Ensemble-based estimates of information content in observational data extractable by data assimilation methods

Dusanka Zupanski
CIRA/Colorado State University
Fort Collins, Colorado

EMC Predictability Meeting
Camp Springs, MD, January 24, 2006

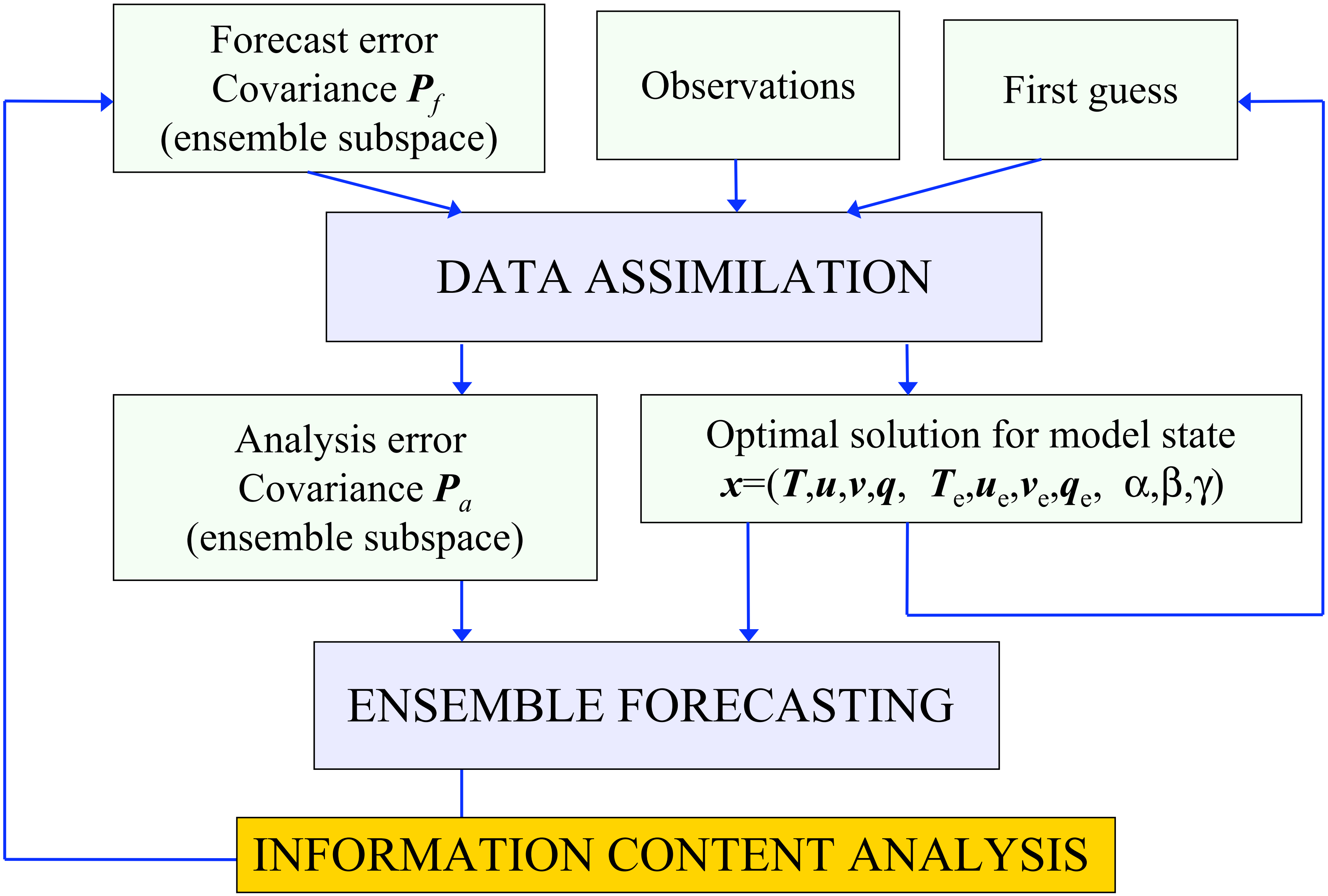
Collaborators
- *S. Denning, M. Uliasz, R. Lokupitiya, L. Grasso, M. DeMaria, and M. Zupanski (Colorado State University)*
- *A. Y. Hou, S. Zhang (NASA/GMAO)*

OUTLINE

- Information measures in ensemble subspace
- Experimental results employing various dynamical models
- Conclusions
- Future plans

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Ensemble Data Assimilation



Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Information measures in ensemble subspace

(Bishop et al. 2001; Wei et al. 2005; Zupanski et al. 2005, 2006)

$$\boldsymbol{C} = \boldsymbol{Z}^T \boldsymbol{Z}$$

$\boldsymbol{C}$ **- information matrix in ensemble subspace of dim** $Nens$ **x** $Nens$

$$z^i = R^{-1/2} H[M(x + p_f^i)] - R^{-1/2} H[M(x)]$$

z^i **- are columns of Z**

$$\boldsymbol{x} - \boldsymbol{x}_b = \boldsymbol{P}_f^{1/2} (\boldsymbol{I} + \boldsymbol{C})^{-1/2} \zeta$$

ζ **- control vector in ensemble space of dim** $Nens$

$\boldsymbol{x}$ **- model state vector of dim** $Nstate >> Nens$

Degrees of freedom (DOF) for signal *(Rodgers 2000)*:

$$d_s = tr[(\boldsymbol{I} + \boldsymbol{C})^{-1} \boldsymbol{C}] = \sum_i \frac{\lambda_i^2}{(1 + \lambda_i^2)}$$

λ_i^2 **- eigenvalues of C**

Shannon information content,
or entropy reduction

$$h = \frac{1}{2} \sum_i \ln(1 + \lambda_i^2)$$

Errors are assumed Gaussian in these measures.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Basic characteristics of Maximum Likelihood Ensemble Filter (MLEF)

(Zupanski 2005; Zupanski and Zupanski 2006)

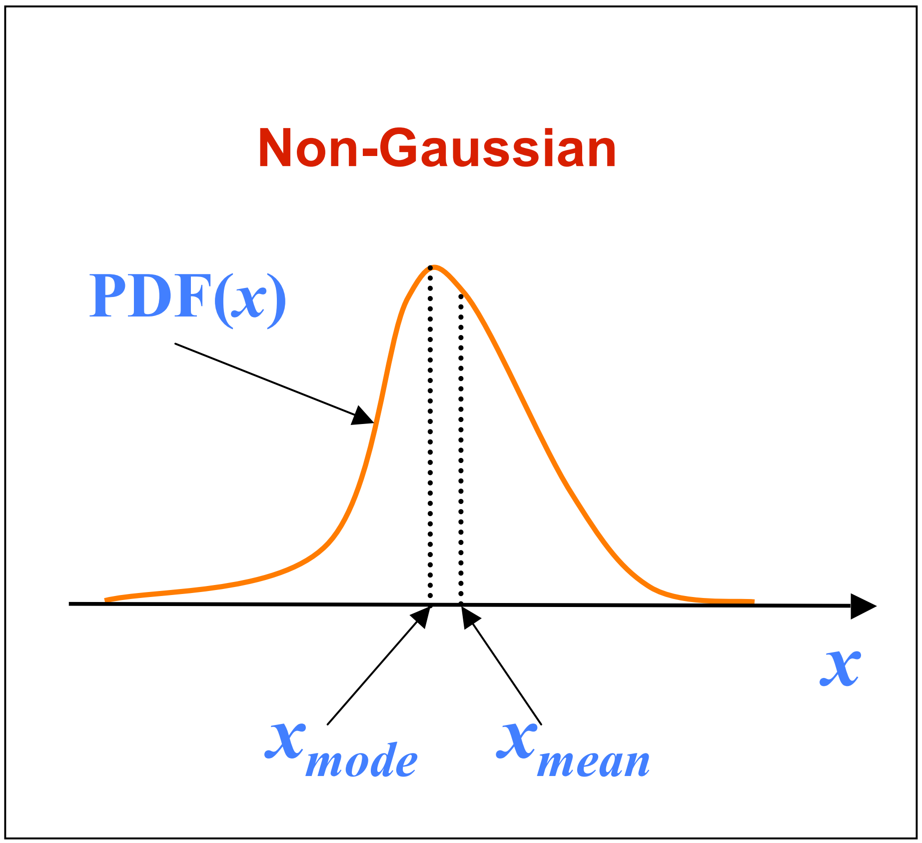
- **MLEF is similar to 4dvar because it seeks a maximum likelihood solution (i.e., minimum of J).**
- **It is also similar to EnKF methods because it uses ensembles to calculate forecast error covariance.**
- **MLEF uses the same definition of matrix C as in the ETKF (Bishop et al. 2001).**
- **It has a built-in capability to estimate and reduce several major sources of forecast uncertainties simultaneously: Initial conditions, model error, boundary conditions, and empirical parameters.**

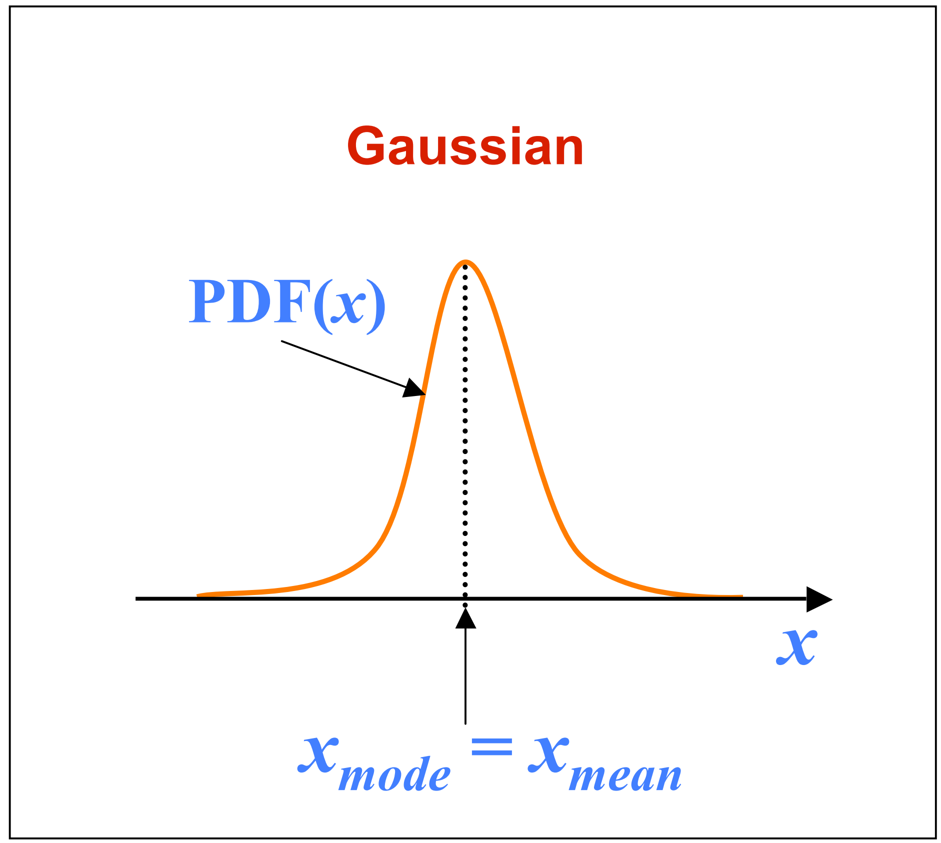
Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

MODE vs. MEAN

MLEF involves an iterative minimization of functional *J* $\Rightarrow x_{mode}$

Minimum variance methods (EnKF) calculate ensemble mean $\Rightarrow x_{mean}$





Different results expected for non-Gaussian PDFs

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Experiments

Atmospheric models:

- **GEOS-5** single column model: Assimilation of T and q
  (In collaboration with Athur Hou and Sara Zhang, NASA)

- **RAMS** model: Assimilation of u,v,w,p,th, and r
  (In collaboration with Louie Grasso and Mark DeMaria, CSU)

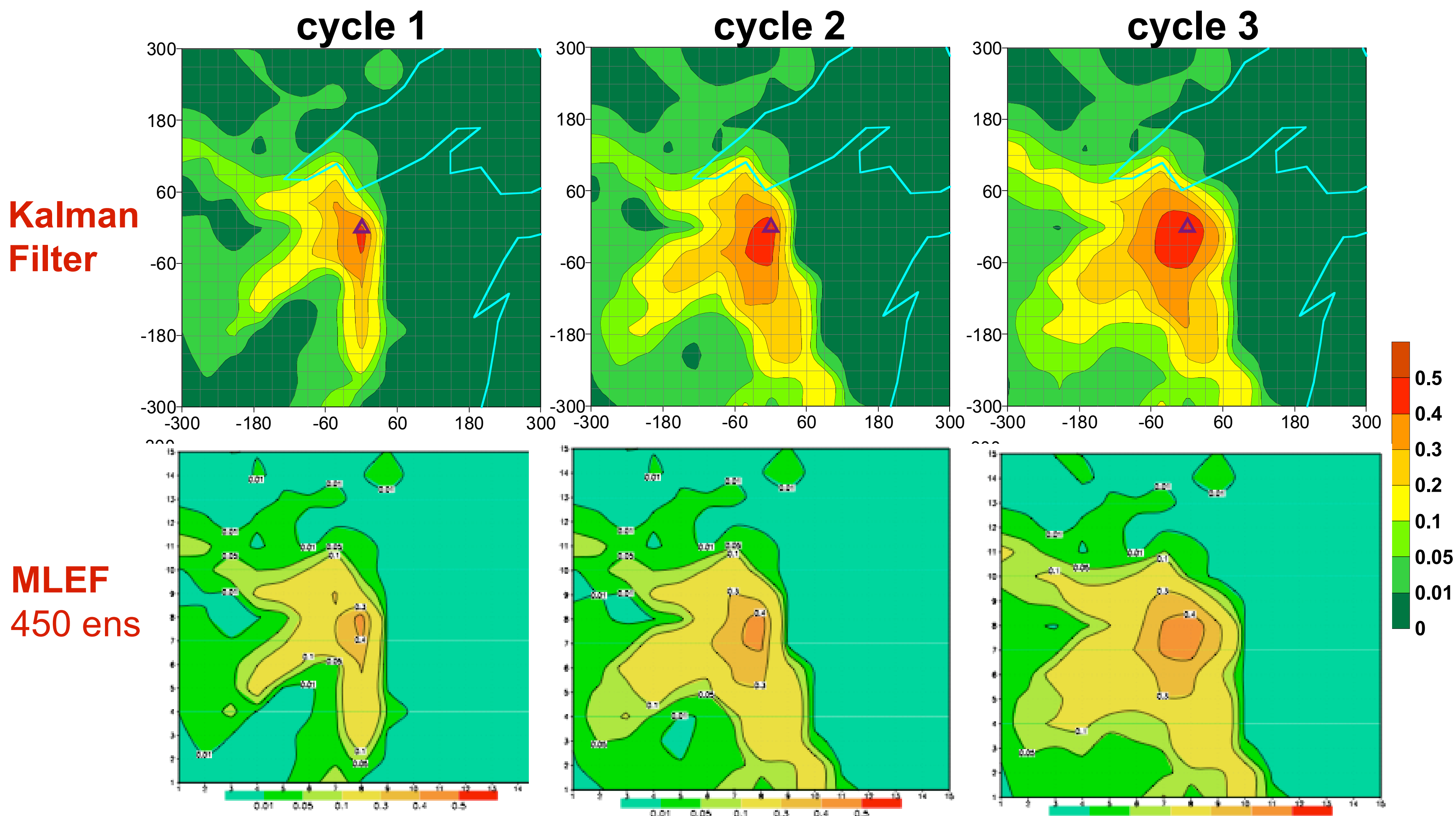
Carbon transport models:

- **PCTM** model: Global CO2-flux inversion (estimation of weekly CO2-fluxes)
  (In collaboration with Scott Denning and Ravi Lokupitiya, CSU)

- **LPDM** model: Regional (mesoscale) CO2-flux inversion (estimation of model bias in daily CO2-fluxes)
  (In collaboration with Scott Denning, Marek Uliasz and Andrew Schuh, CSU, and Peter Rayner, CEA/LSCE France)

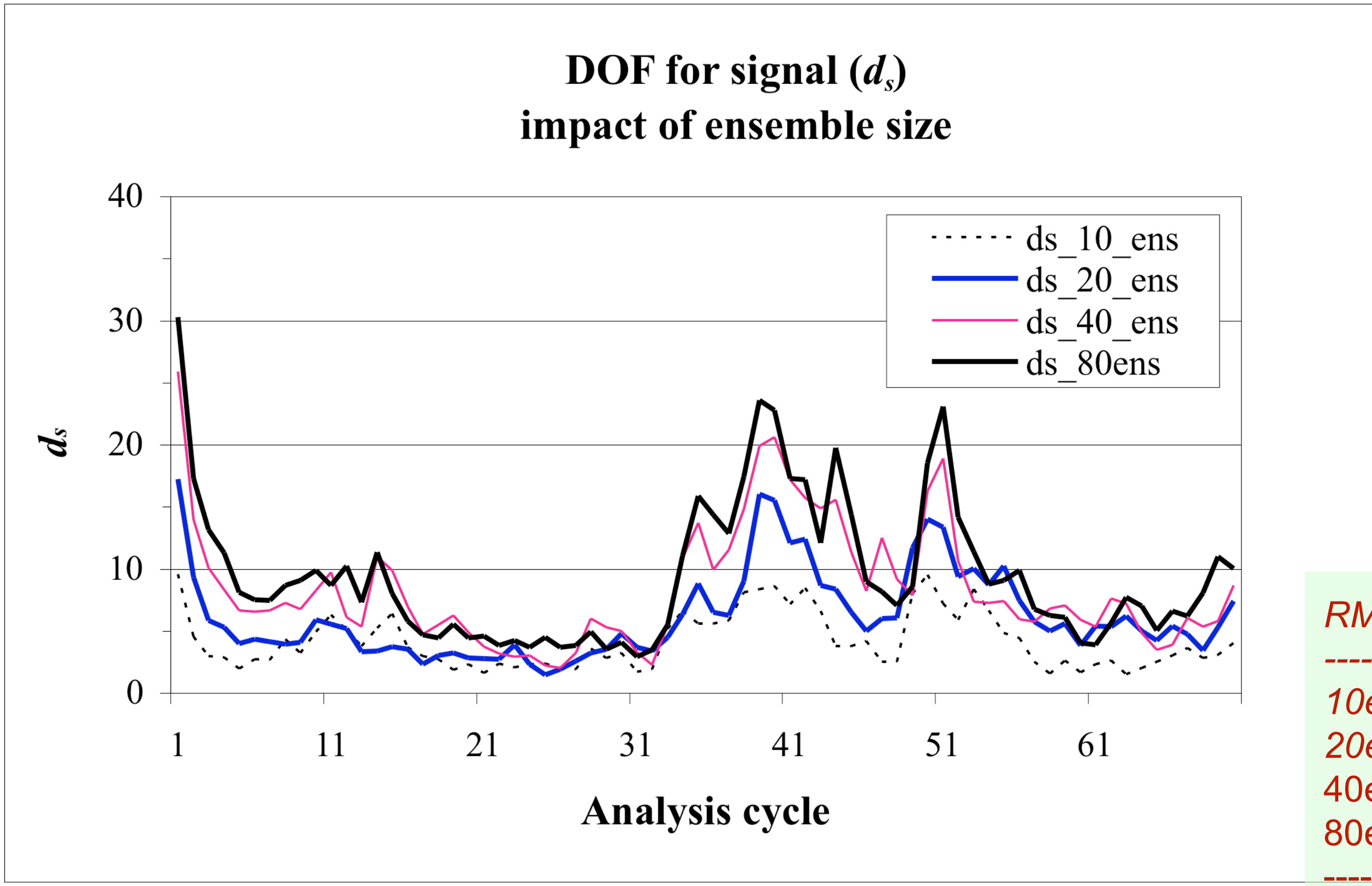
Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

LPDM model: Estimation of respiration bias
Reduction of uncertainty (σ_0-σ), Nstate=450, Nobs=600, three 5-day data assimilation cycles



This is a sanity check of the full-rank MLEF solution: it is equal to the Kalman filter solution for linear models (e.g., LPDM model).

GEOS-5 Single Column Model: DOF for signal (Nstate=80; Nobs=80, seventy 6-h DA cycles, assimilation of simulated T,q observations)



DOF for signal varies from one analysis cycle to another due to changes in atmospheric conditions.

RMS Analysis errors for T, q:

10ens ~ 0.50K; 0.566g/kg

20ens ~ 0.32K; 0.462g/kg

40ens ~ 0.27K; 0.417g/kg

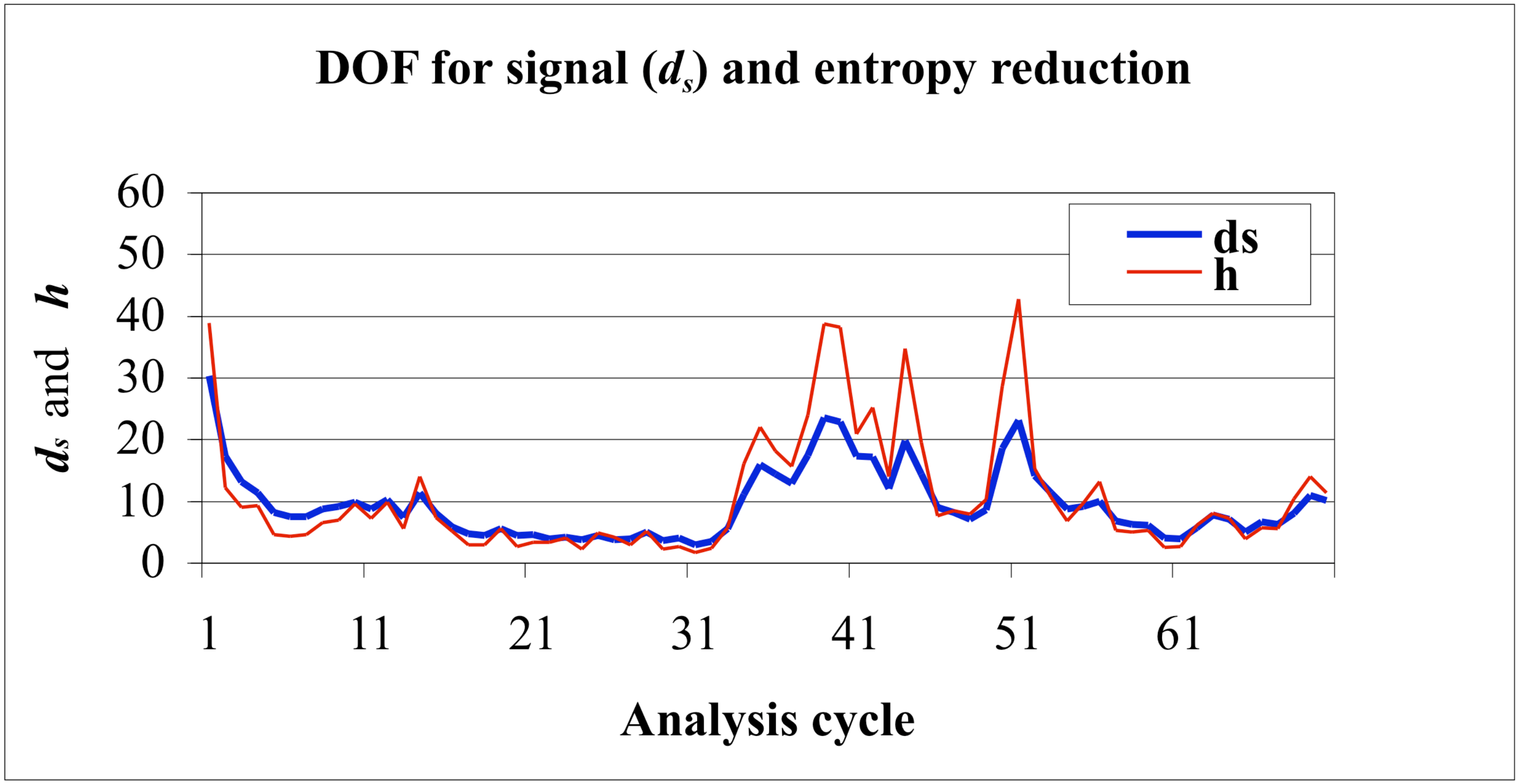
80ens ~ 0.20K; 0.362g/kg

No_obs ~ 0.82K; 0.656g/kg

Small ensemble size (10 ens), even though not perfect, captures main data signals.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

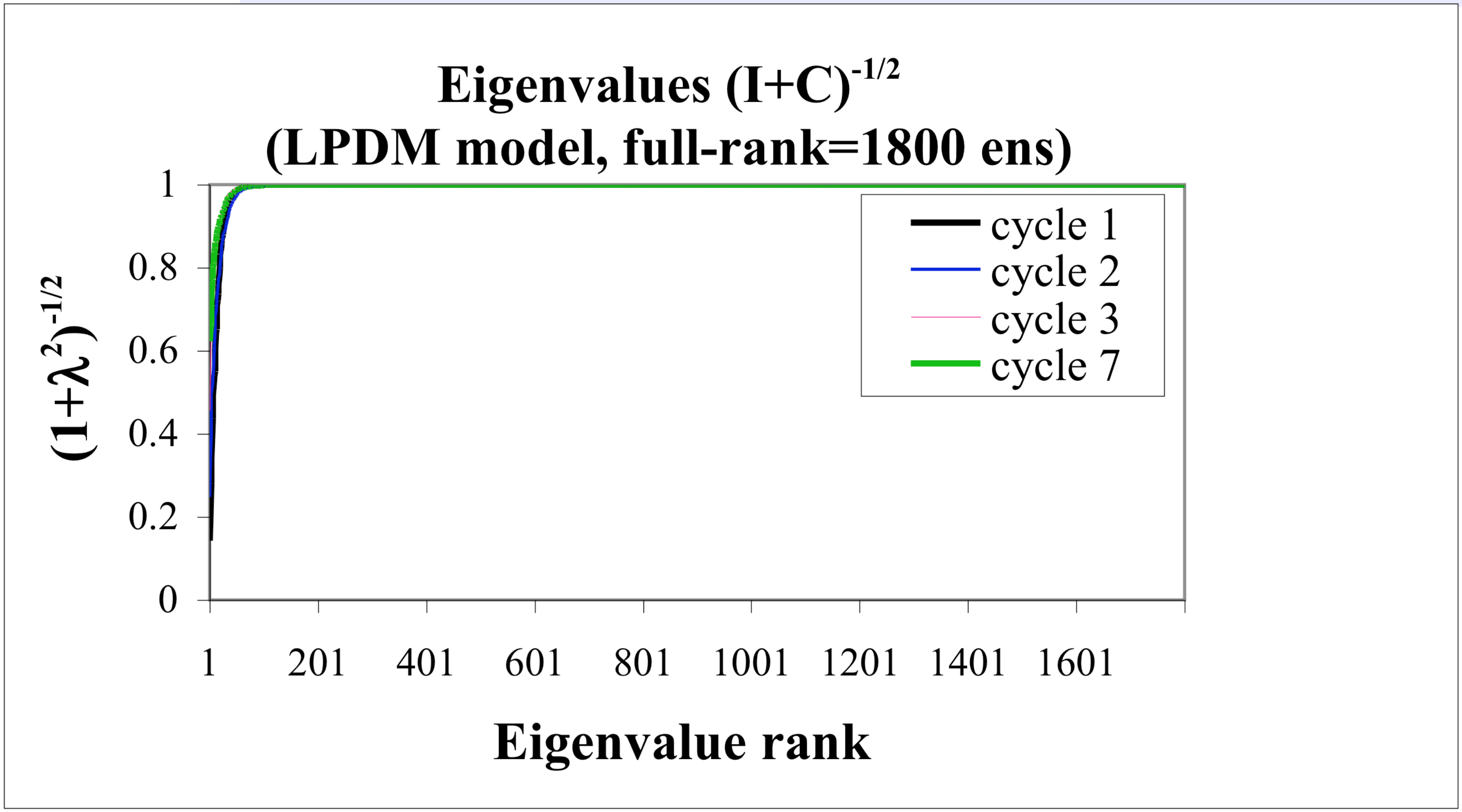
GEOS-5 Single Column Model: DOF for signal (Nstate=80; Nobs=80, seventy 6-h DA cycles, assimilation of simulated T,q observations)



DOF for signal and entropy reduction are very similar information measure. Main difference: the valued of DOF are always ≤ Nens.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

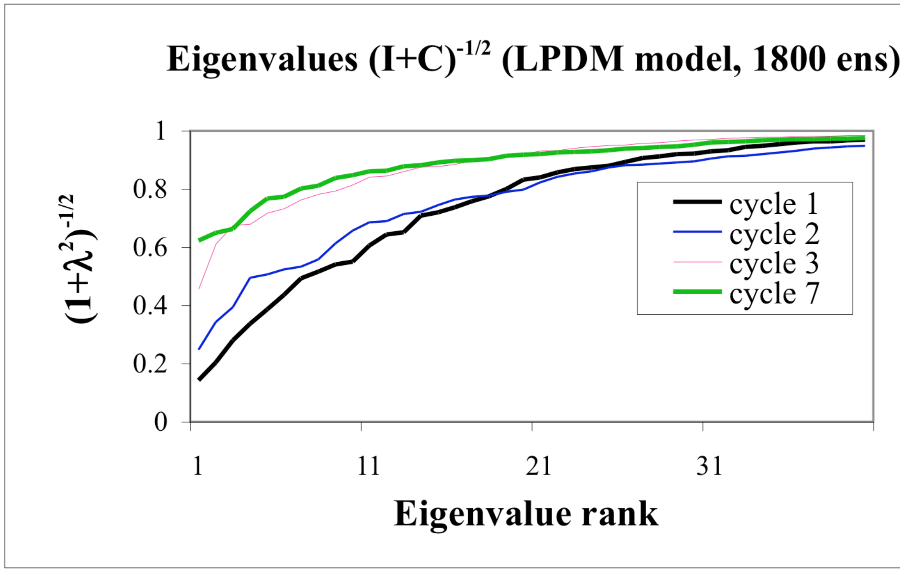
LPDM Model CO2-flux BIAS estimation: Eigenvalue spectrum of $(I+C)^{-1/2}$ (Nstate=1800; Nobs=1200, Nens=1800, seven 10-day DA cycles, assimilation of simulated C02 observations from a tall tower)

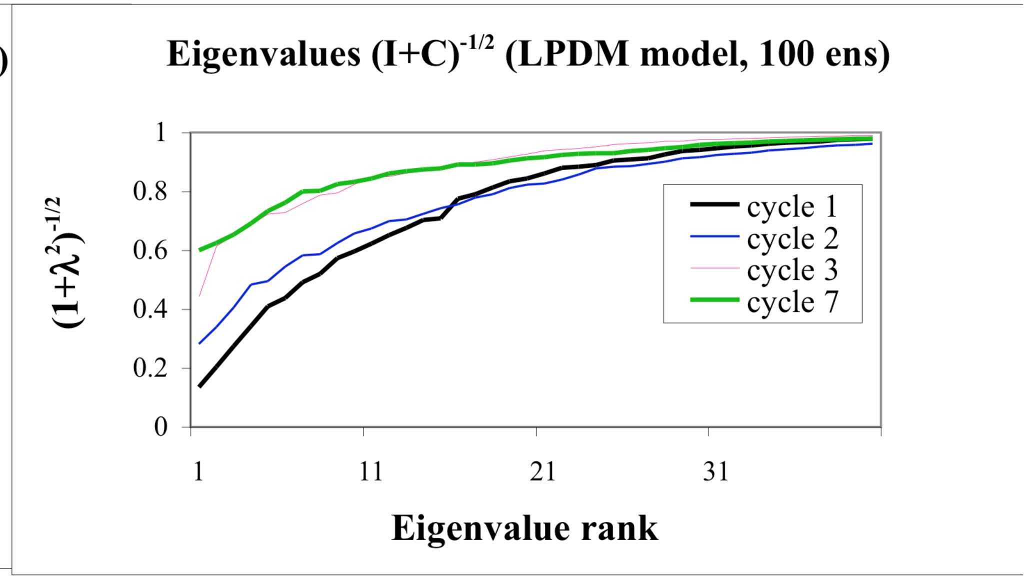


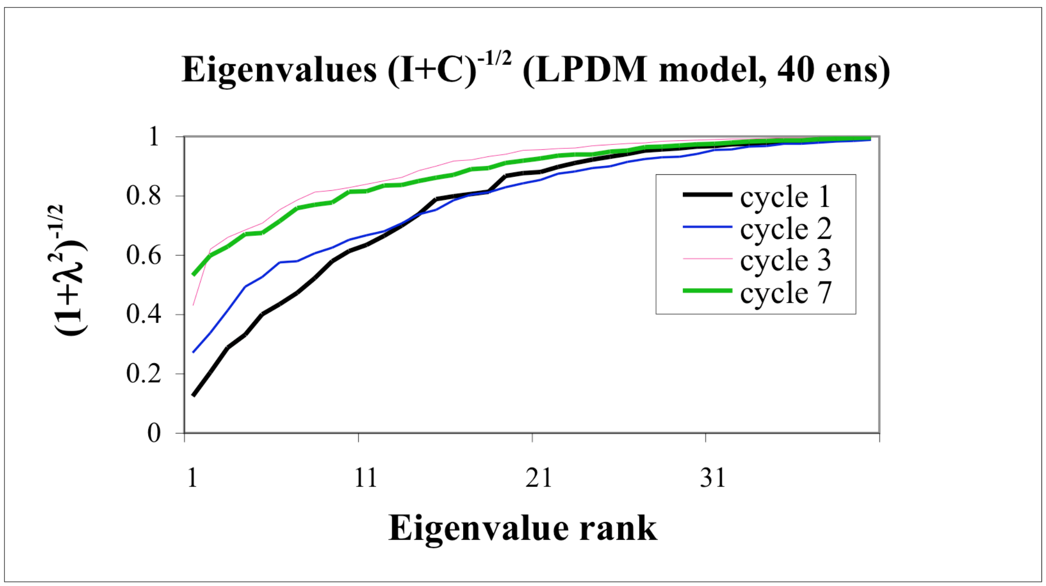
The the number of effective DOF of this system is between 10 and 20. We do not need 1800 ensembles!

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

LPDM Model CO2-flux BIAS estimation: Eigenvalue spectrum of $(I+C)^{-1/2}$ (First 40 eigenvalues, Nens = 1800, 100, and 40)



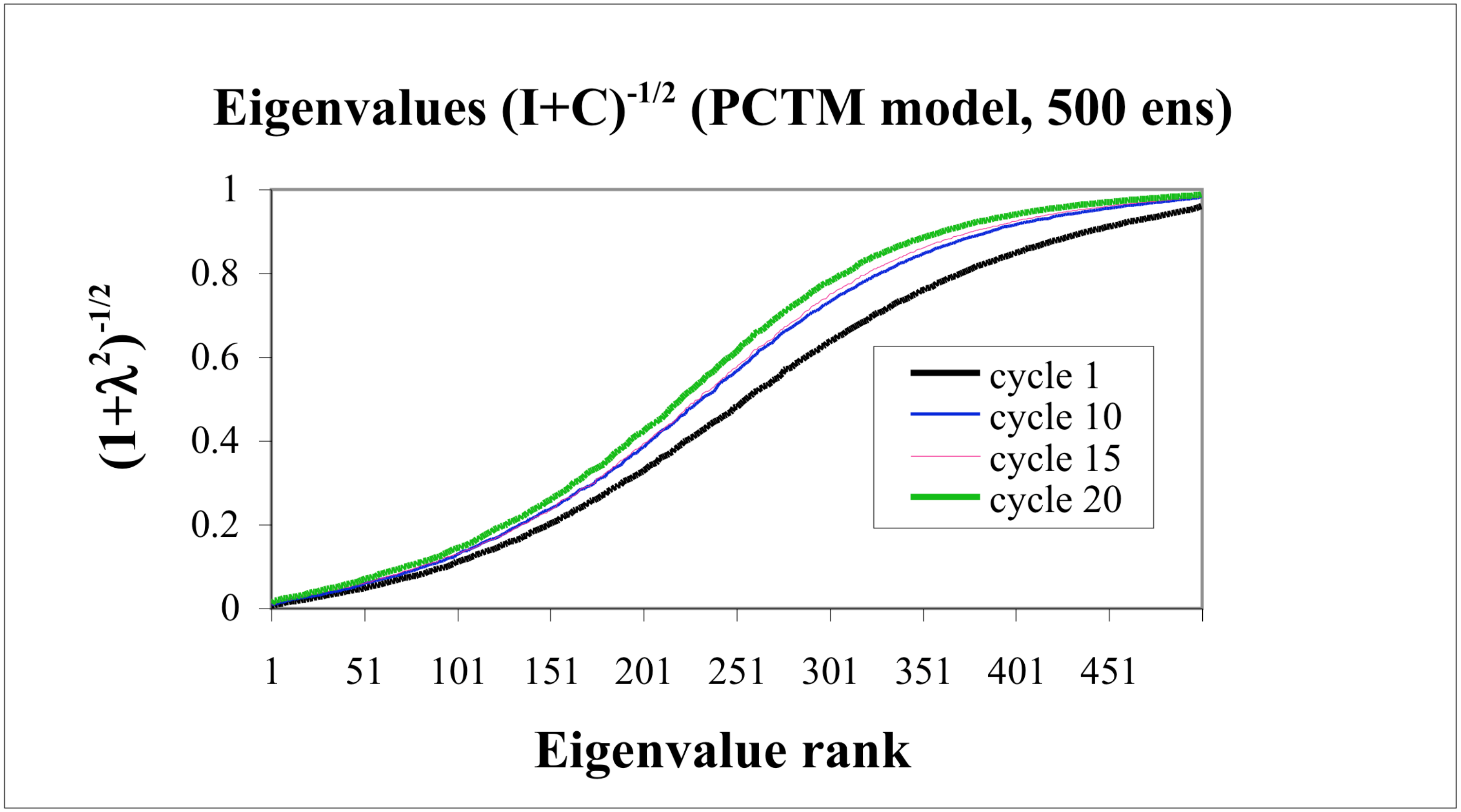




Eigenvalue spectrum is very similar for all 3 ensemble sizes!

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

PCTM Global Model CO2-flux estimation: Eigenvalue spectrum of C (Nstate=13104, Nobs=13104, fully observed system, Nens= 500)

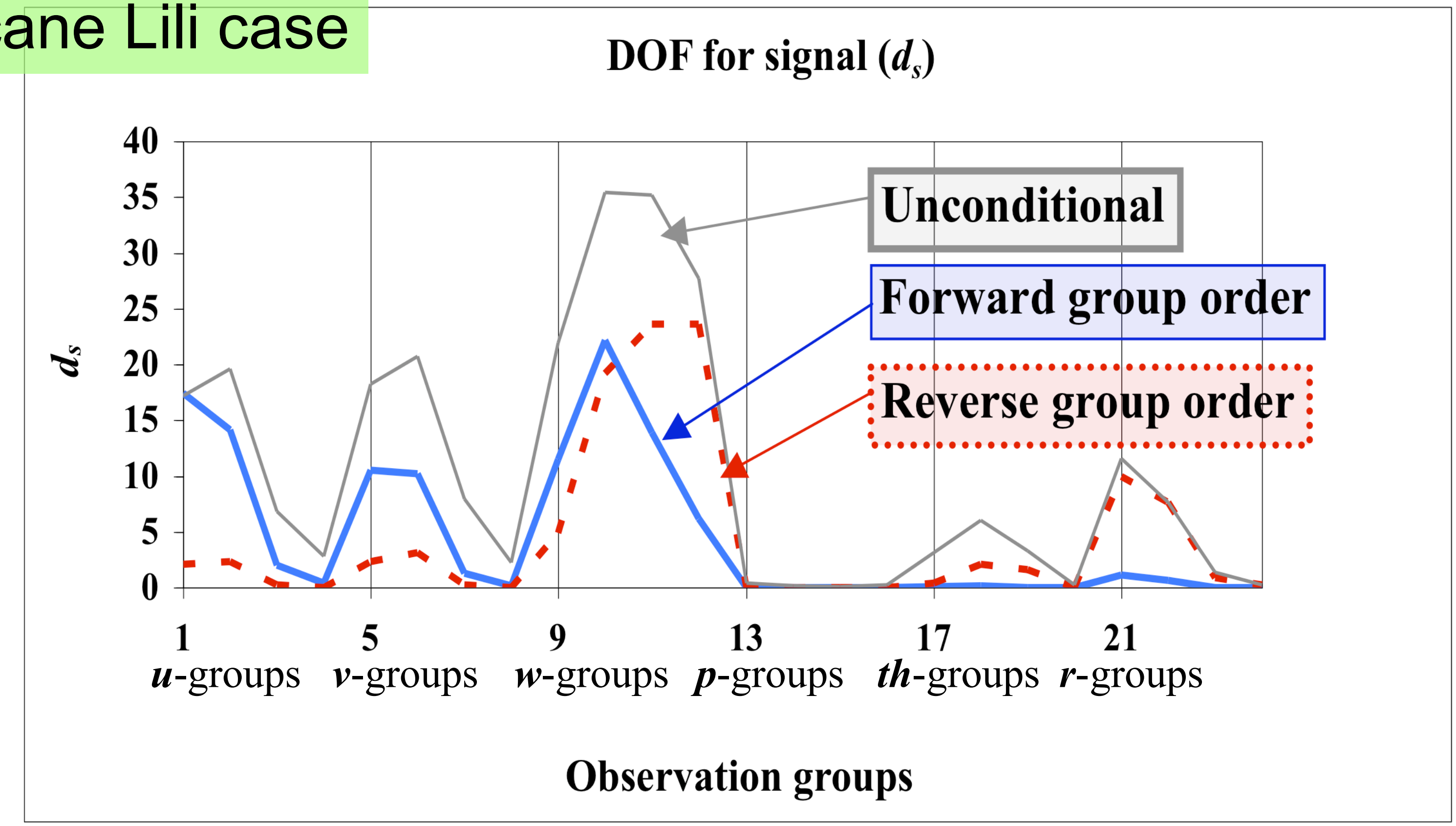


Ensemble size of 500 is adequate for describing all DOFs of this fully observed system.
In later cycles more eigenvalues are approaching value 1 (no information).

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

RAMS Model:
Assimilation of simulated observations: u, v, w, p, th, and r groups of observations assimilated successively (Nstate=54000, Nobs=7200, Nens= 50)

Hurricane Lili case



Conditional information content analysis depends on the group order. Unconditional information content analysis produces largest information content, and does not depend on the group order.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Conclusions and Future Plans

❑ Experience from different dynamical models (e.g., atmospheric and carbon transport models) indicates that information measures, defined in ensemble subspace, are reliable measures of effective DOF.

❑ These measures can be used for many different applications: estimation of information content of data, defining adequate ensemble size, defining adequate control variables for data assimilation, optimally combining different observations, quality control, and data thinning.

❑ Main advantages of using ensemble-based approaches for information content analysis are: flow-dependent error covariance, and small dimensions of information matrix C (*Nens* x *Nens*).

❑ There are indications that a relatively small ensemble size might be sufficient for meaningful information content analysis.

❑ Future Plans: Collaboration with NCEP/EMC on estimating information content of NCEP operational data.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu