Heteroscedastic Ensemble Post-Processing

Liz Satterfield and Craig Bishop

Naval Research Laboratory, Monterey CA

ESPC & NUOPC Workshop

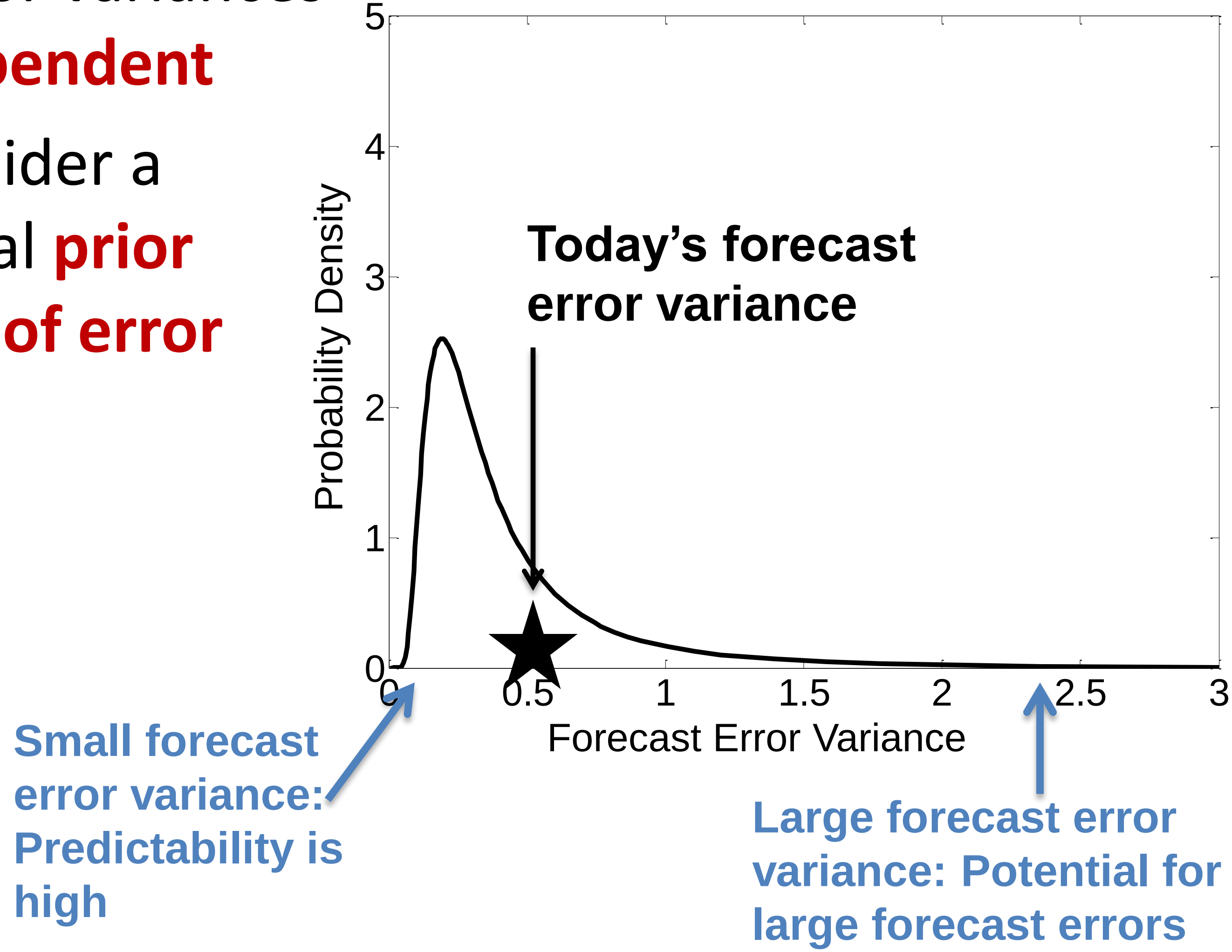
San Diego, CA 2013

Introduction

- Probabilistic predictions provide information that can increase the **socio-economic value of weather and climate forecasts**
- **Ensemble based probabilistic forecasts** of events are typically based on the relative frequencies of events in the ensemble
- Accurate **predictions of forecast error variances** are vital to probabilistic prediction

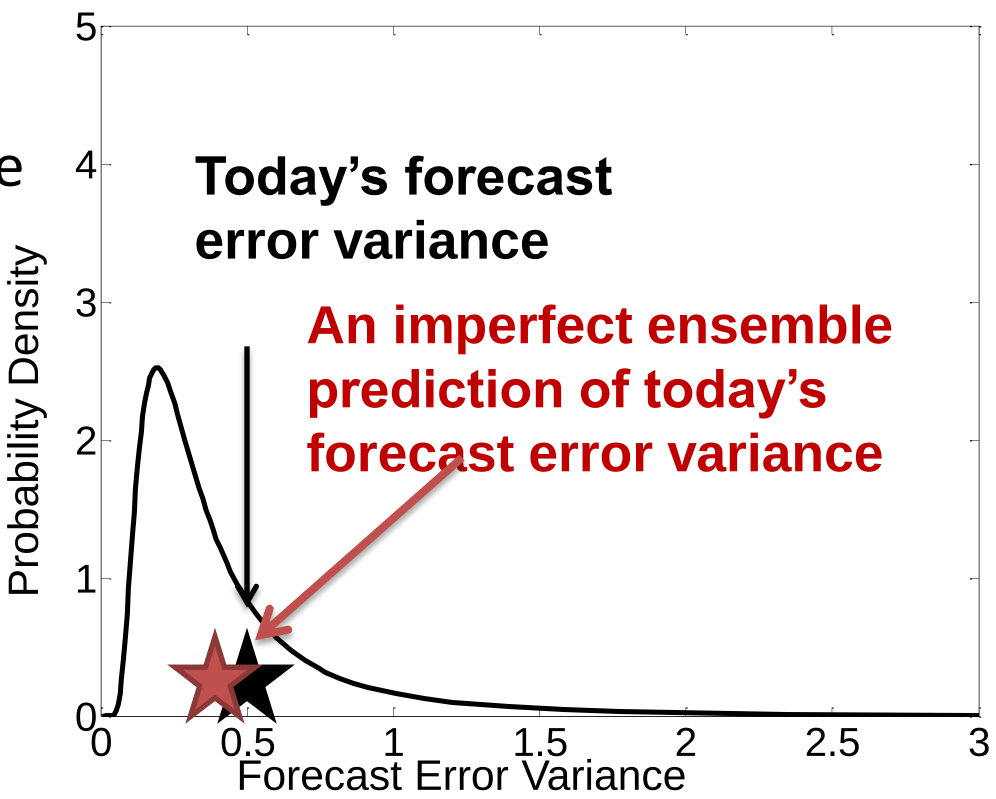
Forecast Error Variances

- Forecast error variances are **flow dependent**
- We can consider a climatological **prior distribution of error variances**



Ensemble Forecasts

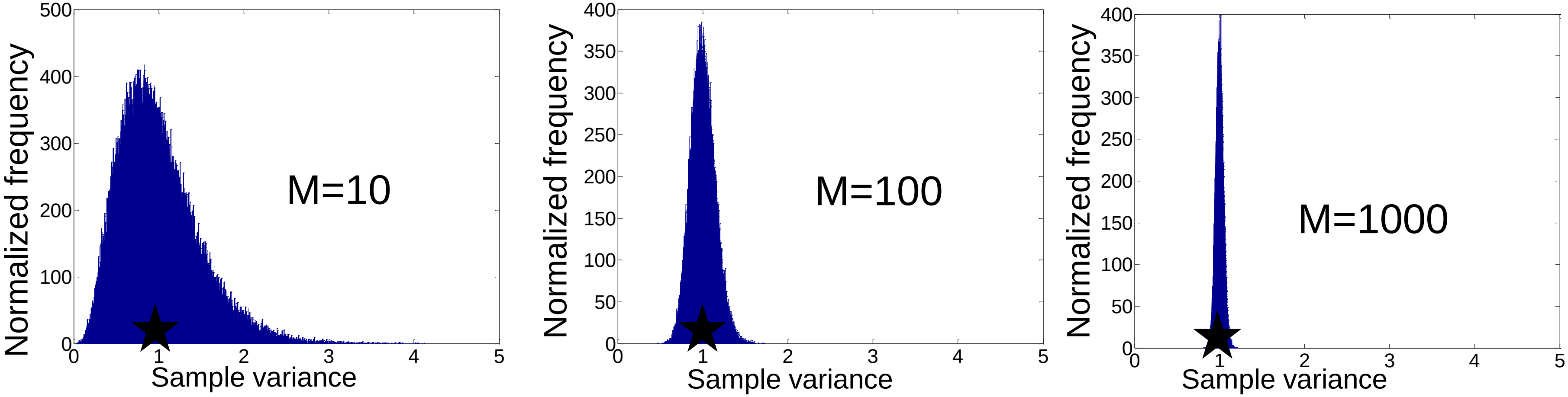
- Ensemble forecasts provide estimates of forecast error variance as a function of the **flow of the day**
- Ensemble **variances are imperfect**
  - Finite Size
  - Imperfect Model
  - Imperfect Initial Conditions



Inflating the ensemble perturbations can correct systematic under dispersion. How can we account for random fluctuations about the actual variance due to sampling error?

Distribution of sample variances

- Consider a normal distribution with mean zero and variance 1.
- Now consider the distribution of possible variances based on M-member draws from this distribution.

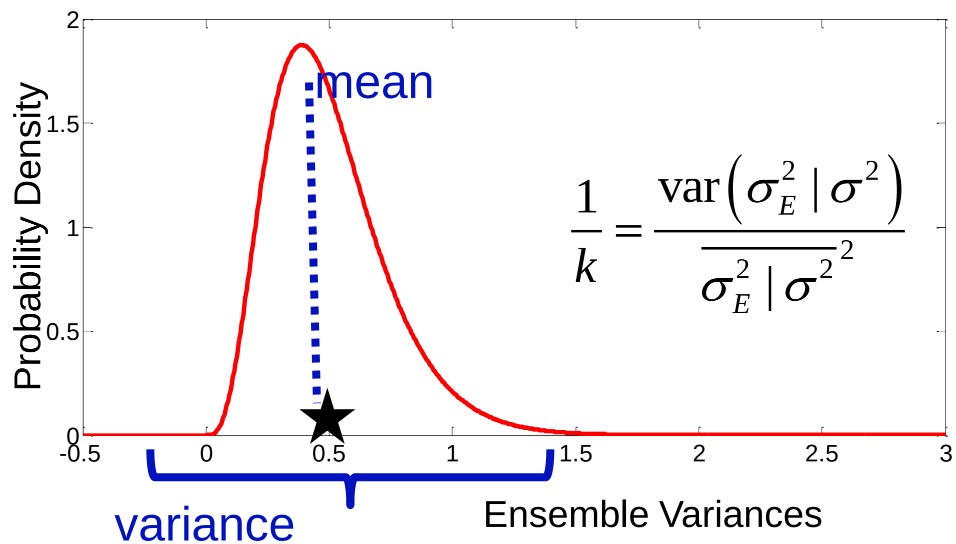


As the sample size increases the distribution of possible variance narrows

"Effective Ensemble Size"

- We want to account for random fluctuations due to sampling error in the ensemble.
- Because of deficiencies in the ensemble the ensemble size is not necessarily the sample size that we want to account for.
- In this sense, an 80 member ensemble could produce sampling errors like a 10 member ensemble.
- Theory (Bishop *et al.* 2013) shows how we can deduce this property, from long time series of pairs of innovations and ensemble variances, in terms of an ***effective ensemble size***.

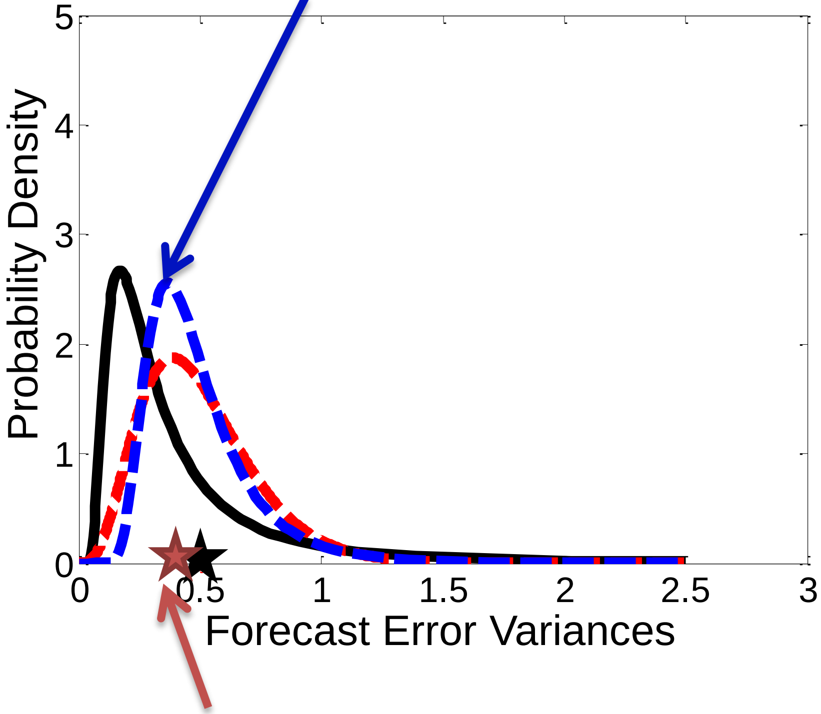
  - Effective ensemble size is inversely proportional the relative variance of the distribution of ensemble variances
  - We calculate this property in terms of an expected value over a time series



Distribution of forecast error variances given an ensemble prediction

- Recent work has developed an analytic approximation to a ***distribution of possible forecast error variances*** given an imperfect ensemble variance.
- Parameters defining such a distribution can be recovered from a long time series of pairs of innovations and ensemble variances
  - Bishop, Craig H., Elizabeth A. Satterfield, 2013: Hidden Error Variance Theory. Part I: Exposition and Analytic Model. *Mon. Wea. Rev.*, **141**, 1454–1468.
  - Bishop, Craig H., Elizabeth A. Satterfield, Kevin T. Shanley, 2013: Hidden Error Variance Theory. Part II: An Instrument That Reveals Hidden Error Variance Distributions from Ensemble Forecasts and Observations. *Mon. Wea. Rev.*, **141**, 1469–1483.

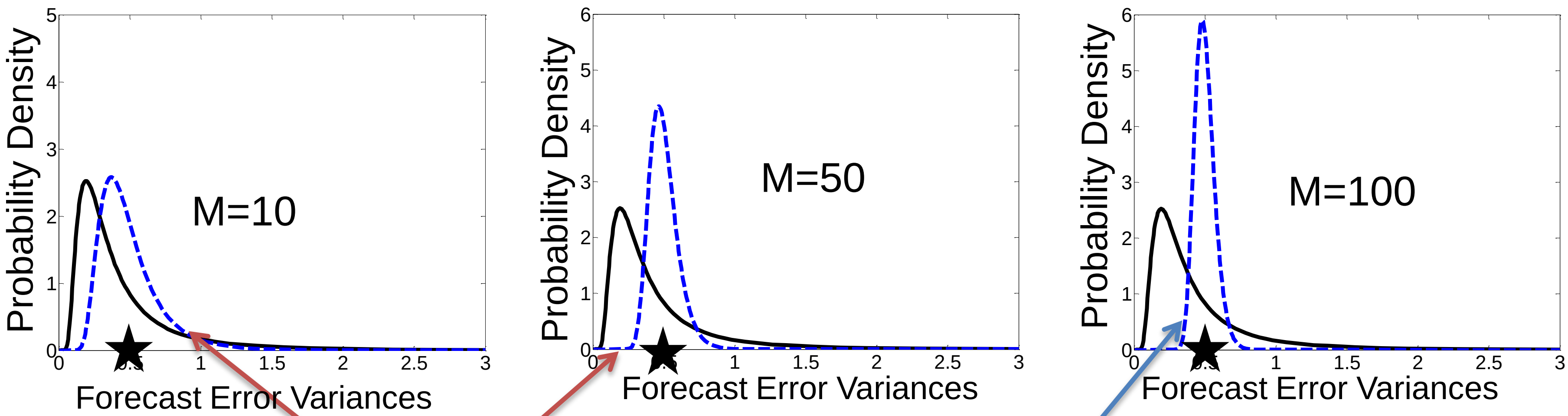
Posterior distribution of variances given an imperfect ensemble variance



An imperfect ensemble prediction of today's forecast error variance

"Effective Ensemble Size"

An effective post-processing scheme will account for stochastic fluctuations of the ensemble variance about today's actual forecast error variance.



Posterior is broad: Important to account for a distribution of variances

As the effective ensemble size increases the posterior distribution narrows and variance is better approximated by a single value

Heteroscedastic Post-Processing

How do we post-process in a way that is consistent with a distribution of variances?

We want to incorporate climatological (i.e. means, statistical prediction, persistence) information about our analyzed variable to re-center the mean

$$x_{post} = a_1 \overline{x^f} + a_2 \overline{x^c}$$

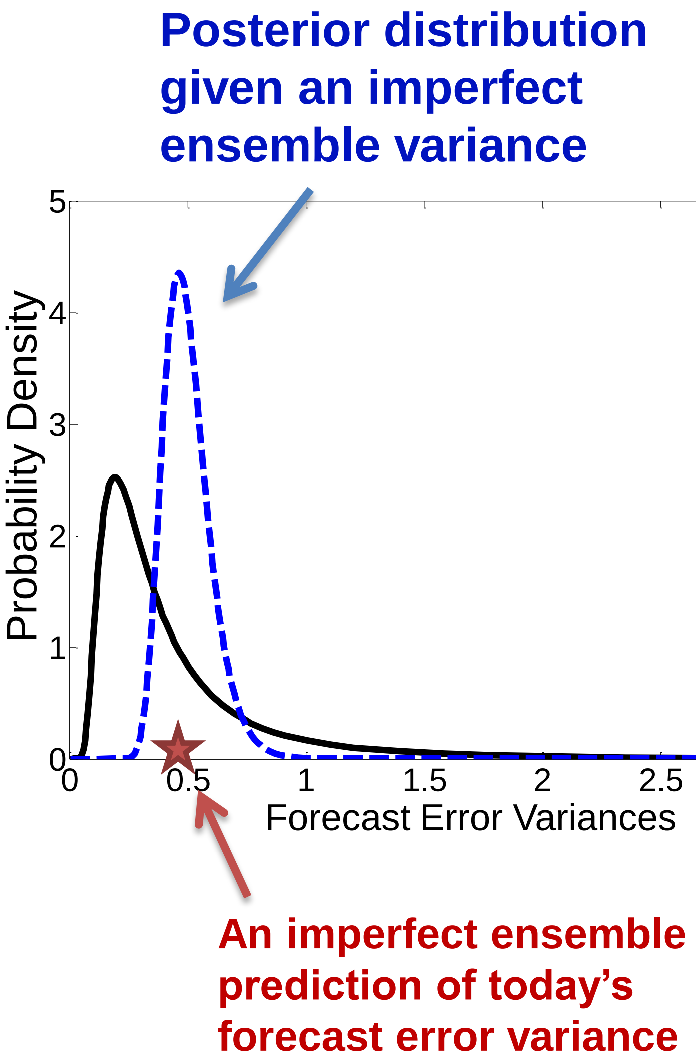
$$a_1 = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_f^2}, a_2 = \frac{\boxed{\sigma_f^2}}{\sigma_c^2 + \sigma_f^2}$$

If the forecast error variance is not a single value, but a distribution, the **degree to which climatological information should be incorporated within the ensemble is itself uncertain**

$$\sigma_{post}^2 = \frac{\sigma_f^2 \sigma_c^2}{\sigma_c^2 + \sigma_f^2}$$

Heteroscedastic Post-Processing

- **Step1**
  - Draw a variance from the posterior distribution



Heteroscedastic Post-Processing

- **Step1**
  - Draw a variance from the posterior distribution
- **Step 2**
  - Use that variance along with climatological information to re-center the forecast as a weighted combination of the climatological mean and the ensemble mean

$$x_j^{post} = \left[\frac{\sigma_C^2}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^f}\right) + \frac{\left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^C}\right) \right] + \zeta_j$$

$$j = 1, 2, .., M$$

$$\zeta_j \sim N(0, \sigma^2{}_P)$$

$$\sigma_p^2 = \frac{\sigma_C^2 \left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2}$$

Weighting given to climate mean is dependent on the value of error variance, which is randomly drawn from the posterior

Heteroscedastic Post-Processing

- Step1
  - Draw a variance from the posterior distribution
- Step 2
  - Use that variance along with climatological information to re-center the forecast as a weighted combination of the climatological mean and the ensemble mean
- **Step 3**
  - Use the variance drawn from the posterior along with climatological information to perturb the re-centered forecast

$$x_j^{post} = \left[\frac{\sigma_C^2}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^f}\right) + \frac{\left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^C}\right) \right] + \zeta_j$$

$$j = 1, 2, .., M$$

$$\zeta_j \sim N(0, \sigma^2{}_P)$$

$$\sigma_p^2 = \frac{\sigma_C^2 \left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2}$$

The distribution from which the perturbations are drawn has a variance that changes with each draw from the posterior

Heteroscedastic Post-Processing

- **Step1**
  - Draw a variance from the posterior distribution
- **Step 2**
  - Use that variance along with climatological information to re-center the forecast as a weighted combination of the climatological mean and the ensemble mean
- **Step 3**
  - Use the variance drawn from the posterior along with climatological information to perturb the re-centered forecast
- **Step 4**
  - Repeat until a sufficient ensemble size is achieved. We will refer to this ensemble as “Fully Processed”.

$$x_j^{post} = \left[\frac{\sigma_C^2}{\left(\sigma_S^2\right)_j + \sigma_C^2}\left(\overline{x^f}\right) + \frac{\left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2}\left(\overline{x^C}\right) \right] + \zeta_j$$

$$j = 1, 2, .., M$$

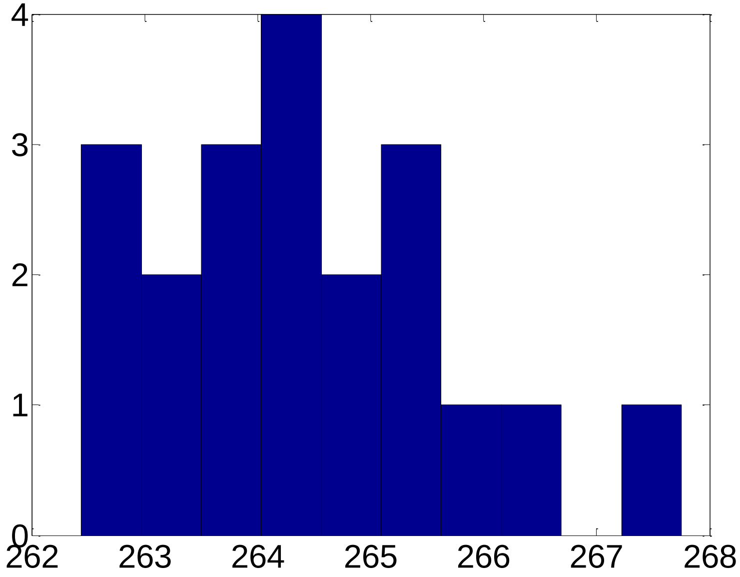
$$\zeta_j \sim N(0, \sigma^2{}_P)$$

$$\sigma_p^2 = \frac{\sigma_C^2 \left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2}$$

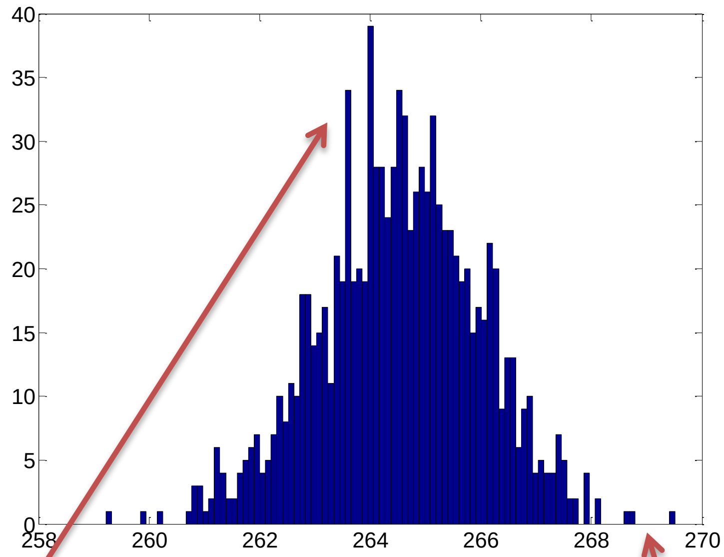
For each ensemble member this variance is drawn from the posterior

Heteroscedastic Post-Processing

72-hr forecast of virtual temperature at 500hPa



Original 20 member ensemble



1000 member post-processed ensemble

Higher kurtosis

Broader tails

"Straw Man" Homoscedastic Methods

Use the same post-processing algorithm to adjust the mean and add a perturbation.

This time, <u>use a single value of variance (homoscedastic) associated with each forecast</u>

$$x_j^{post} = \left[\frac{\sigma_C^2}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^f}\right) + \frac{\left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2} \left(\overline{x^C}\right) \right] + \zeta_j$$

$$j = 1, 2, .., M$$

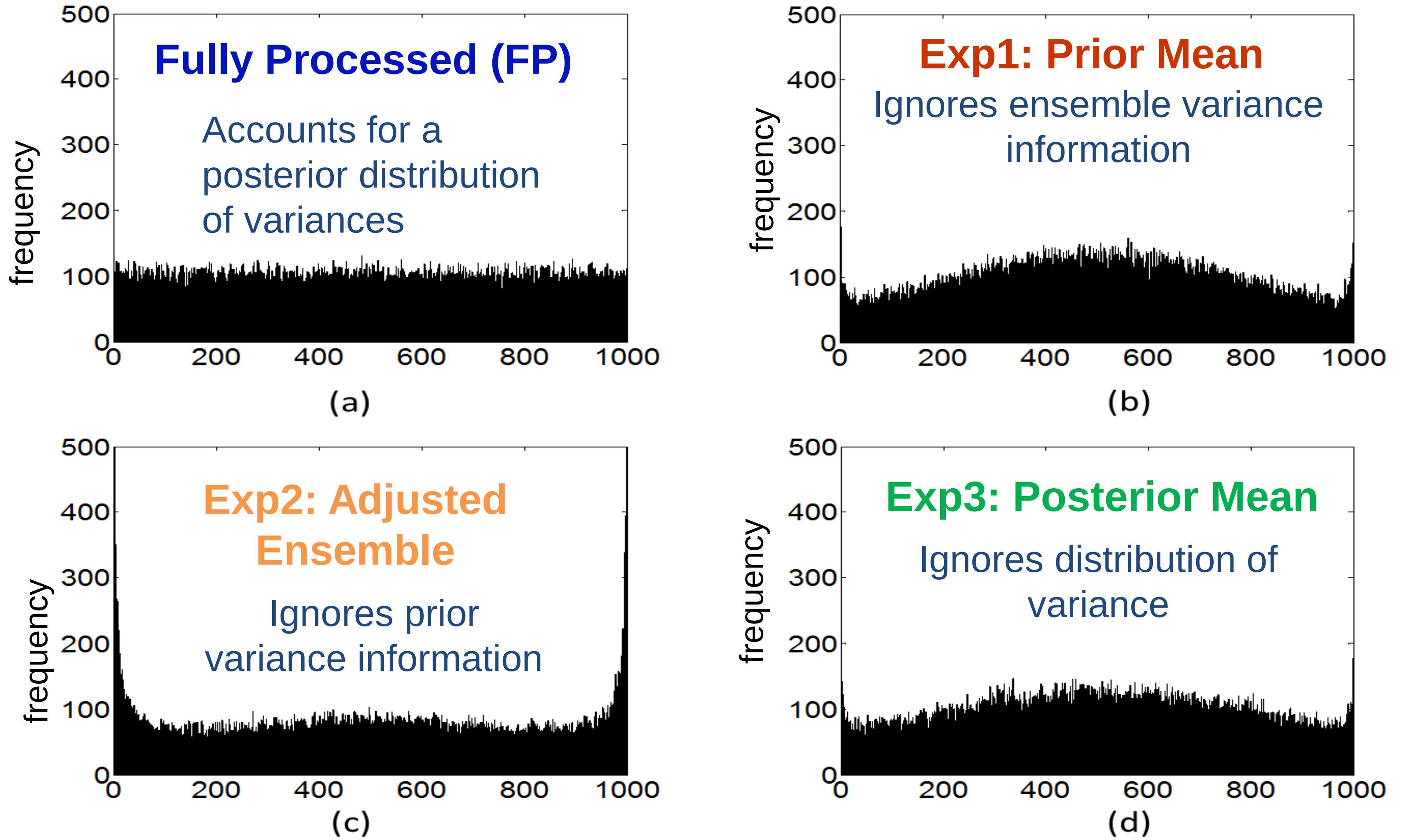
$$\zeta_j \sim N(0, \sigma^2{}_P)$$

$$\sigma_p^2 = \frac{\sigma_C^2 \left(\sigma_S^2\right)_j}{\left(\sigma_S^2\right)_j + \sigma_C^2}$$

We define variances in three different ways, each with differing amounts of information:

1. **Exp1: Prior Mean**
   - Ignore ensemble prediction of variance
   - Variance is the same for every forecast
2. **Exp2: Adjusted Ensemble**
   - Use only the ensemble prediction of the forecast error variance
   - Variances are inflated or attenuated to be consistent with innovation statistics
   - Ignore information from the prior distribution of variances
3. **Exp3: Posterior Mean**
   - Use the mean of the posterior distribution.
   - Ignore the variable nature of the forecast error variance

Synthetic Data Results
Rank Frequency Histograms



Even though all ensembles have the correct variance on average, only the FP ensemble accounts for the variable nature of possible error variances.

Weather Roulette (Hagedorn and Smith, 2008)

- Player **A** opens up a weather roulette casino
- Player **A** creates a roulette wheel where each slot is an equally probable state based on climatology
- Player **A** sets the odds in each slot based on the FP ensemble
- Player **B** places bets based on their ensemble
- Each starts with a $1 bet and reinvests their winnings every round
- Value of the ensemble is interpreted using an ***effective daily interest rate*** *(the rate at which **A** goes broke or gets rich!)*

Synthetic Data Results – Effective Daily Interest Rate

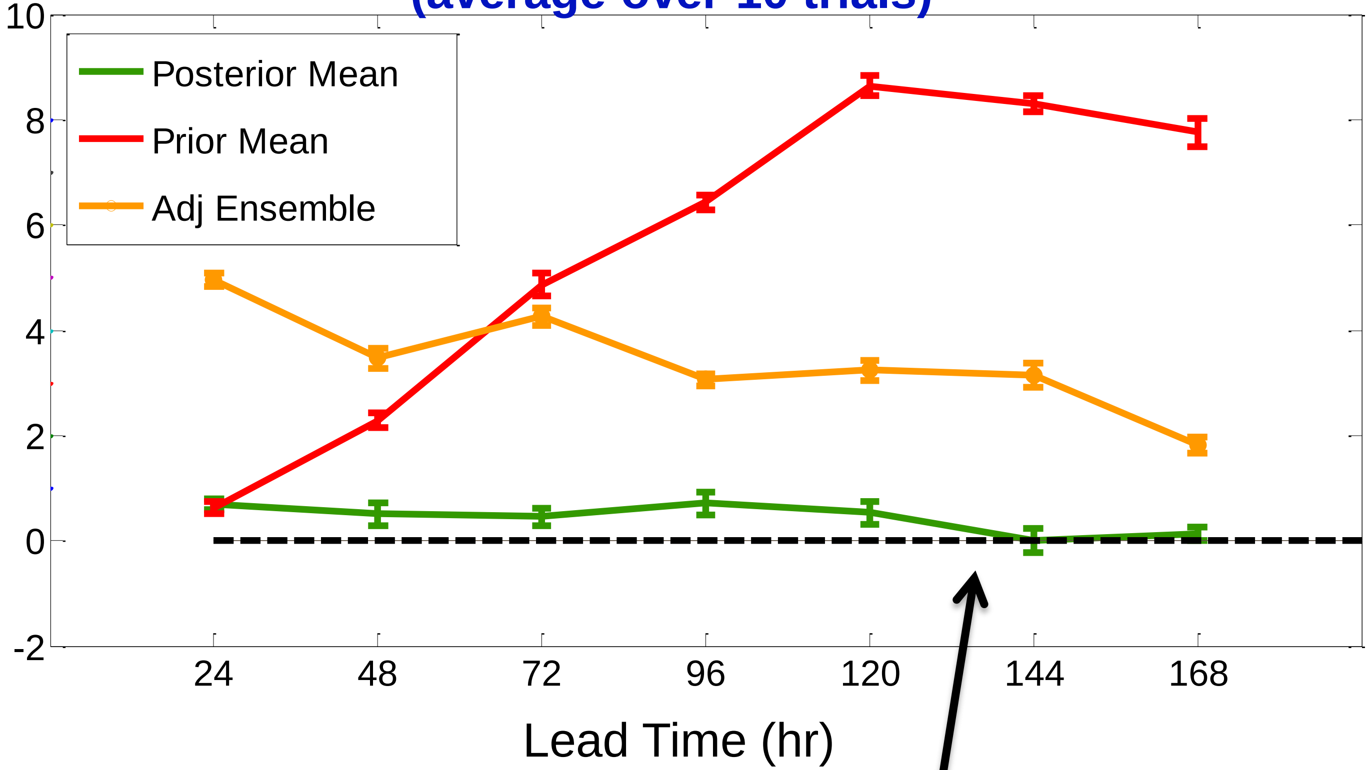
Weather Roulette Interest Rate (%) Earned by the FP Ensemble when "straw man" ensembles are played (average over 10 trials)

Effective Ensemble Size	M=2	M=4	M=6	M=8	M=10
Exp1: Prior Mean	3.95	6.31	7.66	8.50	9.11
Exp2: Adjusted Ensemble	69.90	13.49	5.56	3.06	1.88
Exp3: Posterior Mean	1.06	0.44	0.16	0.08	0.04

As ensemble size increases, the posterior distribution narrows.

500mb virtual temperature
May 2012 FNMOC post-processed ensemble

Weather Roulette Interest Rate (%) Earned by the FP Ensemble when "straw man" ensembles are played (average over 10 trials)



Accounting for a distribution shows benefit up to 144-hr leads

Exp1: Prior Mean

FP ensemble wins more money as the lead time increases and the distribution of the prior broadens.

Exp2: Adjusted Ensemble

The effective ensemble size increases with lead time, reducing the amount the FP ensemble wins.

Exp3: Posterior Mean

Combining ensemble and prior estimates of variance is optimal.

The FP wins money until the posterior distribution becomes sufficiently narrow

Application to intraseasonal to interannual time scales

- Forecasts may benefit from the inclusion of non-NWP information (climatological means, etc.)
- The extent to which this information should be included is itself uncertain.
- Accounting for the variable nature of variances matters when:
  - Distribution of prior (climatological) error variances is broad
  - Ensemble based variance prediction is uncertain (small effective ensemble size)
- The recovery of parameters that define the posterior distribution depends on a series of ensemble and innovation pairs
  - Archives of observations, hindcasts

Main Conclusions

- We have presented a new post-processing scheme which accounts for a distribution of possible variances given an imperfect ensemble predictions
- Bishop *et al* 2013 introduce a new diagnostic, the *effective ensemble size*, which measures the ability of the ensemble to track fluctuations in actual error variance.
- Application to synthetically generated data and 500hPa forecasts of virtual temperature from the operational FNMOC ensemble demonstrate that accounting for the variable nature of forecast error variances leads to improved probabilistic skill scores.

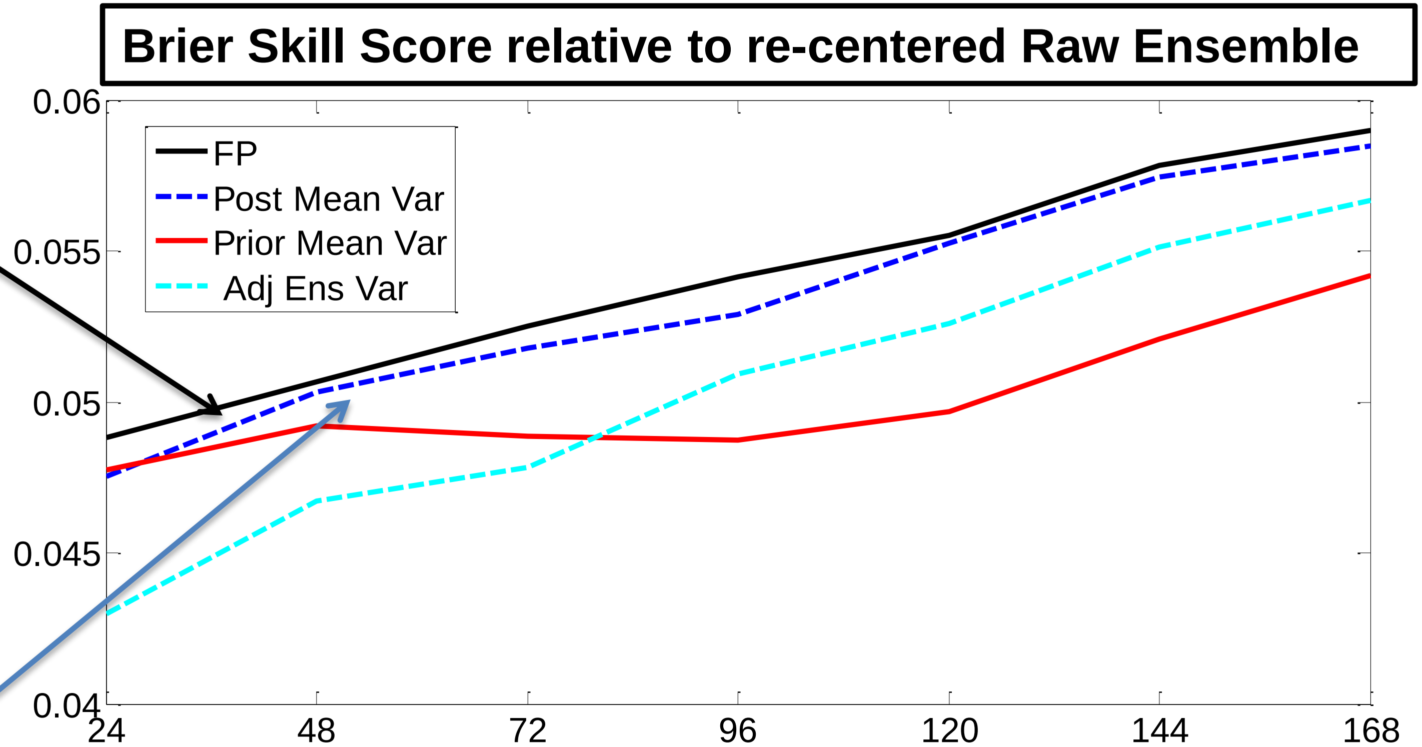
THANK YOU

EXTRA SLIDES

Brier Skill Score

FP is the only method that accounts for a range of variances, which is important until posterior narrows at 144-hrs

Both FP and Posterior Mean weight variance information from ensemble and prior



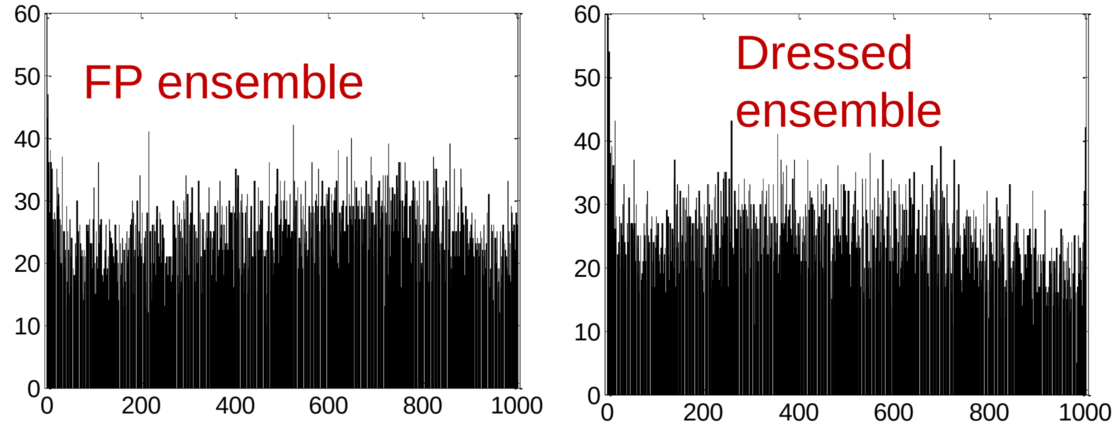
- A Brier score was computed for 100 equally likely climatological bins
- The Brier Score was then averaged over all bins
- Shown here is the associated Brier Skill Score where the (re-centered on deterministic) raw ensemble is used as a reference value
- Results shown are averaged over 10 trials each with a random selection of 5000 observations

Comparison to Wang and Bishop dressing method

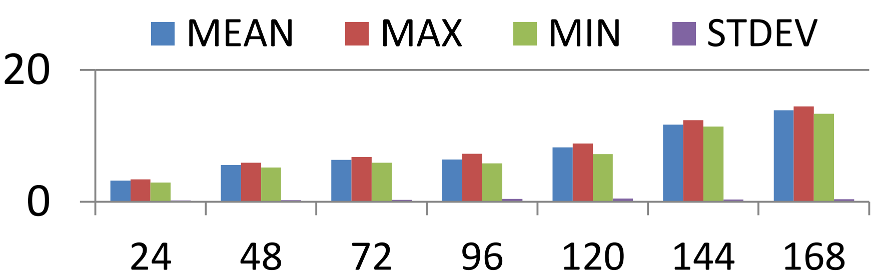
Wang and Bishop (2005) Dressing Method

- Spread is corrected by adding a perturbations to raw ensemble members
- Perturbation is taken from normal distribution with variance derived from innovation statistics
- A larger ensemble can be formed by dressing each raw ensemble member multiple times

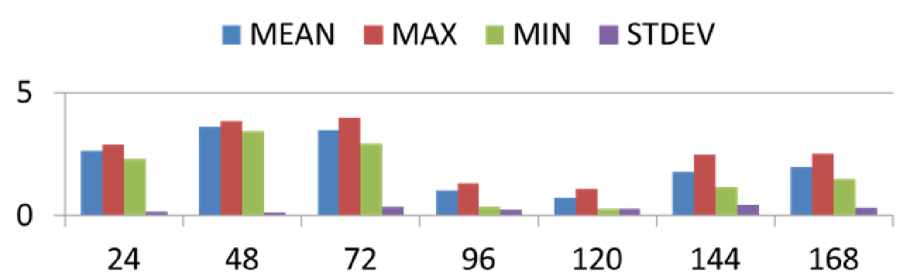
RFH 96-hr lead time 500hPa virtual temperature



Re-center on Deterministic



Weight Climatology and Deterministic



Effective Ensemble Size Derivation

the mean of the distribution of j ensemble variances for a particular σ_i^2 at some spatial-temporal point i :

$$\left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j = a(\sigma_i^2 - \sigma_{min}^2)$$

the mean of the distribution of ensemble variances over some spatial-temporal series,

$$\left\langle \left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j \right\rangle_i = a\left[\left\langle \sigma_i^2 \right\rangle - \sigma_{min}^2\right]$$

the square of the mean over some spatial-temporal series,

$$\left\langle \left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j^2 \right\rangle_i = a^2 \left\langle (\sigma_i^2 - \sigma_{min}^2)^2 \right\rangle_i = a^2 \left\langle (\left\langle \sigma_i^2 \right\rangle - \sigma_{min}^2 + \sigma_i'^2)^2 \right\rangle_i$$

$$= a^2 \left(\left\langle \sigma^2 \right\rangle - \sigma_{min}^2\right)^2 + a^2 \operatorname{var}(\sigma^2)$$

the variance of the ensemble variances given σ_i^2 over some spatial-temporal series,

$$\left\langle \operatorname{var}\left(s_j{}^2 \mid \sigma_i^2\right)\right\rangle_i = \left\langle \left\langle \left(s_j{}^2 - s^2{}_{min} \mid \sigma_i^2\right)^2 \right\rangle \right\rangle_i - \left\langle \left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j^2 \right\rangle_i =$$

$$= \left\langle \left\langle \left(s_j{}^2 - s^2{}_{min} \mid \sigma_i^2\right)^2 \right\rangle - \left\langle \left(s_j{}^2 - s^2{}_{min} \mid \sigma_i^2\right) \right\rangle^2 \right\rangle_i - a^2 \operatorname{var}(\sigma^2)$$

$$= \left\langle \operatorname{var}\left(s_j{}^2 - s^2{}_{min} \mid \sigma_i^2\right)\right\rangle_i - a^2 \operatorname{var}(\sigma^2)$$

relative variance is

$$\left\langle \frac{\operatorname{var}\left(s^2 \mid \sigma_i^2\right)}{\left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j^2} \right\rangle_i = \frac{\left\langle \operatorname{var}\left(s^2 \mid \sigma_i^2\right)\right\rangle_i}{\left\langle \left\langle s_j{}^2 - s^2{}_{min} \mid \sigma_i^2 \right\rangle_j^2 \right\rangle_i} = \frac{\left\langle \operatorname{var}\left(s_j{}^2 - s^2{}_{min} \mid \sigma_i^2\right)\right\rangle_i - a^2 \operatorname{var}(\sigma^2)}{a^2\left[\left(\left\langle \sigma^2 \right\rangle - \sigma_{\min}^2\right)^2 + \operatorname{var}\left(\sigma^2\right)\right]} = \frac{1}{k}$$

The squared mean of the distribution of ensemble variances given a true error variance

Variance of ensemble variances given a true error variance

Parameter Recovery

- In order to apply this post-processing technique to real data, we must obtain the parameters defining the *climatological prior, likelihood, and posterior*
- Bishop et al (2012) show how these parameters can be deduced from time series of innovations and ensemble variance predictions (v , s^2)

Parameters describing the climatological distribution of variances

1. $\langle \sigma^2 \rangle = \langle v^2 - R \rangle$

2. $\mathrm{var}(\sigma^2) = \left(\langle \sigma^2 \rangle + \langle R \rangle\right)^2 \left[\frac{kurtosis(v) - 3}{3} \right] - \mathrm{var}(R)$

3. $a = \frac{\mathrm{covar}(v^2, s^2)}{\mathrm{var}(\sigma^2)}$

Inflates or attenuates ensemble variances

4. $\sigma^2_{\min} = \langle \sigma^2 \rangle - \frac{\langle s^2 \rangle - s^2_{\min}}{a}$

5. $k^{-1} = \frac{\mathrm{var}(s^2) - a^2\,\mathrm{var}(\sigma^2)}{a^2\left[\left(\langle \sigma^2 \rangle - \sigma^2_{\min}\right)^2 + \mathrm{var}(\sigma^2)\right]}$

relative variance of the ensemble variances