Accounting for hidden error distributions in ensemble based state estimation and prediction

Dr. Craig H Bishop
Code 7531
Marine Meteorology Division
(831) 656 - 5715
bishop@nrlmry.navy.mil
Overview

- **Aim:** To optimally predict flow dependent forecast error variances using smoothed ensemble variances and climatological information.
- **Motivation:** Quality control, DA, isolating error/ensemble correlations
- **Prior Work:**
  a. Ensemble variance smoothing used operationally at Meteo-France and ECMWF
  b. Climatological innovation variance information used at both Meteo-France and ECMWF to inform their variance prediction.
  c. Bishop et al. (2013) provides methods for optimally combining ensemble variance and climatological error variance information at a point but not for a spatial field of variances
- **Current Work:** Extends (c) to spatial fields of variances.
Idealized experiments by Raynaud et al. (2008) showed that provided the true variance field has a variation length scale much longer than the error correlation length scale then spurious fluctuations in ensemble variance have a much shorter scale than fluctuations in the true error variance field.
Improve quality of the ensemble
Tests of new variance smoother

Take away points:
1. Smoothed variance is closer to true variance than raw ensemble variance
2. New NRL method outperforms Meteo-France and ECMWF approach
   a. new smoothed variance is closer to true variance
   b. new approach includes full posterior distribution of ensemble variances

**Yellow** gives true error variance
**Red** gives the ensemble variance from an ensemble of \( M_{\text{prior}}=16 \) members
**Green** gives new filtered estimate of true error variance
**Cyan** gives Berre et al. (2007) filtered estimate of the true error variance
**Thin-colored** lines give climatological distribution of true error variances from \( K_{\text{clim}}=10 \)
**Black** lines give random draws from posterior distribution of forecast error variances.
Concluding remarks

• Previous approaches to ensemble variance smoothing ...
  i. Do not explicitly incorporate information about seasonally averaged error covariances
  ii. Do not attempt to represent the distribution of true error variances given an imperfect error variance prediction

• Variance estimation approach described here ..
  a. Optimally incorporates information about seasonally averaged error covariances
  b. Estimates the posterior distribution of true error variances
Concluding remarks

• Previous approaches assumed that the ensemble variance is equal to the true variance plus a noise term whose variance was solely determined by the actual size of the ensemble.

• Here, a new approach has been described in which the statistical analysis of (observation, forecast, ensemble variance) yield estimates of
  – The factors required to debias the ensemble
  – The actual noise in the debiased ensemble and an associated effective ensemble size
  – The variance and covariance of the climatological distribution of true error variances.

• Performance of previous methods is poor when length scale of $P^C$ is smaller than that of $R^I$. Optimality of new method is independent of this parameter.
Assumption 1: The error of the deterministic forecast is a random draw from a Gaussian distribution, whose true variance $\sigma_i^2$ is a random draw from a prior climatological inverse gamma pdf of error variances.

$$x^f = x^t + \varepsilon^f, \varepsilon^f \sim N(0, \sigma_i^2)$$

$$\sigma_i^2 \sim \rho_{prior}(\sigma^2) = \gamma^{-1} \left[ \langle \sigma^2 \rangle, \text{var}(\sigma^2), \sigma_{min}^2 \right]$$

The histogram of the climatological distribution of true error variances shown on the left was obtained by running 25,000 independent replicate ETKF data assimilation cycles for model 1 of Lorenz (2005). Each cycle suffered independent observation errors but the truth was identical in each cycle. The 25,000 replicates enabled the true error variance given a particular true state and observing network to be computed at each grid point.
An analytic model of hidden error variance

Assumption 2: Ensemble variances are drawn from a likelihood gamma pdf of ensemble variances with mean $a(\sigma_i^2 - \sigma_{\text{min}}^2) + s_{\text{min}}^2$

\[
s_i^2 = a(\sigma_i^2 - \sigma_{\text{min}}^2)\eta + s_{\text{min}}^2, \quad \text{where} \quad \eta \text{ is a (+ve) stochastic variable}
\]

\[
\left\langle s_i^2 \mid \sigma_i^2 \right\rangle = a(\sigma_i^2 - \sigma_{\text{min}}^2) + s_{\text{min}}^2
\]

\[
s_i^2 \mid \sigma_i^2 \sim \mathcal{L}(s_i^2 \mid \sigma_i^2) = \gamma \left[ \left\langle s_i^2 \mid \sigma_i^2 \right\rangle, \text{var} \left( s_i^2 \mid \sigma_i^2 \right), s_{\text{min}}^2 \right]
\]

\[
\text{M} = 4 \quad \text{M} = 8
\]

$s_i^2$ is ensemble variance, $\sigma_i^2$ is the true flow dependent error variance

\[
\frac{\text{var} [ (s_i^2 - s_{\text{min}}^2) \mid \sigma_i^2 ]}{\left[ \text{mean} [ (s_i^2 - s_{\text{min}}^2) \mid \sigma_i^2 ] \right]^2} = \frac{\text{var} [ (s_i^2 - s_{\text{min}}^2) \mid \sigma_i^2 ]}{\left[ a\sigma_i^2 \right]^2} = \frac{1}{k}, \quad \text{Effective ensemble size} = 2k - 1 = M
\]
Green lines give mode
Blue lines give mean

Given an ensemble variance, there are a broad range of possible true error variances.

Current DA schemes require a single value.

For the minimum error variance estimate, use the posterior mean.

For the maximal likelihood estimate, ...

For QC, ...
New equations have been derived that estimate hidden error variance parameters from a long time series \( (v_i, s_i^2), i = 1, 2, ..., n \) of (innovation, ensemble-variance) pairs:

\[
\langle \sigma^2 \rangle = \langle v^2 - R \rangle, \text{ where } R \text{ is the observation error variance}
\]

\[
\left\langle \left( \sigma^2 - \sigma_{\min}^2 \right)^2 \right\rangle P^R = \text{var} \left( \sigma^2 \right) = \frac{\langle v^4 \rangle}{3} - \left( \langle \sigma^2 \rangle + \langle R \rangle \right)^2 - \text{var} \left( R \right)
\]

\[
a = \frac{\text{covar} \left( v^2, s^2 \right)}{\text{var} \left( \sigma^2 \right)}, \quad \left[ \text{recall that } \langle s^2 | \sigma^2 \rangle = a \left( \sigma^2 - \sigma_{\min}^2 \right) + s_{\min}^2 \right]
\]

\( s_{\min}^2 \) - assigned by designer or \( s_{\min}^2 = \min \left( s_i^2 \right) \) over all \( i \) if sample is large

\[
\sigma_{\min}^2 = \langle \sigma^2 \rangle - \frac{\langle s^2 \rangle - s_{\min}^2}{a}
\]

\[
R_r^L = \frac{\text{var} \left( s^2 | \sigma^2 \right)}{\left( \langle s^2 | \sigma^2 \rangle - s_{\min}^2 \right)^2} = \frac{\text{var} \left( s^2 \right) - a^2 \text{var} \left( \sigma^2 \right)}{a^2 \left( \langle \sigma^2 \rangle - \sigma_{\min}^2 \right)^2 + \text{var} \left( \sigma^2 \right)} = \frac{2}{M - 1}
\]
Posterior mean error variance is a Hybrid combination of static and ensemble variances

\( \sigma^2 : \text{true error variance} \)
\( s^2 : \text{ensemble variance} \)

\( P^R : \) Relative variance of prior climatological pdf of error variances

\( R^R : \) Relative variance of likelihood pdf of ensemble variances \( L(s^2 \mid \sigma^2) \)

Flow dependent ensemble variance

\[
\langle \sigma^2 \mid s^2 \rangle = \langle \sigma^2 \rangle + \left[ \frac{P^R}{P^R + R^R} \right] \left\{ \left[ \frac{s^2 - s_{\min}^2}{a} + \sigma_{\min}^2 \right] - \langle \sigma^2 \rangle \right\}
\]

\[
= \left[ \frac{P^R}{P^R + R^R} \right] \left[ \frac{s^2 - s_{\min}^2}{a} + \sigma_{\min}^2 \right] + \left[ \frac{R^R}{P^R + R^R} \right] \langle \sigma^2 \rangle
\]

i. As the stochastic variation of ensemble variance about the true variance goes to zero, the weight on the ensemble variance goes to 1.

ii. If there is any imperfection in the flow-dependent ensemble variance, the optimal error variance estimate gives weight to the climatological covariance.

iii. If there is no variance of the true error variance, the weight on the static variance goes to 1.

All of this work pertains to variances at a point.
It does not pertain to fields of variances.
Extension of hidden error variance theory to fields

\( \hat{\sigma}^2 \) \( \sim \) denotes true flow dependent error variance

\( \mathbf{P}^C \) \( \sim \) denotes covariance matrix of climatological realizations of true error variances

\( \hat{s}^2 \) \( \sim \) denotes vector-field of ensemble variances

\[
\hat{\sigma}_n^2 = (\hat{s}^2 \cdot / \hat{a}) + \left[ \hat{\sigma}_{\text{min}}^2 - (\hat{s}_{\text{min}}^2 \cdot / \hat{a}) \right] \sim \text{denotes "debiased" ensemble variance}
\]

\( \mathbf{R}^L \sim \) denotes error covariance matrix of \( \hat{\sigma}_n^2 \)

Single variable work suggests

\[
\hat{\sigma}^2 | \hat{s}^2 = \left[ \langle \hat{\sigma}^2 \rangle + \mathbf{K} \left( \hat{\sigma}_n^2 - \langle \hat{\sigma}^2 \rangle \right) \right] \cdot \eta
\]

It can be shown that optimal \( \mathbf{K} \) given by

\[
\mathbf{K} = \mathbf{P}^C \left( \mathbf{P}^C + \mathbf{R}^L \right)^{-1} \left( \hat{\sigma}_n^2 - \langle \hat{\sigma}^2 \rangle \right)
\]

\( \eta \) is stochastic with \( \langle \eta \rangle = [1,1,\ldots,1]^T \)

\[
\langle (\eta - 1)(\eta - 1)^T \rangle = \left[ \mathbf{P}^C - \mathbf{P}^C \left( \mathbf{P}^C + \mathbf{R}^L \right)^{-1} \mathbf{P}^C \right] \cdot \left[ \langle \hat{\sigma}^2 \rangle \langle \hat{\sigma}^2 \rangle^T + \mathbf{P}^C \left( \mathbf{P}^C + \mathbf{R}^L \right)^{-1} \mathbf{P}^C \right]
\]
Extension of hidden error variance theory to fields

Now

\[
\hat{\sigma}^2 \mid \hat{s}^2 = \left[ \langle \hat{\sigma}^2 \rangle + P^C \left( P^C + R^L \right)^{-1} \left( \hat{\sigma}^2_n - \langle \hat{\sigma}^2 \rangle \right) \right] \quad \eta
\]

\[
= \left[ P^C \left( P^C + R^L \right)^{-1} \hat{\sigma}^2_n + R^L \left( P^C + R^L \right)^{-1} \langle \hat{\sigma}^2 \rangle \right] \quad \eta
\]

Previous work misses the last climatological error variance term in this equation entirely. It also assumes that the noise responsible for \( R^L \) is entirely due to the sampling noise associated with an ensemble of size \( N \) and it does quantify the uncertainty of the smoothed estimate.

How can the terms in the above equation be estimated?
There is a climatological distribution of true forecast error variance fields $\hat{\sigma}^2$ with pdf denoted $\rho^C(\hat{\sigma}^2)$. This distribution has a mean $\langle \hat{\sigma}^2 \rangle$, covariance matrix $P^C$ and relative covariance matrix

$$P_r^C = \langle \hat{\sigma}^2 \hat{\sigma}^2 \rangle^{-1} P^C$$
There is a climatological distribution of true forecast error variance fields $\hat{\sigma}^2$ with pdf denoted $\rho^C(\hat{\sigma}^2)$. This distribution has a mean $\langle \hat{\sigma}^2 \rangle$, covariance matrix $P^C$ and relative covariance matrix

$$P_r^C = \langle \hat{\sigma}^2 \hat{\sigma}^{2\tau} \rangle^{-1} \Box P^C$$

Intermediate
There is a climatological distribution of true forecast error variance fields with pdf denoted $\tilde{\Sigma}$. This distribution has a mean $\tilde{\mu}$, covariance matrix $\tilde{\Sigma}$, and relative covariance matrix $\tilde{\Sigma}_r$.

$\langle \tilde{\Sigma} \rangle$, covariance matrix $P^C$ and relative covariance matrix $P^C_r = \langle \tilde{\Sigma} \tilde{\Sigma}^T \rangle^{-1} \tilde{\Sigma}^T P^C$

Could one use a historical set of (forecast, observation) pairs to find historical covariance $P^C$ of true forecast error variance fields?

Realizations of fcst error (narrow case)
Recently derived equation enables an isotropic approximation to $P^C$ to be computed from long list innovations $v_i = y_i - H_i \left( x^f \right)$. The equation is valid provided the distribution of forecast errors given a true error variance is Gaussian. The data set and the ordering of innovations required to apply the method is identical to the Hollingsworth Lohnberg approach.

The key recovery equation is

$$0 = \left( 2 \frac{\langle P_{ij} \rangle^2}{\langle \sigma_i \sigma_j \rangle^2} + 1 \right) \frac{\langle \sigma_i^2 \sigma_j^2 \rangle}{\langle \sigma_i^2 \rangle \langle \sigma_j^2 \rangle} - \frac{\langle v_i^2 v_j^2 \rangle}{\langle \sigma_i^2 \rangle^2 \langle \sigma_j^2 \rangle}.$$
Recovery of column of $P^c$ from $(y, x^f)$ pairs

Black line is covariance function of historical fields of true error variance

Green line covariance function recovered from from $(y, x^f)$ pairs.
Recovery of column of $\mathbf{P}_c$ from $(y, x^f)$ pairs

Historical covariance function of true error variances

Black line is covariance function of historical fields of true error variance
Green line covariance function recovered from from $(y, x^f)$ pairs.

Medium width covariance function
Recovery of column of $P^C$ from $(y,x^f)$ pairs

New method for extracting covariance function of historical distribution of *hidden* error variance fields achieved its objective in these three cases. Unlike previous approaches – no ensemble required.
As in Raynaud et al. (2009), we use
\[ R^L = \left[ \hat{\sigma}^2 \hat{\sigma}^{2T} \right] \otimes R^L_r = \left\langle \text{covar} \left( \hat{\sigma}^2_n \mid \hat{\sigma}^2 - \hat{\sigma}^2 \right) \right\rangle = \frac{2}{M-1} \left\langle P^f \otimes P^f \right\rangle \]
\[ \approx \left\langle \hat{\sigma}^2 \hat{\sigma}^{2T} \right\rangle \otimes R^L_r \quad \text{(exact when \( C^f \otimes C^f \) is constant)} \quad (12.1) \]

However, instead of using the actual ensemble size for \( M \) - which likely overestimates the accuracy of ensemble variances - we will use the effective ensemble size empirically defined by \( (y, x^f, s^2) \) data triplets using the method described in Bishop et al. (2013, MWR).

This approach will only work if the seasonally averaged ensemble covariance is equal to the true seasonal average of error covariances. Since the seasonally averaged forecast error covariance is not hidden, this is not difficult to achieve.
Comparison of old approach with new approach with all assumptions satisfied

**Blue line** gives true error variance
Thick red gives the ensemble variance from an ensemble of $M_{\text{prior}}=16$ members
Green gives posterior mean ensemble variance
Cyan gives Berre et al. (2007) estimate of the true error variance (blue line)
*Thin-colored* lines give climatological distribution of true error variances from $K_{\text{clim}}=10$
**Black** lines give random draws from posterior distribution of forecast error variances.

Posterior effective ensemble size, $M_{\text{post}}=37$

Posterior effective ensemble size, $M_{\text{post}}=49$
Posterior variance proportional to analyzed variance
Smoothing helpful when ens variance and climo mean variance have differing error correlation length scales

**Blue** gives true error variance  
**Red** gives the ensemble variance from an ensemble of **5 members**  
**Green** gives posterior mean ensemble variance  
**Black** lines give random draws from posterior distribution of forecast error variances. Climatological distribution of true error variances from **$K_{clim}=12$**
Smoothing helps a lot when ens variance and climo mean variance have differing error correlation length scales.

Blue gives true error variance
Red gives the ensemble variance from an ensemble of 5 members
Green gives posterior mean ensemble variance
Black lines give random draws from posterior distribution of forecast error variances.
Climatological distribution of true error variances from $K_{clim}=10$
Concluding remarks

• Previous approaches to ensemble variance smoothing ...
  i. Do not explicitly incorporate information about seasonally averaged error covariances
  ii. Do not attempt to represent the distribution of true error variances given an imperfect error variance prediction

• Variance estimation approach described here ..
  a. Optimally incorporates information about seasonally averaged error covariances
  b. Estimates the posterior distribution of true error variances
Concluding remarks

• Previous approaches assumed that the ensemble variance is equal to the true variance plus a noise term whose variance was solely determined by the actual size of the ensemble.

• Here, a new approach has been described in which the statistical analysis of (observation, forecast, ensemble variance) yield estimates of
  – The factors required to debias the ensemble
  – The actual noise in the debiased ensemble and an associated effective ensemble size
  – The variance and covariance of the climatological distribution of true error variances.

• Performance of previous methods is poor when length scale of $P^C$ is smaller than that of $R^L$. Optimality of new method is independent of this parameter.
Extra Slides
A new method

Bishop et al. (2013, MWR) show how to recover diagonal elements of $\mathbf{R}_r^L$ and $\mathbf{P}_r^C$ as well as the debiasing parameters $a$, $\sigma_{\text{min}}^2$ and $s_{\text{min}}^2$ from a long time series of $(s_i^2, y_i - h_i(x^f)), i = 1, 2, ..., n$.

Raynaud et al. (2009) shows how to get off diagonal elements of $\mathbf{R}_r^L$.

Currently, I do not know of a method to get off-diagonal elements of $\mathbf{P}_r^C$.

In the absence of a tractable method to get off diagonal elements of $\mathbf{P}_r^C$, the following method is proposed:

1. Smooth $\hat{s}^2$ using $\tilde{s}_j^2 = S_j \hat{s}^2$ for smoothers $[S_1, S_2, ..., S_n]$ of $n$ different strengths.
2. Find the smoother that minimizes $\mathbf{R}_{rj}^L = \frac{2}{M_j - 1}$. Let $\tilde{s}_o^2 (j = o)$ denote best smoother.
3. Use optimal variance predictor from hidden variance theory to combine climatological variance with smoothed ensemble variance. (This is very close to the constraint applied by Bonavita et al. 2013 before smoothing).
Test of new method

• Tested new approach using synthetic data in which the scales of the true forecast error correlation length scale and the length scale of true variances could be independently changed.
• As anticipated, it was found that the smoother that maximizes effective ensemble size \( M \), minimizes the error of the Hybrid error variance prediction obtained using hidden error variance theory.
• Without combination with the climatologically mean error variance, stronger smoothers than the smoother that maximized \( M \) worked better. This demonstrated that the smoother that is optimal with statistical correction is not necessarily the same as the optimal smoother without statistical correction.
Conclusions

• The empirical measure of “effective ensemble size” (patent pending) described in Bishop et al. (2013) enables the simultaneous tuning of variance smoothing and statistical correction. This was not possible before.

• The hidden error variance perspective makes it clear that
  – Statistical correction of ensemble variances as proposed in Bonavita et al. (2012) is equivalent to combining ensemble variance information with a climatological estimate of the error variance
  – Ideally, variance smoothing should preserve the ensemble variance in directions where the climatological true variance can vary greatly and/or where the ensemble variance is accurate. Once statistical correction is applied, it is not just a question of distinguishing noise from signal in the ensemble.