Error Variance Prediction and Deriving Optimal Weights for Combining Static and Flow-Dependent Variances

Elizabeth Satterfield[1,3], Craig Bishop[1], David D. Kuhl[2,3], Tom Rosmond[4]

[1]Naval Research Laboratory, Monterey CA, USA; [2]Naval Research Laboratory, Washington, DC, USA

[3]National Research Council, Washington, DC, USA; [4]Science Application International Corp., Forks, WA, USA

Definition of the Problem

- Accurate predictions of forecast error variance are vital to ensemble weather forecasting.
- It is difficult to find good historical analogs with which one could attempt to characterize the mean square error of forecasts similar to the particular forecast of interest.
- Ensemble forecast systems attempt to predict error variance as a function of the flow of the day.

Goals

- Develop theory for the distribution of error variances given an imperfect ensemble variance prediction.
- Derive formulae for deducing the parameters defining the prior, likelihood and posterior distributions from a large archive of innovations and corresponding ensemble variances.
- Provide a theoretical justification for Hybrid data assimilation schemes that linearly combine a climatological forecast error variance estimate with an ensemble based estimate of the forecast error variance.

A Simple Model of Innovation Variance Prediction

- The "true state" is a random draw from a climatological Gaussian distribution $x^t \sim N\left(0,\sigma_C^2\right)$

- The error of the deterministic forecast is a random draw from a Gaussian distribution whose variance is a random draw from a climatological inverse Gamma distribution of error variances

Innovation Variance

$$\overline{x^f}=x^t+\varepsilon^f,\varepsilon^f\sim N\left(0,\sigma_f^2\right)$$

$$\sigma_f^2=\omega-R,\omega\sim\Gamma^{-1}(\alpha_{prior},\beta_{prior})$$

An "Imperfect" Ensemble Prediction

$$s^2=a(\omega-R)*\eta$$

Sensitivity **Stochastic**

- represents the sample variance of M random draws from a normal distribution with zero mean and variance1.
- We assume s² is drawn from a Gamma distribution $s^2 \sim \Gamma(k,\theta)$ with mean $a(\omega-R)$

Define a Posterior Distribution of Error Variances Given An Ensemble Variance

Distribution of s² given ω **Prior climatological distribution**

$$\rho_{\text{post}}(\omega|s^2)=\frac{L(s^2|\omega)\rho_{\text{prior}}(\omega)}{\int_0^\infty L(s^2|\omega)\rho_{\text{prior}}(\omega)d\omega}$$

Naïve ensemble prediction **Climatological mean**

$$\overline{\omega}_{\text{post}}\left(s^2\right)=w_e\omega_n+\left(1-w_e\right)\overline{\omega_{\text{prior}}}$$

- This result shows optimal error variance is a combination of a static climatological prediction and a flow dependent prediction: A theoretical justification for Hybrid DA.
- Bishop and Satterfield (2011) show how parameters defining the prior and posterior distributions can be deduced from a time series of innovations paired with ensemble variances

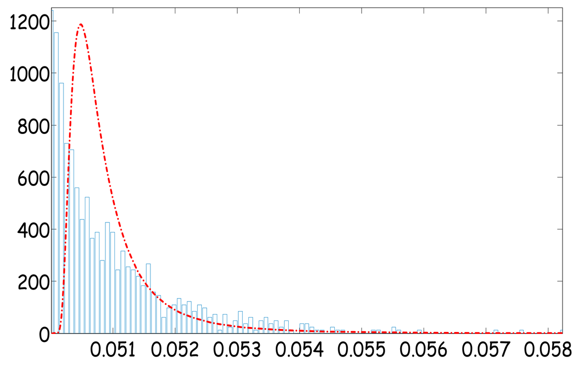
Results from Synthetic Data

	<ω>	<ω'²>	a	k	α_{PRIOR}	β_{PRIOR}	w_e
TRUE VALUE	2	1	1	4.5	3	2	0.6923
RECOVERED VALUE*	2	0.993	0.9999	4.8519	3.0033	2.0143	0.6889
TRUE VALUE	2	5	1	4.5	2.2	1.2	0.7895
RECOVERED VALUE*	2	3.875	0.9998	4.3916	2.3024	1.3022	0.831
TRUE VALUE	2	1	1	10	3	2	0.8333
RECOVERED VALUE*	2	0.981	0.9996	10.537	3.0226	2.0233	0.8296

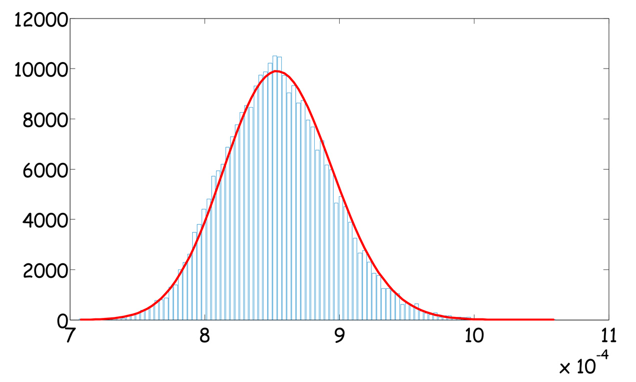
*Mean of 100 independent trials

Results from the Lorenz '96 Model

- We use 10 variable Lorenz '96 model and Ensemble Transform Kalman Filter (ETKF)
- We create 100,000 independent time series of analyses and forecast, each having the same true state and differing only in random draws of observation error.
- Error Variances are computed for each spatio-temporal point by averaging the squared error across each independent forecast.



PDF of innovation variances computed over N=25,000 trials for 100 time steps, using 100% observation coverage. The dashed line shows an inverse gamma function whose shape and scale parameters were derived from the sample data used to create the PDF.



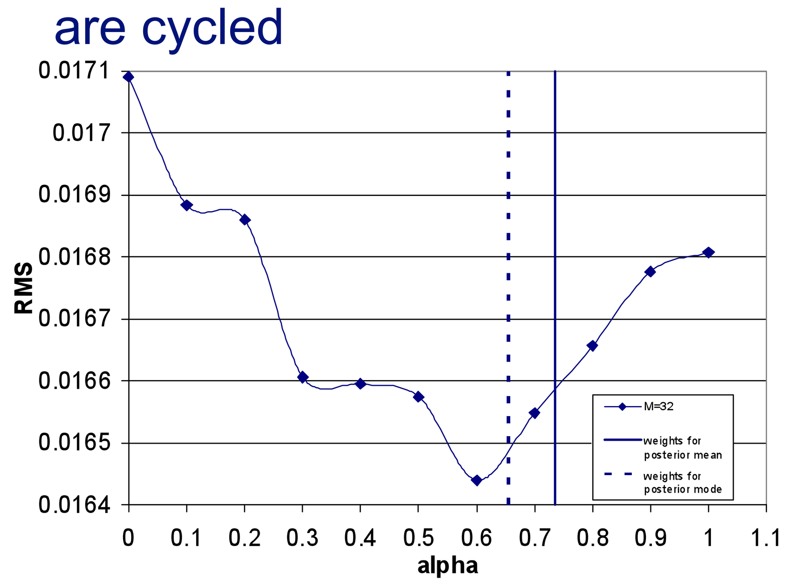
This distribution of ETKF ensemble variances given a particular error variance. Shown for the 50th percentile of error variances computed over N=25,000 trials for 100 time steps, with 100% observation coverage. The dashed line shows the Gamma function whose shape and scale parameters were computed based on the sample mean and variance.

Combining Static and Flow Dependant Variances

$$\mathbf{P}^{\mathbf{f}}=w_e\mathbf{P}^{\mathbf{f}}_{\text{ENSEMBLE}}+\left(1-w_e\right)\mathbf{P}^{\mathbf{F}}_{\text{CLIMATOLOGY}}$$

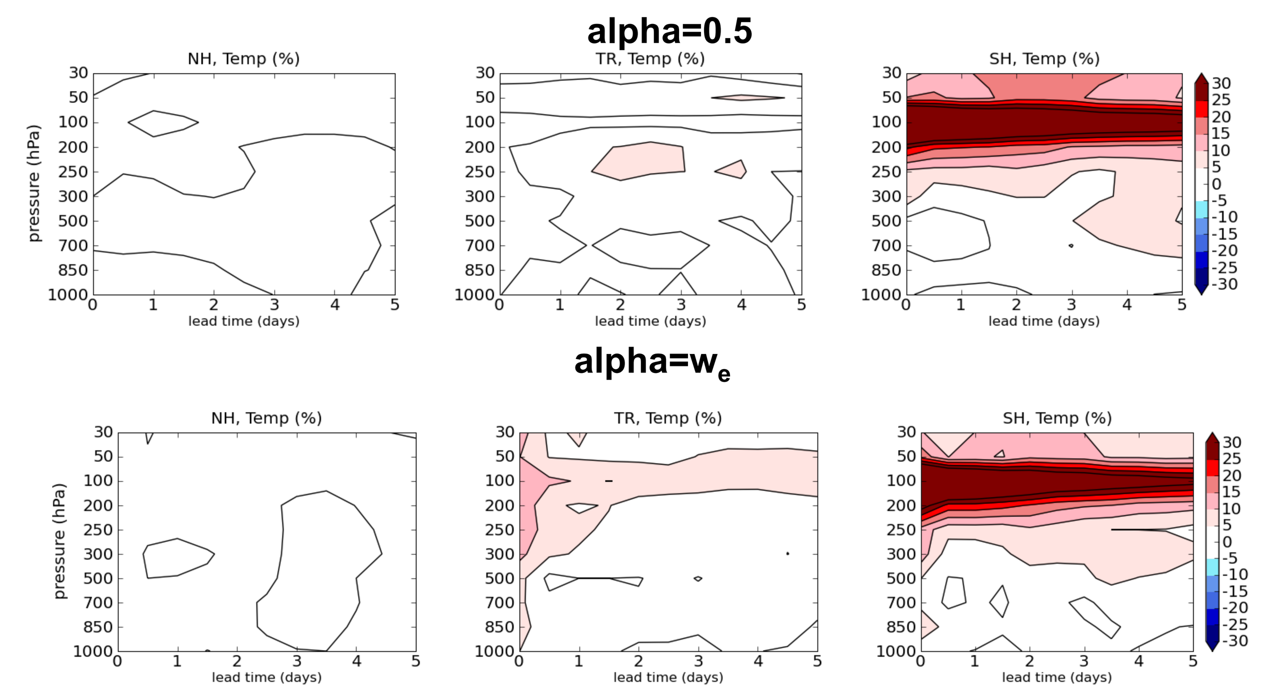
Lorenz Model Results:

- The ETKF ensemble is re-sampled to create a suboptimal M member ensemble to form the error covariance matrix and produce a suboptimal analysis.
- A climatological error covariance matrix (P^f) is formed by collecting forecast errors for 100,000 time steps
- Only the full ETKF analysis and analysis ensemble are cycled



- Adjust to posterior mean or mode?
- Modify theory to account for a non-zero minimum ensemble variance?
- Modify theory to account for model error?
- Future work is necessary to see how the size of M impacts the recovery

NAVDAS-AR-Hybrid Results:



Top three plots alpha=0.5, bottom three plots alpha computed for six different regions and linearly interpolated between. All plots are percentage reduction/increase of RMS temperature error relative to low resolution operational 4D-Var. RMS error is computed relative to radiosondes. Forecasts were launched every 12 hours from Nov. 28, 2008 to Dec. 31, 2008

Conclusions

- The distribution of error variances given an imperfect ensemble variance depends on both the climatological distribution of flow dependent error variances and the accuracy of the ensemble based error variance prediction.
- Our approximations of an inverse-gamma PDF for the prior climatological distribution of innovation variances and a gamma PDF for the likelihood distribution of ensemble variances given an innovation variance are reasonably accurate for the Lorenz '96 system.
- Recovery of the parameters defining the prior and likelihood PDFs has been demonstrated using both synthetic data and data obtained using the Lorenz '96 system (results from the Lorenz '96 system not shown here).
- A theoretical justification for Hybrid DA systems, which linearly combine static and flow-dependent covariances, has been derived.
- Recovery and application of optimal weights of flow dependent and climatological variances has been demonstrated in both the Lorenz '96 and NAVDAS-AR hybrid systems

Acknowledgments: CHB gratefully acknowledges support from the U.S. Office of Naval Research Grant# 9846-0-1-5 . This research was performed while ES and DK both held National Research Council Research Associateship Awards