show periods of weaker fit, with the end of May being especially notable. These jumps are caused by regime changes in the temperature, but the large gap in observations may have contributed. Time series of gridded estimates are filtered by the estimation process and, thus, do not react as quickly to regime changes as do the 7-day running-mean errors $\beta_k$. This does not necessarily imply a poorer prediction of error, because $\beta_k$ can be strongly affected by a single outlier in the recent history. Rather, the value of systematic error prediction should be ascertained in the context of total forecast error.

The goal is to reduce the daily forecast error, and the remainder of the paper will focus on evaluating forecast innovations. Before correction for the predicted bias, the innovations are $y_k^* - \mathbf{Hx}_k^*$; after correction for the predicted bias, they are $y_k^* - \mathbf{Hx}_k^* - \mathbf{Hb}_k^*$. Typical RMSE statistics can be computed on these quantities, either daily to produce a time series of errors or by considering the entire experiment in one sample.

Histograms of the forecast innovations before and after correction with the predicted systematic error show a reduction in error for both ICC and RMCC (Figs. 8 and 9). Figures 8a and 9a show histograms of the raw innovations, without any correction for the predicted error. Figures 8b and 9b show histograms of innovations when the forecast is adjusted for the predicted error on the grid, $\beta_k$, and the stations at which the innovations are evaluated are used to estimate the gridded bias. Figures 8c and 9c show innovations accumulated over each withholding experiment, where the forecast at a site not used in the estimation is adjusted by the error predicted from all other stations. Figures 8d and 9d are the same as Figs. 8c and 9c, but innovations at stations S09 and S17 are not included in the sample. Standard deviations of the innovations are also reported inset.

Figures 8 and 9 show that both ICC and RMCC effectively reduce errors in the forecasts. When all sta-
tions are included (Figs. 8b and 9b), the standard de-
viation of the errors is reduced by approximately 42%,
and neither method shows a clear advantage over the
other. Given the good fit to $\beta_k$ shown in Fig. 7, this
suggests that $\beta_k$ is a reasonable predictor of systematic
error.

Innovations at withheld stations also show an error
reduction (Figs. 8c and 9c). The peaks in the histograms
are much lower than at the included stations, and the
distributions are much broader. But the standard de-
viation of the innovations are still reduced by about
9% and 12% for ICC (Fig. 8c) and RMCC (Fig. 9c),
respectively. These results suggest a slight advantage to
RMCC, but both are affected by the outliers visible in
the left tails of the distributions.

The left tails of the distributions in Figs. 8a and 8c
and 9a and 9c are innovations almost exclusively at
stations S09 and S17, and the tail disappears when those
are removed (Figs. 8d and 9d). Errors at S09 and S17
are successfully reduced when they are included in the
estimation process (Figs. 8b and 9b), but are not well
estimated when they are withheld from the estimation
process. This is an expected result given the generally
uncorrelated errors at those stations shown in Fig. 5.
The systematic errors are approximately 6.3°C and 4.5°C
at stations S09 and S17, respectively, as estimated from
Fig. 2. But the estimated errors are order 0.1°C at those
locations when withheld. Virtually no correction is ap-
plicated there, and the corrected temperatures are nearly
the same as the uncorrected ones. Thus, the tails in the
distributions in Figs. 8c and 9c result from essentially
uncorrected innovations, which are outliers in the dis-
tributions that can dominate summary statistics. We
note that comparison of the standard deviations in Figs.
8d and 8a and 9d and 9a is misleading because the distri-
butions in Figs. 8a and 9a include innovations at stations S09 and S17, and the correct distribution for
the comparison is not shown here.

A summary metric useful for evaluating the overall
effects on forecast error is the RMSE across observing
stations, as a function of prediction day. This reveals
any potential for occasionally degrading the forecast
verification when attempting to account for systematic
errors. Figures 10 and 11 shows that errors are reduced
overall, but regime changes or missing data can pro-
duce occasional busts. Stations S09 and S17 are not
included in these computations because they are rein-
introduced in the next section to better evaluate their
impacts.

The RMSE as a function of day for experiments ICC
and RMCC is shown in Figs. 10 and 11, respectively.
Except that stations S09 and S17 are not included, the
dotted curves correspond to the distributions of uncor-
rected forecast innovations in Figs. 8a and 9a. The
dashed curves, showing the least overall error, corre-
spond approximately to the distributions of innovations
corrected with error predictions in Figs. 8b and 9b. The
overall results of the cross validation are shown with the
solid curves, and correspond exactly to Figs. 8d and 9d.

Figures 10 and 11 show the time dependence of the
error estimates. The difference between the dotted and
solid curves shows the effects of imperfect error cova-
riance models in both ICC and RMCC. The solid curve
lies in between the other two curves and shows less
variability. Error estimates at unobserved sites are es-
sentially damped, and errors in the covariance models
generally do not result in an overall negative effect.
Note that if stations S09 and S17 were included, all
three curves would move upward, and the solid curve
would lie closer to the dotted curve, but their relative
ranking would remain intact.

To summarize, both the ICC and RMCC covariance
models are effective. It appears that the dynamic error
covariances used in RMCC slightly outperform the iso-
tropic covariances used in ICC. In the next section, we
consider an example of the diurnal evolution of error
estimates, explore the sensitivity to error covariances,
and discuss wind speed errors.

4. Additional considerations
a. Diurnal variation of systematic error

Near-surface errors exhibit a diurnal cycle, motivat-
ing the prediction of systematic error based on the er-
rors at a single time of day. The results presented in the last section are all valid at 1800 UTC (1100 LST). Without further parameter tuning, experiments at several other times of day were also completed, with mostly positive results. Here, we present an example of the systematic error estimate as it evolves through the diurnal cycle.

Figure 12 shows the $T_2$ systematic error at four times of day on an arbitrarily chosen day (29 June 2005). The error fields change in both shape and magnitude throughout the day, and in many places the error changes sign. For example, consider that the location marked with the X is over the western slopes of the Sacramento Mountains and north of station S17. At 0300 UTC (2000 LST), the forecasts are estimated to be too cold by approximately $-4^\circ$C. In the early morning (0900 UTC, 0200 LST) the estimated error is nearly the same, but the nearby gradient has steepened with a cold–warm dipole across the X. After sunrise (1500 UTC, 0800 LST) the bias changes sign indicating a forecast that is $4^\circ$C too warm. Large magnitudes near the boundaries of the domain indicate that different tuning may be needed for this time of day. By midafternoon (2100 UTC, 1400 LST), the cool bias has returned. This type of transition may be difficult to follow if the diurnal cycle was not removed from the estimation procedure.

The fields in Fig. 12 also show that the errors are not strictly tied to elevation. Rather, differential heating depending on the aspect of the slope may contribute. An example is the west–east bias gradient across the San Andres Mountains at 2100 UTC. The sunny western slope is biased warm, while the Tularosa Basin to the east remains too cold. This suggests that bias estimates based solely on elevation are inappropriate. The different structures evident in the estimated error also suggest that independent tuning at each time of day will be necessary to capture the diurnal behavior of error.

Bias estimates in Fig. 12 are only examples and the parameters used in the process were tuned on forecasts

![Fig. 10. RMSE of $T_2$ across reporting stations at 1800 UTC for ICC. The dotted curve shows the results when no correction is applied, the dashed curve shows the results when correcting for the predicted error and including all stations in the estimates, and the solid curve shows the results when correcting for the predicted error at every station that is withheld.](image1)

![Fig. 11. RMSE of $T_2$ across reporting stations at 1800 UTC for RMCC. The dotted curve shows the results when no correction is applied, the dashed curve shows the results when correcting for the predicted error and including all stations in the estimates, and the solid curve shows the results when correcting for the predicted error at every station that is withheld.](image2)
valid at 1800 UTC. The stations chosen to omit from the estimation, localization radii, and sampling lengths may vary. Large biases estimated near the corners of the domain at 1500 UTC suggest that the regressions may not be accurate because of improper tuning. A sense of some of these sensitivities can be gained from the experiments reported next.

b. Sensitivities

As with any practical estimation algorithm, some tuning based on measured errors and model characteristics is unavoidable. Here, we report sensitivities to each of the chosen or tuned parameters. We denote the “baseline” experiments examined in section 3 as RMCC A and ICC A, and use subsequent letters to denote different sensitivity experiments. For RMCC these are B for the ridge factor in the regression; C for the number of running means in the sample to compute covariances; I; D for the effect of the covariance localization; and E for the effect of withholding stations S09 and S17 from the estimation process. For ICC the sensitivities tested are the same in B and D, and localization is not an issue. The length scale of the covariance model can be changed, and a longer length scale is tested (ICC C). Because the most difficult test of the gridded error estimation algorithm is constructed by withholding data, the cross-validation experiments are repeated with these configurations. Sensitivity experi-
result because the standard deviation of the histograms does not include the systematic component of RMSE.

The RMCC approach demonstrates an advantage under all circumstances tested here, although in some cases the advantage is small. Baseline experiments A show a 3% difference between RMCC and ICC, whether or not S09 and S17 are included. ICC C, which used a longer correlation length scale derived from the model instead of the observed errors, results in larger forecast errors. RMCC C, which uses a smaller sample in the covariance computations, shows the negative effect of increased sampling error. Tighter localization could potentially correct this, and further experimentation is needed.

Outliers resulting from almost no correction being applied at S09 and S17 dominate all of the results in the middle column of Table 3. We omit the relative improvements for ICC D and RMCC E in the middle column because uncorrected results that exclude S09 and S17 in the estimation process are not shown. In the third column the uncorrected error is lower, and the relative improvement is greater than in the middle column. These results highlight the importance of identifying outliers in samples used for summary statistics that are not resistant.

Differences in the third column between ICC A and ICC D, and RMCC A and RMCC E, measure the effects of including S09 and S17 in the estimation process. Differences of approximately 3% (ICC) and 2% (RMCC) are much smaller than the effect of simply excluding those stations from the summary statistics. This suggests that it is beneficial to include those stations in the estimation, even though a lack of correlated stations prevents a true cross validation.

Increasing the ridge regression factor also has the effect of reducing the positive impact of the corrections (ICC B and RMCC B). This is computationally equivalent to increasing the error variance in $\beta_k^o$ under the assumption that those errors are uncorrelated in time and space. Ridge regression introduces an estimation error because the gridded estimates $\mathbf{b}_k^o$ are less constrained by $\beta_k^o$. These results suggest that a smaller value, namely 0.01, is superior, and constraining the estimates more strongly by $\beta_k^o$ results in better estimates at unobserved locations.

To ensure that these results are not dominated by the dense group of stations in the Tularosa Basin, stations are divided into two groups: 1) those outside of the basin that are also retained in the analysis (LRU, S05, S06, S08, S19, SRR, and TCS), and 2) the remaining stations. Here, we include stations S09 and S17 in the estimation process, but do not include them in either result because they constitute approximately 12% of

---

Table 2. Sensitivity experiment descriptions.

<table>
<thead>
<tr>
<th>Expt</th>
<th>$P^o_k$</th>
<th>Ridge factor</th>
<th>$C$ (km)</th>
<th>$I$</th>
<th>All obs?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICC A</td>
<td>38 km</td>
<td>0.01</td>
<td>—</td>
<td>—</td>
<td>Yes</td>
</tr>
<tr>
<td>ICC B</td>
<td>38 km</td>
<td>0.10</td>
<td>—</td>
<td>—</td>
<td>Yes</td>
</tr>
<tr>
<td>ICC C</td>
<td>350 km</td>
<td>0.01</td>
<td>—</td>
<td>—</td>
<td>Yes</td>
</tr>
<tr>
<td>ICC D</td>
<td>38 km</td>
<td>0.01</td>
<td>—</td>
<td>—</td>
<td>No</td>
</tr>
<tr>
<td>RMCC A</td>
<td>Eq. (8)</td>
<td>0.01</td>
<td>350</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>RMCC B</td>
<td>Eq. (8)</td>
<td>0.10</td>
<td>350</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>RMCC C</td>
<td>Eq. (8)</td>
<td>0.01</td>
<td>350</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>RMCC D</td>
<td>Eq. (8)</td>
<td>0.01</td>
<td>—</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>RMCC E</td>
<td>Eq. (8)</td>
<td>0.01</td>
<td>—</td>
<td>30</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 3. Sensitivity experiment showing the total RMSE in forecast $T_o$ (°C) at withheld stations. Column RMSE* reports the results that exclude stations S09 and S17 from the RMSE calculations. The relative improvement over the uncorrected forecasts is shown following the slashes.

<table>
<thead>
<tr>
<th>Expt</th>
<th>RMSE</th>
<th>RMSE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncorrected</td>
<td>2.95/0%</td>
<td>2.34/0%</td>
</tr>
<tr>
<td>ICC A</td>
<td>2.71/8%</td>
<td>2.03/13%</td>
</tr>
<tr>
<td>ICC B</td>
<td>2.80/5%</td>
<td>2.21/6%</td>
</tr>
<tr>
<td>ICC C</td>
<td>3.07/-4%</td>
<td>2.50/-7%</td>
</tr>
<tr>
<td>ICC D</td>
<td>1.97/—</td>
<td>1.97/16%</td>
</tr>
<tr>
<td>RMCC A</td>
<td>2.61/11%</td>
<td>1.96/16%</td>
</tr>
<tr>
<td>RMCC B</td>
<td>2.78/6%</td>
<td>2.11/10%</td>
</tr>
<tr>
<td>RMCC C</td>
<td>3.14/-7%</td>
<td>2.48/-6%</td>
</tr>
<tr>
<td>RMCC D</td>
<td>2.88/2%</td>
<td>2.11/10%</td>
</tr>
<tr>
<td>RMCC E</td>
<td>1.91/—</td>
<td>1.91/18%</td>
</tr>
</tbody>
</table>
the sample outside of the basin and can dominate even more than in the sample containing all stations. The RMSEs for ICC are $2.00^\circ$ and $2.08^\circ$ inside and outside the basin, respectively, and the results for RMCC are $1.97^\circ$ and $1.95^\circ$. These values can be compared to the results for ICC A and RMCC A in the third column of Table 3. The difference from inside to outside the basin is small compared to the sensitivity of other parameters.

Although exploration of the complete parameter space is intractable, these results suggest that experiments ICC and RMCC represent a nearly maximum benefit of this type of systematic error estimation for this observation network and forecast system. Sensitivities show that including S09 and S17 in the estimation process may be beneficial, but that including them in the summary statistics skews the results.

c. Systematic errors in wind speed

This work focused on $T_2$ for systematic error correction because, as shown in Fig. 2, the error is dominated by the systematic component. This may also be the case for specific humidity $Q_2$, but we were unable to complete the analysis because the archived data were not available. Wind speed errors in this forecast system, and over this domain, are dominated by the random component (Fig. 13). Thus, it would not be fruitful to estimate and apply systematic error corrections with the approaches proposed in this paper. Experiments have confirmed this, showing almost no impact on the overall error statistics in the cross-validation experiments.

5. Summary and discussion

This report explores the potential of a suboptimal filter for estimating systematic error on near-surface mesoscale grids when a good ensemble is not available. Systematic error is defined here as the 7-day running-mean error, measured at available observing stations. This definition is not a mathematically rigorous definition of bias, but for convenience we have referred to the error as bias. The estimation approach relies on solving the statistical-analysis equation for systematic error, and error covariances must then be specified. Two classes of background error covariance estimates are considered: 1) a stationary, isotropic function tuned with observed running-mean errors and 2) dynamic error covariances estimated from a recent history of running-mean forecasts. Systematic error prediction is achieved via a diurnal persistence model, where the error is expected to be the same as the observed 7-day running-mean error valid 1 day prior. Error predictions are tested with an operational forecasting system located over New Mexico, in complex terrain, and spanning multiple seasons. We report results valid at 1800 UTC during the experiment period.

The primary findings of this work are as follows:

1) Errors in 2-m temperature at many locations are dominated by the systematic component (Fig. 2), while errors in wind speed are dominated by the random component (Fig. 13). It is sensible to estimate systematic errors in $T_2$.

2) Both error covariance models reasonably fit the observed running-mean errors (Fig. 7).

3) Both error covariance models result in systematic error predictions that produce an overall RMSE reduction of near 42% at stations where the running-mean error is observed (Figs. 8b and 9b), suggesting that the running-mean error is a reasonable predictor of systematic error at observed locations.

4) Cross-validation experiments show RMSE reductions of approximately 9% and 12% for the isotropic covariance model and the dynamic covariance estimates, respectively, at stations that do not contribute to the bias estimate (Figs. 8c and 9c). This suggests a slight advantage to the dynamic covariance estimates.

5) The corresponding improvements are 13% and 16% when results from stations S09 and S17 are excluded from the summary computations (Table 3, third column ICC A and RMCC A). Results at stations S09 and S17 are outliers, and RMSE is not resistant to them. This result was predicted by computing error correlations in observation space (Fig. 5). A slight further improvement is evident when S09 and S17 are not used in the estimation procedure.

6) The estimated errors evolve with the diurnal cycle, and are not strictly functions of elevation (Fig. 12).
7) Use of a very long length scale in the isotropic error covariance model significantly degrades the results (Table 3, ICC C).
8) Use of a much shorter sampling length, thereby increasing sampling error in the dynamic covariance estimates, significantly degrades the results (Table 3, RMCC C).

These results are all subject to the details of both the forecasting system and the observing network, and any new implementation will require a different specification of the covariances and the localization. This exercise is unavoidable in any case, and should not be viewed as a fundamental barrier. The dynamic approach proposed here is an attempt to estimate recent error, rather than climatological error estimated from long records, although we used a longer record in tuning the correlation and localization functions. Six months of observations and forecasts was sufficient to specify observation error variances and the isotropic correlation function.

The small domain tested here is characterized by few observing locations, and the samples of errors from them are not distributed normally. A few stations show much larger errors than the rest. At times we removed outliers to get fairer estimates of the summary statistics, but also showed the effect of including them. A larger domain with many more observing locations may provide error samples that are less sensitive to outliers.

This work uses a 7-day running-mean error for a definition of systematic error, and the total error reduction is limited by this definition. The gridded estimates are able to fit the observed running-mean error well, and resulted in a positive impact in most places where observations are not available. Presumably, better estimates of the systematic error at the observation locations will result in better gridded estimates as well. One might also expect that an observing network characterized by more uniformly spaced, denser, observations will also result in further improvements.

Although the dynamic approach appears slightly more successful in general, for some implementations choosing between the isotropic function and the running-mean covariance computation may be more a matter of computational and storage capability than actual performance. The computational cost of dynamically computing the error covariance matrices is much greater than simply tuning an isotropic function once, but it is far from prohibitive. Without any algorithmic optimization (i.e., we explicitly computed $\hat{P}_k$), each of these experiments were run on a workstation in a matter of minutes.

More work is required before using this approach in a cycling data assimilation system, which typically assumes a bias-free background forecast. Correcting only the near-surface model layers will change the 3D background error covariance and potentially introduce unphysical modes.

Finally, we have restricted our attention to surface temperature grids because of the dominance of systematic error, and the persistence of error across a diurnal cycle. It is not clear whether this approach will prove useful aloft for two reasons: 1) the systematic error component may not be as large and 2) the persistence model for systematic error propagation may not be appropriate. In the limit of the climate-mean error, the systematic component may be large enough to warrant estimation, and no model for propagation is required. But to estimate flow-dependent systematic error, an observed error distribution conditioned on the “flow of the day” is needed. Such a distribution is difficult to obtain without a good ensemble.

Acknowledgments. This research was supported by the U.S. Army Test and Evaluation Command through an Interagency Agreement with the National Science Foundation. Discussions with, and comments by, C. Snyder were useful in the development of this manuscript. Reviews by T. Hamill and two anonymous reviewers were valuable in strengthening the paper.

REFERENCES


