P2.18 TOWARD A SURFACE DATA CONTINUUM: USE OF THE KALMAN FILTER TO CREATE A CONTINUOUS, QUALITY CONTROLLED SURFACE DATA SET

John A. McGinley
NOAA Research - Forecast Systems Laboratory
Boulder, CO

1. INTRODUCTION

As operational weather support moves in to the new century, we are seeing a vast increase in the number of available observations based on the maturation of internet and web and the expansion of weather observing networks for an ever-increasingly weather-oriented user community.

The modernization of the US National Weather Service has created great opportunities for expanding the amount of surface data for local operational forecasting. Utilizing a system called the Local Data Acquisition and Dissemination System (LDAD: Jesuroga, et al.,1998) the porting of a wide variety of local surface (and in some cases upper air) data is unprecedented. These data originate from state highway departments, agricultural networks, private industry, and schools.

Some weather support functions have dense networks of surface and upper air data aimed at a specific mission (for example, space launch support). These facilities require frequent monitoring of weather conditions mandating a high frequency product cycle. This requires that immense numbers of observations be processed for quality, consistency, and timeliness over small time intervals. In these environments data from many diverse sources have to be amalgamated and time sequenced so that products being developed from the data suite are appropriate for a specific time. When analyzed meteorological products are needed at intervals of a few minutes it is inappropriate to utilize data that may be many product-cycles old.

In the modern data environment one is thus presented with a difficult problem. How can we combine data coming from a wide variety of sensors, with varying error characteristics and operating at differing observational frequencies, to produce time-consistent sets of products that will be trusted by the user? Second, how can we meet the challenge of assimilating vast numbers of data of highly varying quality? Third, can we scale the problem to be economically feasible for local environments where high-speed computers may not be available?

The answer to these questions is the focus of this paper: namely the development of a model running in observation space which uses Kalman filter (Kalman 1960) principles to create a unique short-term prognostic model for <u>every</u> observing location. Once applied this Kalman model creates a surface database that is continuous in time, has predictable error characteristics, and has a full suite of representative observations at any instant. The benefits of such a model will be made evident:: constant data density, a bulletproof quality control, precise error estimates not much larger than the instrument itself, improved analyzed product continuity, short-range forecasts for surface stations, validation of NWP in forecasting surface conditions, and fast computational characteristics. This work is a follow-on to previous work by McGinley and Stamus (1996 and 1998)

2. THE SURFACE DATA PROBLEM

At the NOAA Forecast Systems Laboratory we maintain a surface dataset that resembles a modern surface dataset as described. This abundant and diverse data has a wide range of: quality of maintenance (7day/24h to monthly) ; frequency of observation (1 min. to 3 h.); and frequency of communication (1 min. to 3 h.). The nonstandard nature of much of these data results in highly variable station counts from hour to hour (Fig 1) .

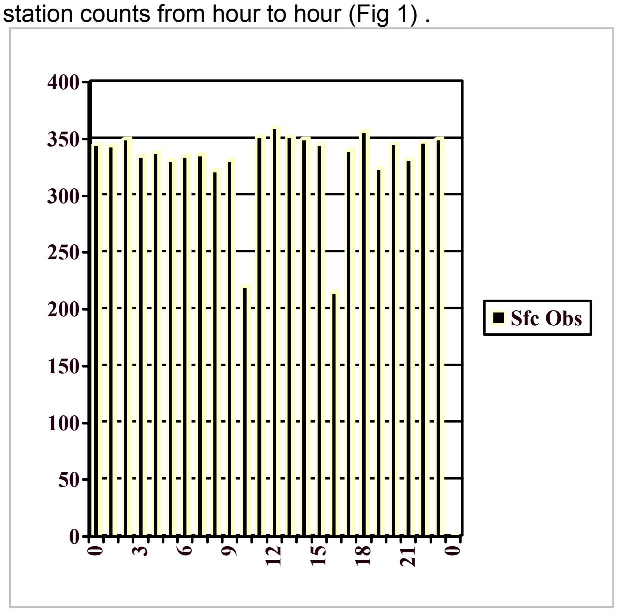


Fig. 1: Number of observation versus time (UTC) on 19 March 2001, for the 11-state area in Fig. 2. Note that due to a data outage no surface data was available at 0000 UTC on 20 March.

3. THE KALMAN FILTER

The Kalman filter (Kalman 1960; Kalman and Bucy, 1961) deals with estimation of stochastic processes that are generated by randomly perturbed difference equations (Daley, 1991). The Kalman filter provides the means for updating estimates of an unknown process by combining observations of that process with a model of the process. The Kalman filter has had wide application in prediction of spacecraft orbits, signal processing, and control, and in meteorology. The most successful applications have been where processes can be modeled with quasi-linear systems of equations. In our application we wish to advance a vector of processed station observations X , forward in time by the linear matrix operator F. The hatted quantity X_t is the Kalman estimate derived from the previous observational cycle $X_{t-\Delta t}$.

$$\hat{X}_t = F_{t(\Delta t)} X_{t-\Delta t}$$

The associated error covariance P, must be advanced forward in time by applying the forward model and adding cycle-averaged error covariance matrix W. Earlier work (McGinley and Stamus 1998) indicated that the best estimate for W comes from averaging more than 12 cycles (24 is chosen).

$$\hat{P}_t = F_t \ P_{t-\Delta t} F_t^T + W$$

$$W = \left\langle \overline{e \bullet e^T} \right\rangle$$

$$e = X_{t-\Delta t} - X_{t-\Delta t}^{truth}$$

The estimate of X is used as the new datum vector until the arrival of the new set of observations Y. When the new observations arrive, estimates of the error can be made and the error matrix W updated. The updated Kalman observation is then determined by adding the first estimate to the product of the Kalman gain K and the innovation vector made up of differences from the observations and the first estimate.

$$K = \hat{P}_t H^T (H \hat{P}_t H^T + V)^{-1}$$

$$X_t = \hat{X}_t + K(Y_t - H\hat{X}_t)$$

$$V = < \overline{o \bullet o^T} >$$

$$o = Y - Y^{truth}; Y = HX_{truth} + o$$

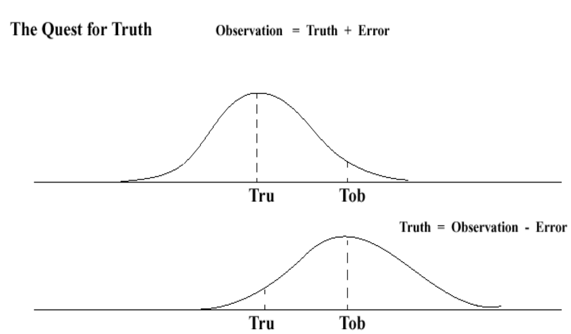
Above, K is the Kalman Gain, V is the observational error matrix, usually a diagonal matrix of known measurement error for surface data. For our application H is the identity matrix; superscript T is the transpose operator.

3.1 The Forward Model F

The forward model possesses components that a human would employ if given the task for manually estimating or predicting an observation based on past information. The method is based on consistent trends from data cycle to data cycle. The sources of trend information in a data environment would be provided by: 1) a trend from a station time series, 2) a trend information by the stations in the vicinity and 3) trends from numerical prediction products. A forecaster estimating data would seek information from each of these sources, and over time would learn what components provide the best estimates. These are the attributes we seek in defining the forward model. We specify that the model produce a composite trend from the three estimates and evolve the relative contribution of each based on past performance. Thus, F varies in time reacting to how well a station value is projected by trend from itself, buddies, or NWP. Limited space precludes showing the full forward operator. The RUC model was used for the NWPcomponent; the Eta if the RUC was unavailable. Resolutions for RUC and Eta were 40-and 32-km, respectively.

3.2 **Estimate of Truth**

Clearly the Kalman model is dependent on a truth estimate if we hope to get realistic error estimates. We employ the principle that for a well-performing, unbiased instrument the observation extracted is distributed about truth as a Gaussian probability density function with zero mean and standard deviation of the RMS instrument error. It is also true is that given an observation, truth is distributed about the observation as a Gaussian PDF. We estimate truth using a Gaussian random number with zero mean and standard deviation of one and multiply this times the RMS observation error. While this is most likely a poor assumption for any given observation, over time for many, many truth extractions we likely get error statistics that are sound. The observation used must past a gross error check prior to the truth estimate.



4. THE KALMAN OBSERVATIONS

The Kalman estimates can be used in many ways. After setting up error thresholds and running the system in parallel, they can serve as rejection criteria for conducting gross and standard error checks. The user may opt for utilizing the Kalman suite of observations X_t as a replacement for Y. Once the new observations Y arrive at time t, X_t will be the optimum value given forward-model and observational accuracy. X_t will also provide values for missing stations. Product generation can be done with simple, efficient analysis schemes, making costly four-dimensional data assimilation unnecessary. The observation projection (hatted-X) can serve as a short range forecast tool for individual stations.

5. CASE STUDY

An example of the Kalman scheme is shown for a time period of low data count for 19 March 2001 at 1600 UTC. The Kalman scheme has processed over 800 observations in the 11-state area shown in figs 2-5 in about 6 minutes on a PC-Linux platform. For this case just over 400 observations are processed.

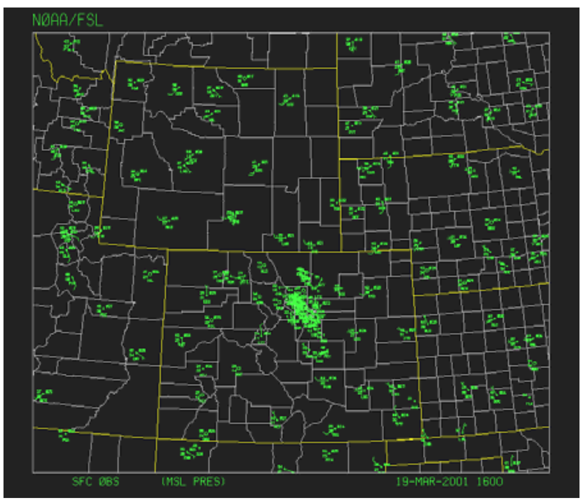
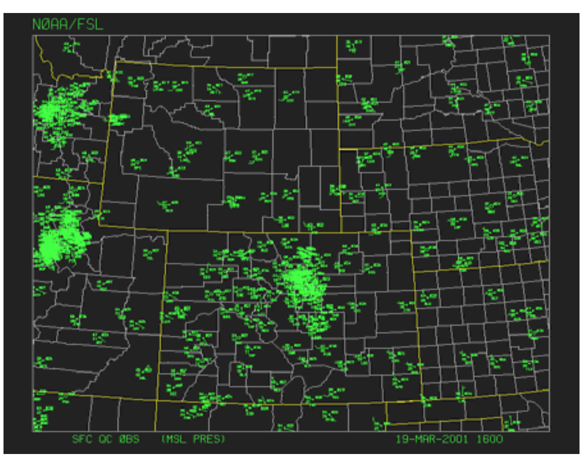


Fig 2. Each plot (not intended to be readable) indicates an observation for 1600 UTC 19 March 2001. A total of 215 observations arrived. Below shows the same time with the complete set of Kalman-generated observations for a total of 432 observations.



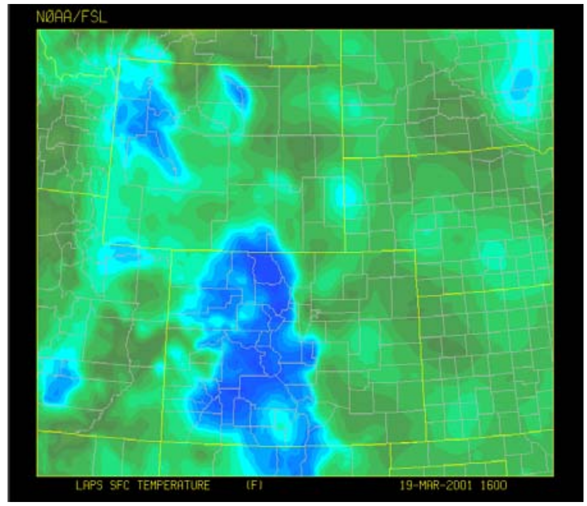
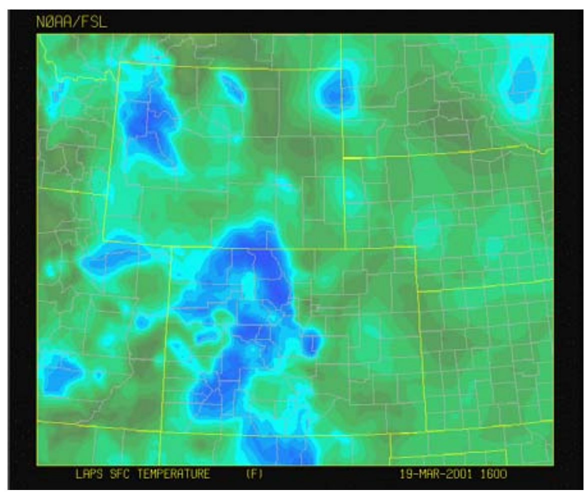


Fig 3a,b. Surface temperature analysis for 1600 UTC 19 March 2001. On a, above, note the lack of detail over the mountain areas where much of the data was lost. Below b shows the detail improved with use of the Kalman observations for the same time.



6. SEASONAL PERFORMANCE OF THE FORWARD MODEL

The Kalman scheme was run for a full year with nearly complete coverage for each season except summer (1.5 months represented). This provided an opportunity to evaluate the performance of the complete Kalman model and the individual components. The data vector contained estimates for temperature, dewpoint, vector winds, and MSL pressure. Fig. 4 illustrates the weight

applied to each component of the forward model. The sum of the weights is normalized so equal performance of each component would result in a weight of 0.33333….
For the results show in Figs. 4,6,7, the target is a 1-h temperature observation projection (one cycle). The target in Fig. 5 is a 1-h wind projection. Fig. 4 shows that the self trend and buddy trend perform close to equally well in all seasons. NWP does not do as well.

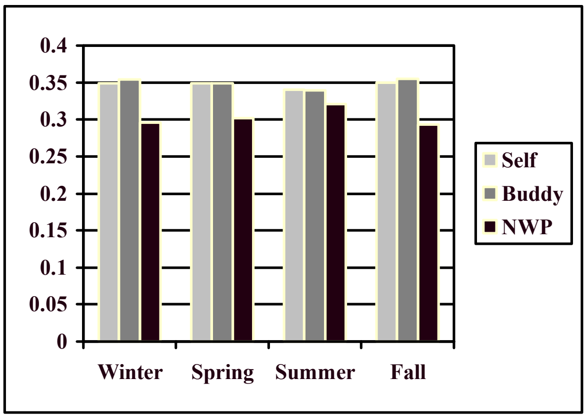


Fig 4 Forward model weights for temperature for each of the three components: self trend, buddy trend and NWP trend. Weights total to 1.0. Low values indicate poor performance. Note that NWP is best during summer but lower than self trend and buddy trend through most of the year.

Figure 5 shows the same characteristics for winds. Here we see much better NWP performance, particularly in spring and summer. The seasons in question have better mixed boundary layers, and better coupling to the rest of the atmosphere may result in better surface prediction of winds.

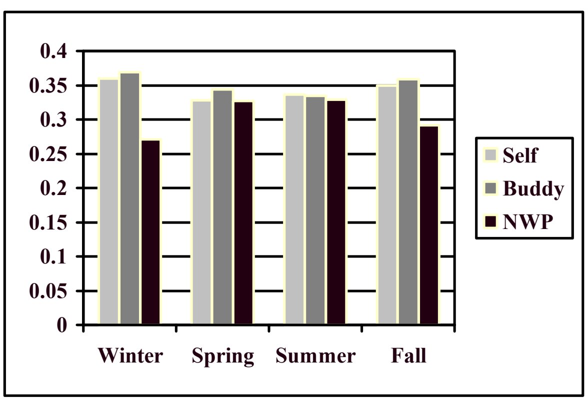


Fig. 5: Kalman model weights for vector wind for each of the three components: self trend, buddy trend, and NWP trend. Weights total to 1.0. Low values indicate poor performance. Note that NWP is best during spring and summer but lower than self trend and buddy trend through most of the year.

With over 400 stations in the suite of data it is difficult to select a subset. However, it is instructive to look at errors for individual stations. In Fig. 6 we consider a set of stations representing conditions in the western plains and mountain regions. Stations selected are Denver (DEN: mountain lee); Salt Lake City (SLC: mountain basin) ; Goodland (GLD: plains) Farmington (FMN: mountain plateau); and Billings (BIL, mountain lee). Note the larger NWP errors in the mountain basin and lee areas and poor performance for temperature in the basin. However, NWP does a good job in at the lee stations.

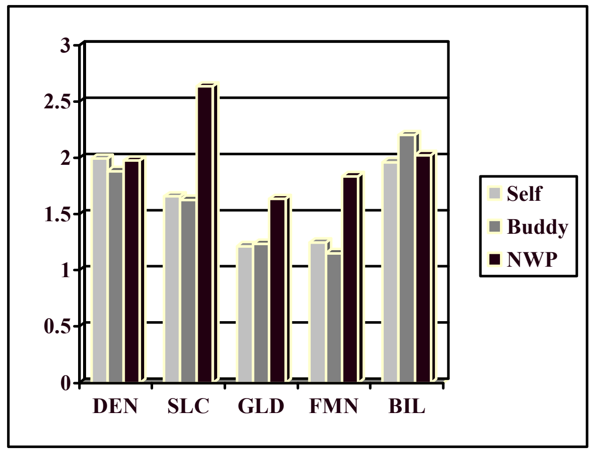


Fig 6. Temperature error (F) for Denver, Salt Lake City, Goodland, Farmington, and Billings for winter 2001.

For certain stations persistence was used as a model component to compare to the buddy and self trends. Figure 7 shows winter results for these stations, Helena (HLN, mountain lee) , Gallup (GUP, mountain plateau), Hutchinson (HUT, plains) ; DPG (Dugway Proving Ground; mountain basin); Las Vegas, NM (LVS, lee). In every case except the mountain basin (DPG) the error using persistence is larger than NWP errors in similar regimes show in Fig. 6. Buddy and self trends perform better than persistence except at Las Vegas.

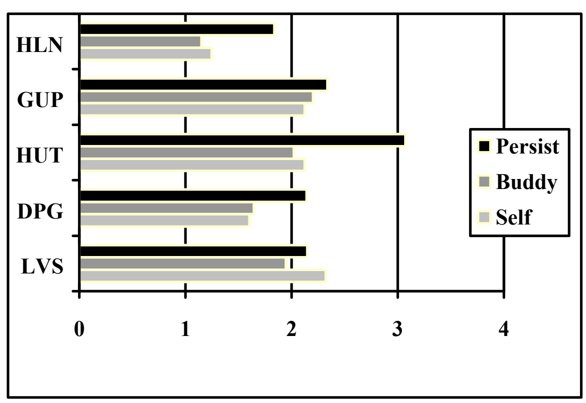


Fig. 7: Temperature error (F, x-axis) for self trend and buddy trend compared with persistence projections (persistence is a zero trend). All show persistence is inferior except at Las Vegas, NM, a lee station.

7.THE KALMAN FILTER AS A SHORT-TERM FORECAST MODEL

The Kalman Filter is able to sustain reasonably accurate observations while a station may be out of action. Figure 8 shows a plot for Aurora NE that had an outage for a number of hours. Note the consistency of the temperature and dewpoint trace based on continuing buddy and NWP trends. In this case hourly NWP trends are taken from the latest model guidance. When the observation reappears after 8 h we see that only minor error is evident in the Kalman observation.

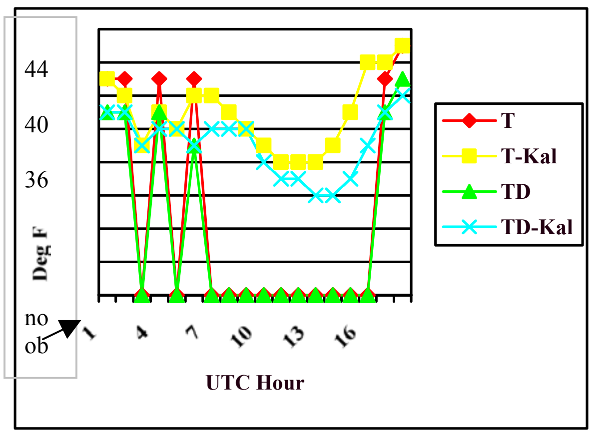


Fig. 8: Time series showing observations at Aurora, NE (temperature, dewpoint, and Kalman-created observations (T-Kal, TD-Kal) over an 18-hour period on 22 March 2001. There are no observations from 0600-1700 UTC yet the model keeps pace with diurnal cooling and when the observation reappears after 11 h only 2F temperature error and 1F dewpoint error result.

A good test for the Kalman filter is to look at the projection error over periods of time longer than the one-hour data cycle. We can do this by compiling statistics in cases where a station vanishes for a period of hours as did Aurora, either routinely (a part-time site) or because of other problems. If we compute the error between the Kalman estimate and "truth" when this station reappears, we create forecast error statistics for any number of hours. The errors are binned by hour for all stations and are averaged for each season. The plots for temperature forecasts out to 12 hours are shown in Fig. 9. Note that errors less than 5F are common out to 4 h.

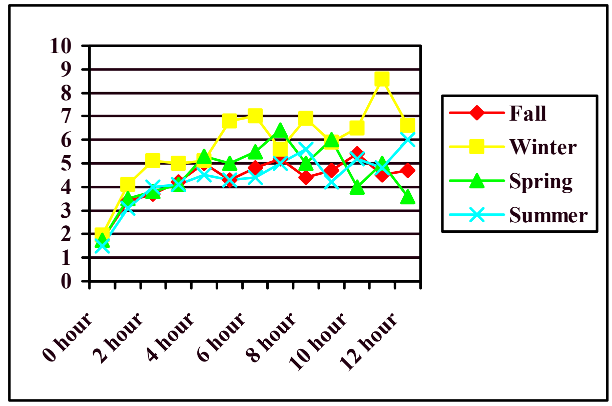


Fig. 9: Temperature error (F) for Kalman forecasts for large numbers of missing stations, as a function of forecast length (abcissa) for seasons as indicated. Spikes in trace indicate small samples for some hour bins.

8. SUMMARY

The Kalman model as discussed here is an appropriate tool for environments where data quality control is a high priority, where computer resources are limited, where data sources are diverse, and where short -range forecasts are needed for point locations.

9. REFERENCES

Daley, R., 1991 : *Atmospheric Data Analysis*, Cambridge Univ. Press, 457pp

Jesuroga, R., C. Subramaniam, and P. Miller, 1998: The AWIPS Local Data Acquisition and Dissemination System, 368-370 Proceedings, *14th Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* Phoenix, AZ, Amer.Met.Soc,, 268-370

Kalman, R., 1960: A new approach to linear filtering and prediction problems, *Trans. ASME, Ser.,D,J, Basic Eng.,* **82**,35-45.

Kalman, R. and R. Bucy, 1961: New results in linear filtering and prediction theory, *Trans. ASME, Ser.,D,J, Basic Eng.,* **83**,95-108.

McGinley J. and P. Stamus, 1996: A quality control scheme for local mesonet observations based on the Kalman Filter, *Preprints 15th Conf on Weather Analysis and Forecasting,* Norfolk VA, Amer. Met. Soc 223-226

McGinley J. and P. Stamus, 1998: A quality control scheme for local mesonet observations based on the Kalman Filter*:* Tests with real data *Second Symposium on Integrated Observing Systems,* Phoeniz, AZ, Amer. Met. Soc, 208-210