Kalman Filter and Analogs to Improve Numerical Weather Predictions

Luca Delle Monache*, Gregory Roux, Yubao Liu, and Thomas Warner
NCAR

Thomas Nipen, and Roland Stull
UBC

Neil Jacobs, and Peter Childs
AirDat LLC

The 4th EnKF Workshop -- April 7, 2010, Rensselaer, NY

* lucadm@ucar.edu
Outline

• Ensemble and Kalman filtering (KF) for air quality predictions
• A new method based on KF and an analog approach (ANKF, AN)
• Test KF, ANKF, and AN to correct 10m wind speed
• Application of new methods within an EnDA framework
• Summary
A Kalman filter bias correction for deterministic and probabilistic ozone predictions

Model Domains Used in Ensemble Forecast Study

(a)

Region of 8 model overlap, location of 358 AIRNow O₃ monitors

(b)

- Summer of 2004 (56 days)
- 8 photochemical models
- 360 ozone surface stations

Sources:
Delle Monache et al. (JGR, 2006b)
Delle Monache et al. (Tellus B, 2008)
Djalalova et al. (Atmospheric Environment, 2010)
Ensemble averaging and Kalman Filtering effects on systematic and unsystematic RMSE components

RMSE decomposition (Willmott, Physical Geography 1981):

\[ \text{RMSE} = \sqrt{\text{RMSE}_s^2 + \text{RMSE}_u^2} \]

\[ \text{RMSE}_s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{P}_i - O_i)^2} \]

\[ \text{RMSE}_u = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{P}_i - P_i)^2} \]

\[ \hat{P}_i = a + bO_i \]

a and b least-squares regression coefficients of P_i and O_i
Kalman Filtering effects on probabilistic prediction reliability
Kalman Filtering effects on probabilistic prediction resolution

![Graph showing the relationship between ROC Area and Concentration Threshold (ppbv)]
NOTE

This procedure is applied independently at each observation location and for a given forecast time.

"Analog" Space
How to find analogs? (1)

We can define a metric:

\[ d_{t_i} = \left\| f_{t_n} - A_{t_i} \right\| = \sum_{\text{var}=1}^{N_{\text{var}}} w_{\text{var}} \frac{1}{\sigma_{f^{\text{var}}}} \sqrt{\sum_{k=-\tilde{\tau}}^{+\tilde{\tau}} (f_{t_n+k}^{\text{var}} - A_{t_i+k}^{\text{var}})^2} \]

Where,

- \( N_{\text{var}} \) is the number of variable to compute the “distance” between \( f_{t_n} \) and \( A_{t_i} \)
- \( w_{\text{var}} \) is the weight given to each variable while computing the metric
- \( \sigma_{f^{\text{var}}} \) is the standard deviation of the set \( \{ f_t^{\text{var}} \} \) with \( 0 \leq t \leq t_0 \)
- \( \tilde{\tau} \) is the half-width of the time window over which differences are computed
How to find analogs? (2)

We can define a metric:

\[ d_{t_i} = \left\| f_{t_i} - A_{t_i} \right\| = \sum_{\text{var}=1}^{N_{\text{var}}} w_{\text{var}} \frac{1}{\sigma_{f_{\text{var}}}} \sqrt{\sum_{k=-\tilde{t}}^{+\tilde{t}} (f_{t_i+k} - A_{t_i+k})^2} \]
How to find analogs? (3)

We can define a metric:

\[ d_{t_i} = \left\| f_{t_n} - A_{t_i} \right\| = \sum_{var=1}^{N_{var}} w_{var} \frac{1}{\sigma_{f_{var}}} \sqrt{\sum_{k=-\tilde{t}}^{+\tilde{t}} (f_{t_n+k}^{var} - A_{t_i+k}^{var})^2} \]
How to find analogs? (4)

We can define a metric:

\[ d_{t_i} = \left\| f_{t_n} - A_{t_i} \right\| = \sum_{\text{var}=1}^{N_{\text{var}}} w_{\text{var}} \frac{1}{\sigma_{f_{\text{var}}}} \sqrt{\sum_{k=-\tilde{t}}^{+\tilde{t}} (f_{t_n+k}^{\text{var}} - A_{t_i+k}^{\text{var}})^2} \]
Modeling settings for wind predictions  
(wind energy project)

D1: 30 km, 128x114  
D2: 10 km, 253x232  
D3: 3.3 km, 541x571  
NOTE: 37 vertical levels, with 12 levels in the lowest 1-km

physics:  
- Lin et al. microphysics  
- YSU for PBL  
- Monin-Oboukov for SL  
- Kain-Fritsch CUP (Domain 1/2)  
- Noah Land Surface Model

WRF Model
Wind speed: statistics as function of time

- 6 months period
- 500 surface stations
- $\tilde{t} = 1$
- $w_{\text{var}} = 1, \forall \text{var}$
- Variables: $u, v, T, P, Q$ @ surface
Wind speed: RMSE (m/s) as function of space
Sensitivity to data set length

- NWP model: MM5
- 1 year period
- 22 surface stations in NM
- $\tilde{t} = 1$
- $w_{\text{var}} = 1, \forall \text{var}$
- Variables: u, v, T @ surface

**KF vs ANKF**

![Graph showing KF vs ANKF skill scores over data set length (months)]
Ensemble filters main steps: Possible sources of error

Sources of error:
- Model error: (1)
- **Forward operator (H) errors** (e.g., interpolation, flow-dependent behavior of representativeness errors, etc.): (2)
- Instrumental errors, retrieval and transmission of obs: (3)
- Gaussian and other (depending on the method) assumptions: (4)
- **Sampling errors** (analogs as a way to populate an ensemble): (4), (5)
- **Errors from linear regression between obs and state variables increments**: (5) (next slide)

Source: Adapted from Anderson (Fig. 1 – Physica D, 2007)
Summary

- Ensemble and Kalman filtering (KF) for air quality predictions
- A new method based on KF and an analog approach (ANKF, AN)
- Test KF, ANKF, and AN to correct 10m wind speed
  - ANKF and AN beats KF over a range of metrics
  - ANKF gain vs KF grows with length of data set
- The analog concept can be explored also within an ensemble data assimilation framework
GRAZIE!
(lucadm@ucar.edu)
KF in analog space (1)

- $f_n$ is a forecast at time $t_n$ and at a given location, with $t_n > t_0$
- $d_i = \|f_n - A_i\|$ is a metric to measure the “distance” between $f_n$ and $A_i$
- $\{A_i\}$ is a set of “analog” forecasts at a time $t_i$, with $t_i < t_0$
  - $\{A_i\}$ are ordered with respect to $d_i : d_{i-1} > d_i$, and $\{i, N \in \mathbb{N} : 2 \leq i \leq N = |\{A_i\}|\$

We can now introduce the Kalman filter bias correction procedure as follows:

- The true unknown forecast bias at time $t_n$ can be modeled by
  
  $$x_{t_n|t_{n-1}} = x_{t_{n-1}|t_{n-2}} + \eta_{t_{n-1}}, \eta \sim \mathcal{N}(0, \sigma_{\eta_{t_{n-1}}}^2)$$

- And the actual forecast error can be expressed as
  
  $$y_{t_i} = A_{t_i} - O_{t_i} = x_{t_i} + \varepsilon_{t_i} = x_{t_{i-1}} + \eta_{t_{i-1}} + \varepsilon_{t_i}, \varepsilon_{t_i} \sim \mathcal{N}(0, \sigma_{\varepsilon_{t_i}}^2)$$
The optimal recursive predictor of $x_t$ can be written as

$$\hat{x}_{t_n|t_{n-1}} = \hat{x}_{t_{n-1}|t_{n-2}} + K_{t_{n-1}|t_{n-2}} (y_{t_{n-1}} - \hat{x}_{t_{n-1}|t_{n-2}})$$

Where $K$, the “Kalman gain” is

$$K_{t_n|t_{n-1}} = \frac{p_{t_{n-1}|t_{n-2}} + \sigma_{\eta_{t_{n-1}}}}{(p_{t_{n-1}|t_{n-2}} + \sigma_{\eta_{t_{n-1}}}^2 + \sigma_{\varepsilon_{t-1}}^2)}$$

And $\rho$, the expected mean square error is

$$\rho_{t_{n}|t_{n-1}} = (p_{t_{n-1}|t_{n-2}} + \sigma_{\eta_{t_{n-1}}}^2)(1 - K_{t_{n-1}|t_{n-2}})$$

NOTE: The system of equations is closed by:

- first running the filter for $\sigma_{\varepsilon}^2$ (with $\sigma_{\varepsilon}^2$, $\sigma_{\eta}^2$ constant)
- $r = \sigma_{\eta}^2 / \sigma_{\varepsilon}^2$
Ensemble filters main steps:
Possible sources of error

(1) background forecasts (“prior”)
(2) “forward operator”
(3) observation + “gross” error
(4) updates
(5) regression into state vector

Sources of error:
• Model : (1)
• Forward operator (H): (2)
• Observation: (3)
• Gaussian and other assumptions: (4)
• Sampling: (4), (5)
• Linear regression: (5)

Source: Adapted from Anderson (Fig. 1 – Physica D, 2007)
Photochemical modeling

**INPUTS**

- Emissions
- Meteorology
- Chemistry

**THE MODEL**

Air Quality Model

**OUTPUTS**

- Advection, Diffusion, Deposition, etc.
- July 18, 1995 23:00:00
  - Min = 0.021 at (241,1)
  - Max = 0.130 at (41,73)
Kalman Filter predictor bias correction

Abbotsford, 9-15 August 2004

\[
FRI = \frac{|Raw Fcsts - KEK|}{|Raw Fcsts - Obs|}
\]

“Fractional Relative Improvement”

Delle Monache et al., Journal of Geophysical Research (2006b)
KF results: Chilliwack
A Kalman filter bias correction for deterministic and probabilistic PM$_{2.5}$ predictions

Source: Djalalova et al. (Atmospheric Environment, 2010)
Error-ratio sensitivity tests
“Global” Statistics

RMSE (m s⁻¹)

Correlation

MAE (m s⁻¹)
Statistics in space: Correlation

[Map Image with Various Correlation Maps: Raw, KF, ANKF, AN]
Statistics in space: MAE (m/s)

Raw

KF

ANKF

AN
Statistics in space: BIAS (m/s)
Wind speed: PDFs (observations vs. predictions)

- **Raw**
- **KF**
- **ANKF**
- **AN**