Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Stochastic description of poorly represented SGS physical processes and truncation errors

Ekaterina Machulskaya

DWD, FE14

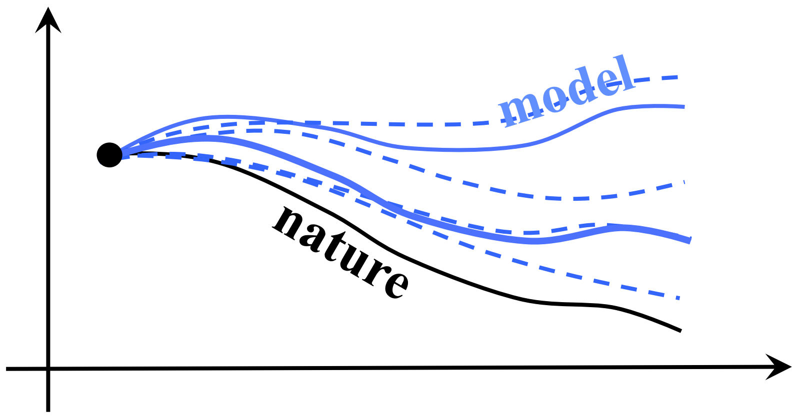
Stochastic Physics Workshop
German Weather Service, Offenbach am Main
26-27 November 2013

Outline

- Motivation: what do we want to achieve?
- Formulation of the problem
- How random variable may be propagated in time
- Methods to construct random model error field
- Outlook

Motivations

Motivation 1: improvement of the deterministic forecast



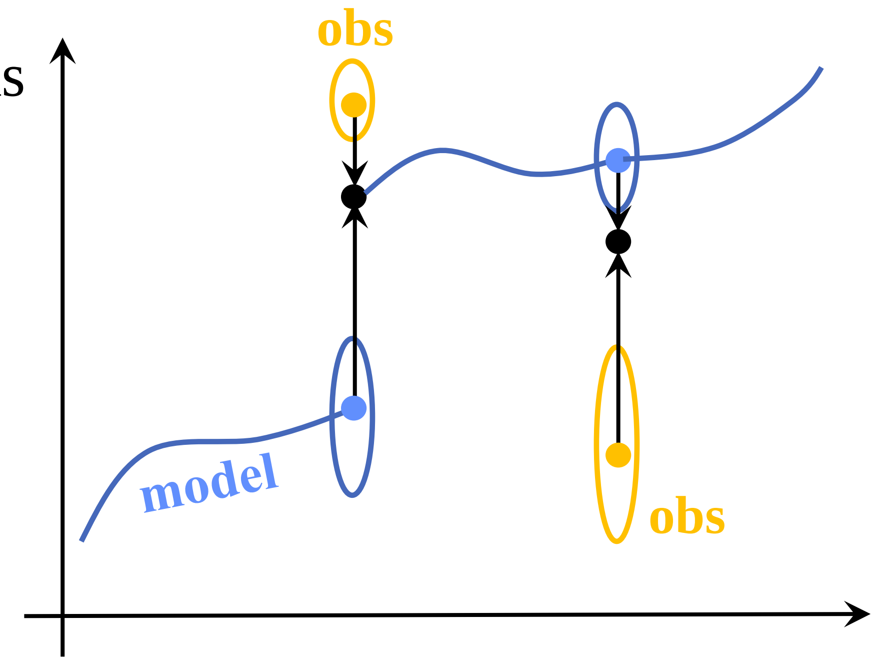
- The model results are imperfect.
- A non-negligible part of errors appears due to the imperfection of a model itself.
- Some errors can be handled deterministically, but some cannot (i.e. from those parameterizations that would require almost infinite resolution, e.g. microphysics, soil, etc.)
- They can be accounted for in a statistical way.

Motivation 2:

- The end-users should be provided with the information how reliable/uncertain the forecast is.

Motivation 3: estimation of the background error

- Most of the data assimilation systems represent an interpolation between the observations and the first guess to provide a new initial condition.



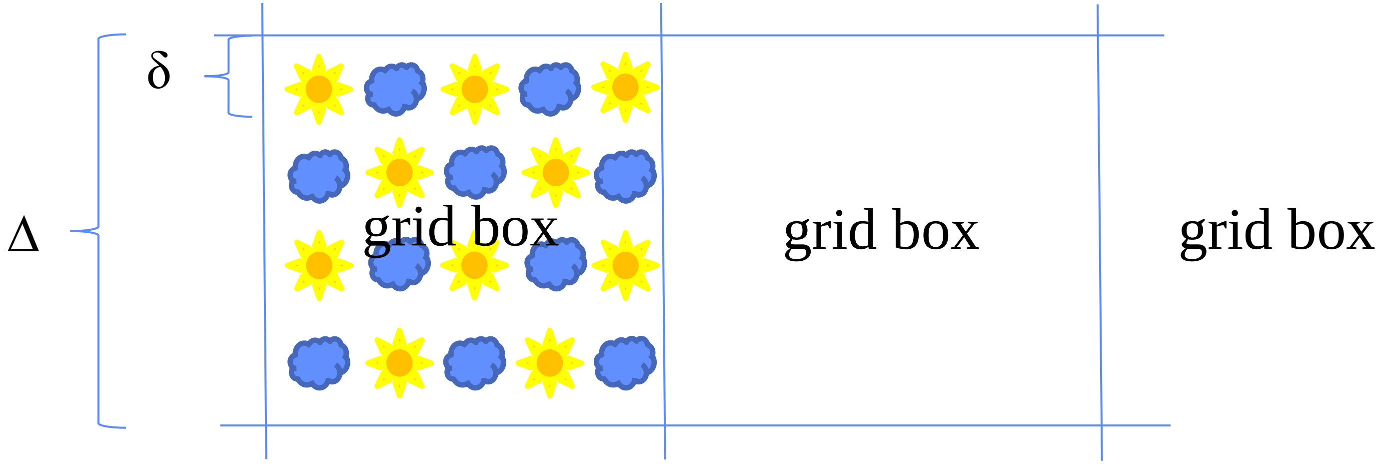
- In KF, the weights for the interpolation between the observations and the model are inversely proportional to the corresponding uncertainties, or possible errors

- An estimate of the model error is needed in order to give an appropriate weight to the first guess. If the model error is underestimated, this weight will be too large and less regard will be paid to the observations than should be.

Formulation of the problem

Example of the solution (see Hasselmann, 1988)

if there exist a clear time and space scale separation between resolved and unresolved processes (Δ >> δ) (= spectral gap!)

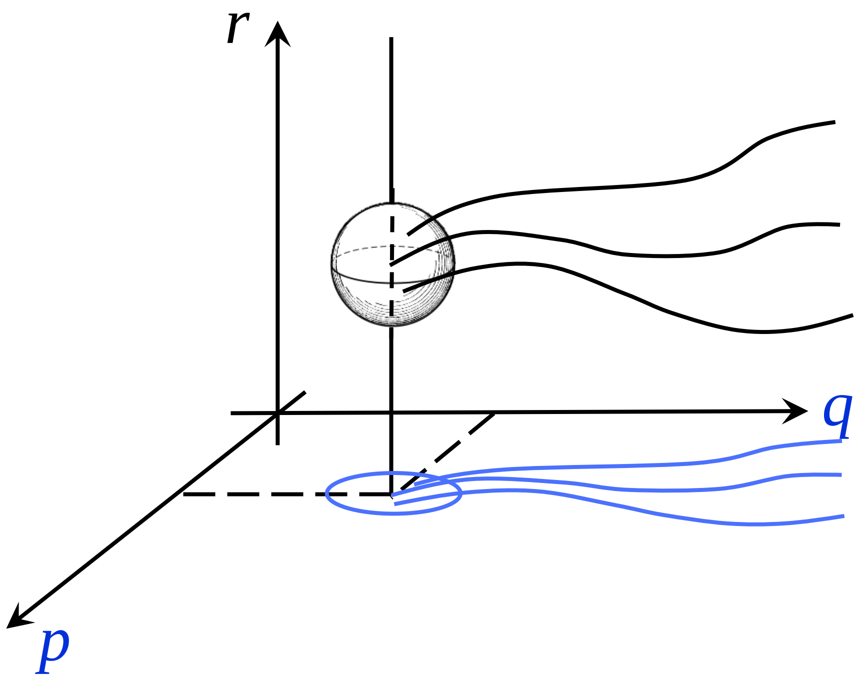


the cumulative effect of the random errors within each grid box may be represented by means of the Central Limit Theorem:

sum of many independent identically distributed random variables is Gaussian
→ the perturbations are the samples of the white noise process

Good for climate studies,
but no spectral gap and independence for the variables in NWP

Formulation of the problem



$\{p, q, r\}$ –

full set of modes (= nature)

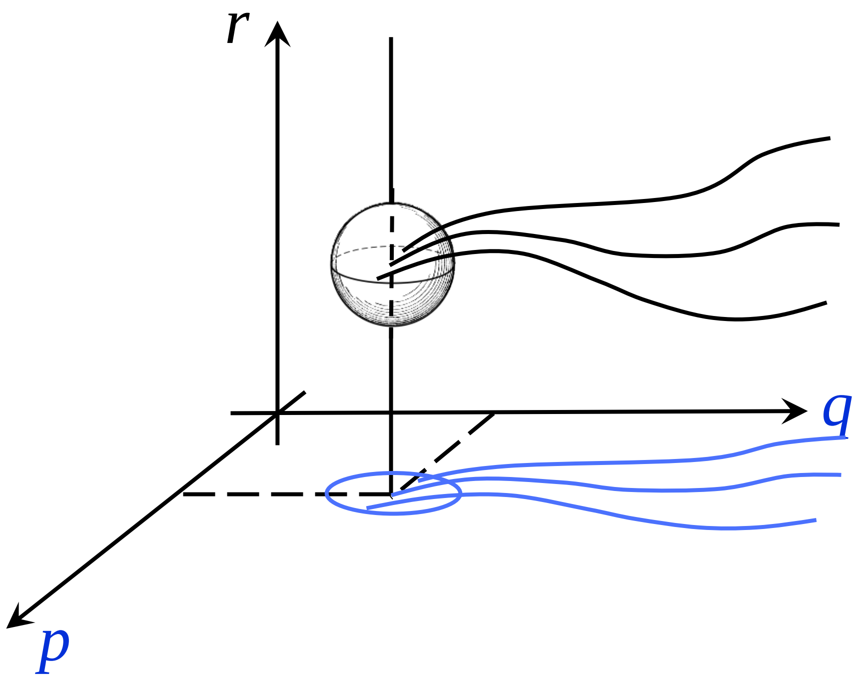
$\{p, q\}$ –

model variables

r – unaccounted degree of freedom

Usually, the exact initial condition is not known.

Formulation of the problem



$\{p, q, r\}$ –

full set of modes (= nature)

$\{p, q\}$ –

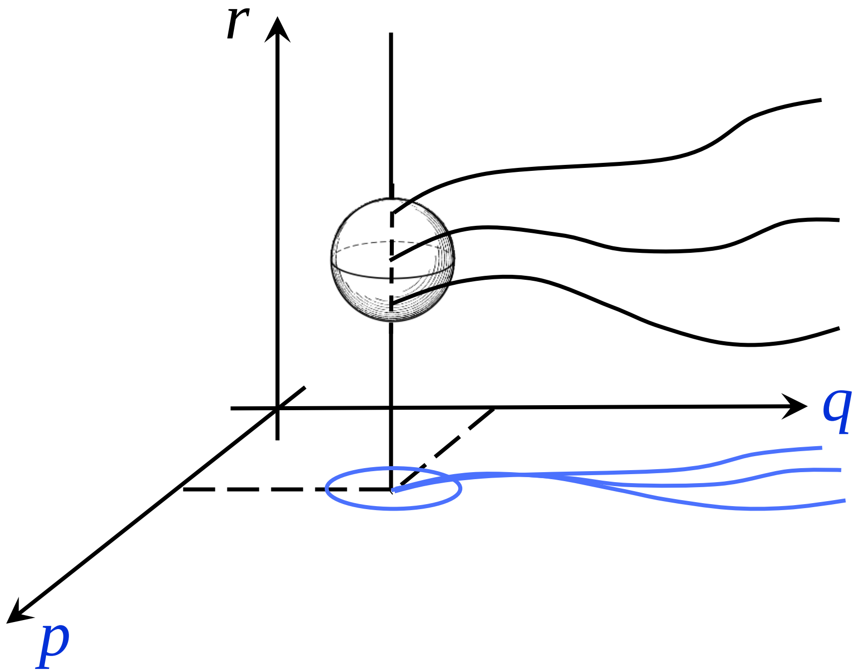
model variables

r – unaccounted degree of freedom

Usually, the exact initial condition is not known.

The lack of knowledge in the model variable's plane (p,q) = the uncertainty in the model's initial conditions.

Formulation of the problem



$\{p, q, r\}$ –

full set of modes (= nature)

$\{p, q\}$ –

model variables

r – unaccounted degree of freedom

Usually, the exact initial condition is not known.

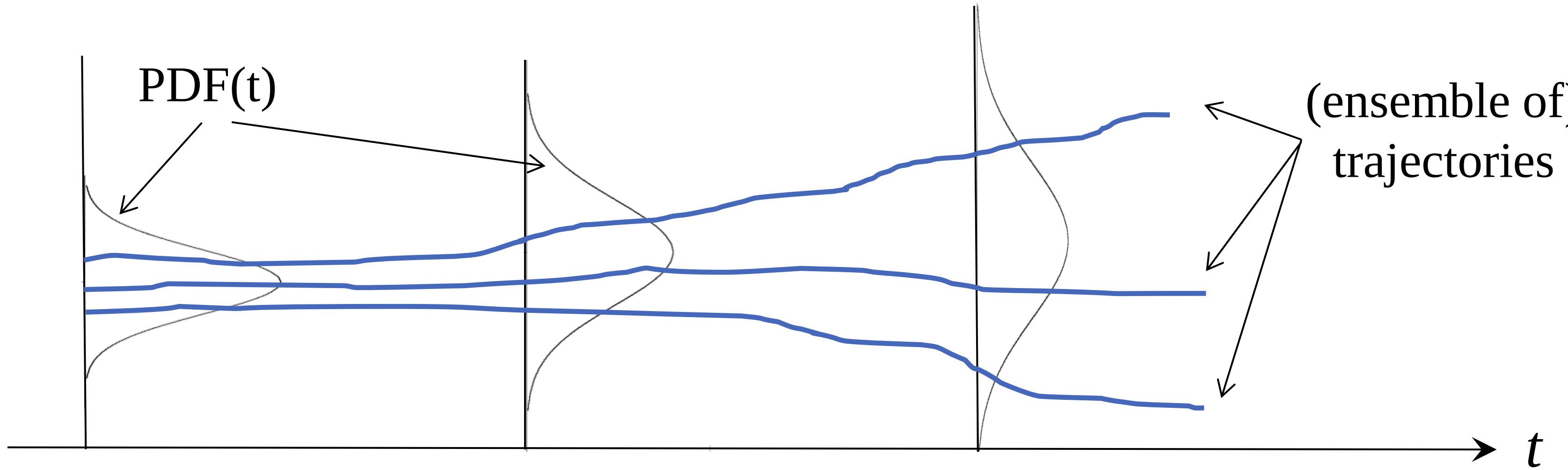
The lack of knowledge in the model variable's plane (p,q) = the uncertainty in the model's initial conditions.

The lack of knowledge in the unresolved mode r = the uncertainty in the model's physics.

Propagation in time

If a (small) part of the model is random, then the model state is a random variable evolving in time (= random process). This evolution may be represented as

- evolution of the probability density function (PDF);
- evolution of all statistical moments of the PDF;
- evolution of all particular realizations of the random process.



Theoretically these ways are equivalent, but practically not necessarily. Which one to choose?

Propagation in time

Example: the error due to the discretization of the advection process

Consider a transport equation of a quantity *f*:

$$\frac{df}{dt} \equiv \frac{\partial f}{\partial t} + u_k \frac{\partial f}{\partial x_k} = S_f$$

Representing a quantity *f* as a sum of the ensemble average (≈ resolved flow) and fluctuations therefrom (≈ unresolved) $f = \overline{f} + f',$ one arrives at an ensemble (≈ spatially) averaged equation

$$\frac{d\overline{f}}{dt} \equiv \frac{\partial \overline{f}}{\partial t} + \overline{u}_i \frac{\partial \overline{f}}{\partial x_i} = -\frac{\partial}{\partial x_i} \overline{u_i' f'} + \overline{S}_f$$

with the second-order subgrid-scale contribution $\overline{u_i' f'}$ subject to a parameterization scheme (= statistical model bias correction due to the interaction between resolved and unresolved flow).

Propagation in time

From the governing equations for *f* and *u* the prognostic equations for all statistical moments can be derived, for example:

$$\frac{d}{dt}\overline{u_i' f'} = -\overline{u_i' u_k'}\frac{\partial \overline{f}}{\partial x_k} - \overline{u_k' f'}\frac{\partial \overline{u_i}}{\partial x_k} - \frac{\partial}{\partial x_k}\overline{u_k' u_i' f'} + \overline{u_i' S_f'} + \overline{f' S_u'}$$

known | requires assumptions | neglected for the most part

$\overline{u_i' S_f'}$ - correlations between subgrid-scale fluctuations of wind velocity and latent heat releases (from microphysics), radiation fluxes, etc.

→ all NWP and climate models account already for a part of the model error stochastically:

the error is the discretization error of the advection equations,
it is accounted for through the estimation of the statistical moments.

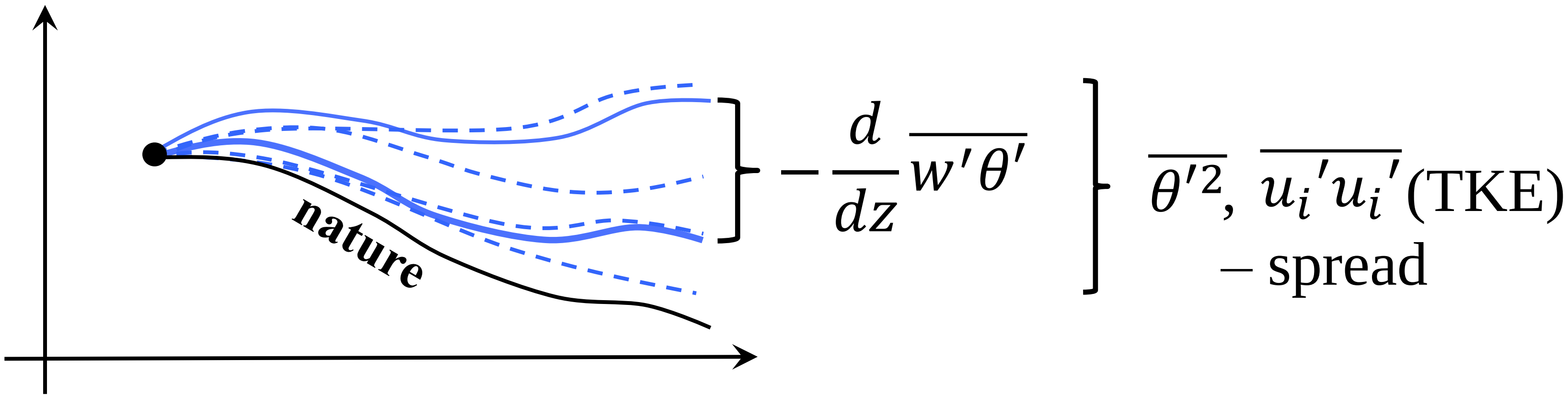
The parts of the model that do this work are called turbulence and convection parameterization schemes.

Propagation in time

If an equation is more complex (the right-hand side includes terms that are highly non-linear, have thresholds, etc.), then the prognostic equations for the moments cannot be easily obtained, if at all.

In this case it might be preferable to use other approaches, e.g. running an ensemble of realizations.

→ The methods can be combined!



How to construct the model error

Two ways are already tried in various studies:

- to derive the statistical properties of the model error **from the available model data** (≈ "top-down")

  Statistical bias correction (e.g. Faller, 1975; Johansson & Saha, 1989; Danforth & Kalnay, 2007; DelSole et al., 2008)

  Linear stochastic models (Nicolis et al., 1997; Achatz & Opsteegh, 2003; Berner, 2005; Sardeshmukh & Sura, 2009)

- to think about "**what can be uncertain** and be the main source of errors and thus **perturbed**" (≈ "bottom-up")

  (Randall & Huffman, 1980; Majda & Khouider et al., 2002; Lin & Neelin et al., 2002; Craig & Cohen, 2006)

How to construct the model error

From the model data:

✓ The entire error is accounted for by construction

– Restricted applicability: parameters of the random processes are fitted to certain regimes, model version, region etc.
Artificial dependencies can appear. Insufficient physical background.

"What should be perturbed":

✓ The choice of the perturbed variables is physically based

– There is no answer to the questions "How it should be perturbed?", "How large is an uncertainty of what is known to be uncertain?"
Final results may not well represent the sought model error field. Danger of double-counting of the errors.

A golden mean is needed

How to construct the model error

<u>Short perspective</u>

use the approach 1 (construct the properties of the model error from the computed data)

Assume the form of how the model error enters the governing equations:

$$\frac{\partial Y}{\partial t} = \left[\frac{\partial Y}{\partial t}\right]_{model} + (a + bY)X(t)$$

Assume the form of the model error equation:

$$\frac{\partial X}{\partial t} = -\gamma X + (\sigma_a + \sigma_b X)\xi(t)$$

damping (time corr.)

noise ($\xi \sim N(0,1)$)

→

Approximate the parameters a, b, γ, σ_a and σ_b by means of the data

Take "forecast–analysis" as a proxy to the model error

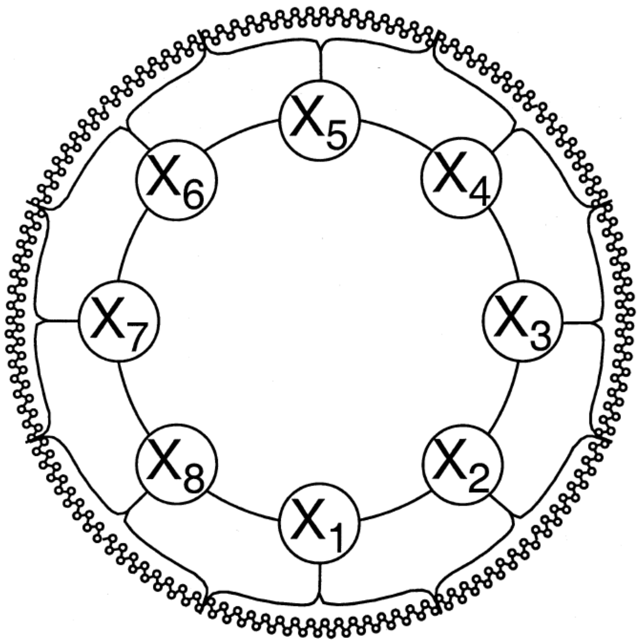
Decrease the dimensionality of the phase space – determine the leading patterns of the model error (spatial correlations are accounted for)

How to construct the model error

Lorenz'96 system (Lorenz 1996)

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} Y_j$$

$$\frac{dY_j}{dt} = -cbY_{j+1}(Y_{j+2} - Y_{j-1}) - cY_j + \frac{hc}{b} X_{int[(j-1)/J]+1}$$



X_k – slow variables
Y_j – fast variables
h, c, b, F – parameters

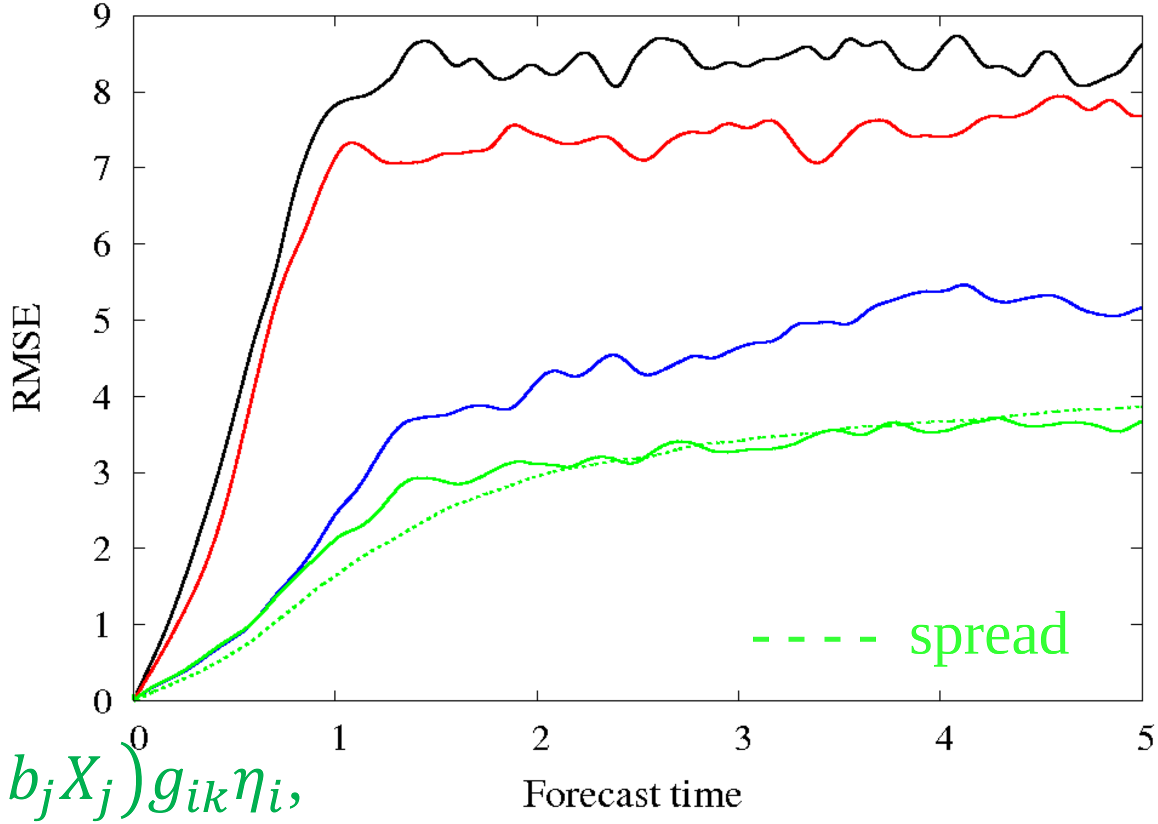
How to construct the model error

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F + ME_k$$

$ME_k = 0$
(no SGS contribution)

$ME_k = A$ (constant)

$ME_k = A + B_{jk}X_j$
(flow dependent, deterministic)



$$ME_k = A + B_{jk}X_j + \left(a + b_j X_j\right) g_{ik}\eta_i,$$

$$\frac{d\eta_i}{dt} = -\gamma_i\eta_i + (\sigma_a + \sigma_b X)\xi_i\,(t), \quad \xi_i \sim N(0,1)$$

g_{ik} – coefficients of POPS decomposition

How to construct the model error

$$\frac{d\boldsymbol{z}}{dt} = F(\boldsymbol{z}, t)$$

$\boldsymbol{z}$ – full set of modes (= **nature**) $\quad \boldsymbol{z} = \{\boldsymbol{x}, \boldsymbol{y}\}$

(e.g. $\boldsymbol{z}$ can be the set of Fourier components of a solution of the equation prior to the discretization procedure)

Let us regard $\boldsymbol{x}$ as a set of resolved components (= **model variables**) and $\boldsymbol{y}$ as a set of unresolved components.

$$\frac{d\boldsymbol{x}}{dt} = f^{rr}(\boldsymbol{x}, t) + f^{ru}(\boldsymbol{x}, \boldsymbol{y}, t)$$

$$\frac{d\boldsymbol{y}}{dt} = f^{ur}(\boldsymbol{x}, \boldsymbol{y}, t) + f^{uu}(\boldsymbol{y}, t)$$

For each value of $\boldsymbol{x}$ there is an ensemble of values of the unresolved degrees of freedom $\boldsymbol{y}$

→ in the equation for $\boldsymbol{x}$ modes the term $f^{ru}(\boldsymbol{x},\boldsymbol{y},t)$ may be represented by an appropriate random process ξ:

$$\frac{d\boldsymbol{x}}{dt} = f^{rr}(\boldsymbol{x}, t) + \xi(\boldsymbol{x}, t)$$

How to construct the model error

"Appropriate" means:

the properties of ξ should not be arbitrary, but consistent with the properties of f^{ru}, f^{ur} and f^{uu}!

Systematic stochastic mode reduction (Majda et al., 2001):

- the interaction f^{uu} is the fastest one among f^{rr}, f^{ru}, f^{ur} and f^{uu}
- replace the fastest interaction terms f^{uu} with a random process
- the equations $\frac{dy}{dt} = \ldots$ for the unresolved components become stochastic
- technics are being developed how to eliminate the unresolved variables in the equations for the resolved variables

See also Kraichnan, 1988; Lindenberg & West, 1984

How to construct the model error

Lorenz'96 system $$\frac{dY_j}{dt} = -cbY_{j+1}\left(Y_{j+2} - Y_{j-1}\right) - cY_j + \frac{hc}{b}X_{near}$$

the fastest terms – interactions between unresolved

$$\underline{-cbY_{j+1}\left(Y_{j+2} - Y_{j-1}\right)} = \boxed{-\gamma_j Y_j + \sigma_j \xi_j\,(t)}$$

random process

Solution

$$Y_j = Y_{j0}e^{-c\beta_j t} + \frac{h}{b\beta_j}\int_0^t X_{near}(\tau)e^{-c\beta_j(t-\tau)}d\tau + \frac{b\sigma_j}{\beta_j}\int_0^t e^{-c\beta_j(t-\tau)}dW$$

where $\beta_j = b\gamma_j + 1$, W – Wiener process

Memoryless approximation $$\int_0^t X_{near}(\tau)e^{-c\beta_j(t-\tau)}d\tau \approx X_{near}(t)$$

$$\int_0^t e^{-c\beta_j(t-\tau)}dW \approx \xi_j\,(t)$$

How to construct the model error

$\rightarrow$ Solution for Y

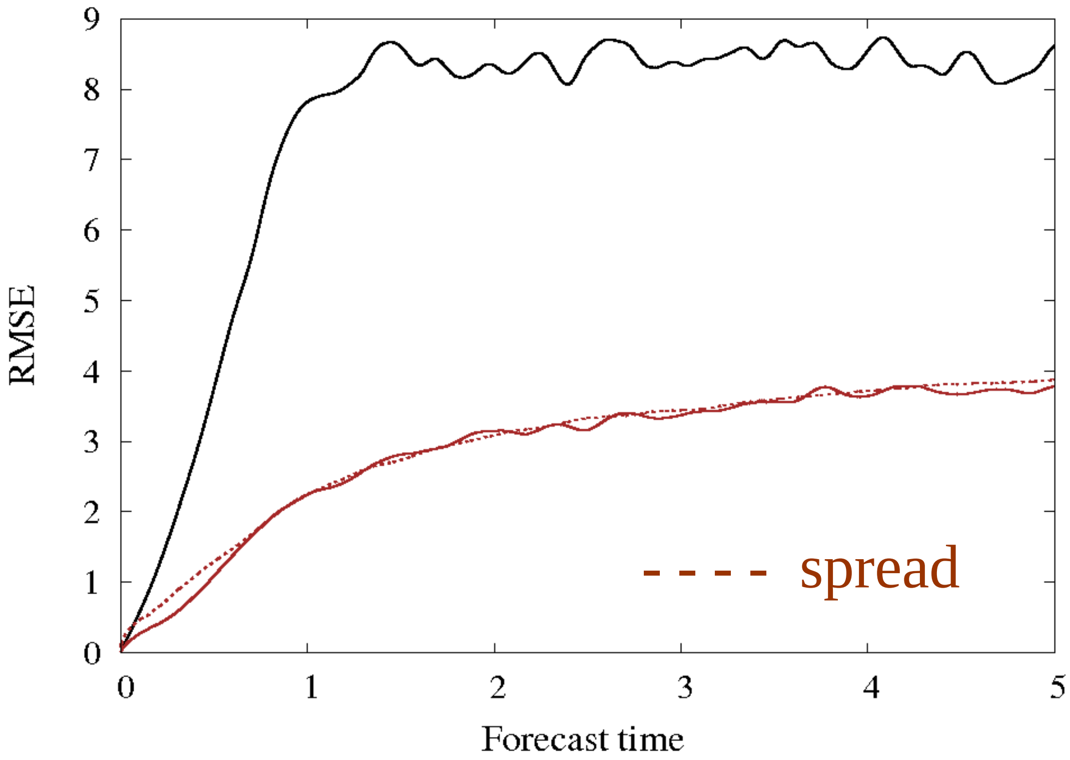
$$Y_j(t) = Y_{j0} e^{-c\beta_j t} + \frac{h}{b\beta_j} X_{near}(t) + \frac{b\sigma_j}{\beta_j} \xi_j(t)$$

Inserting into the equations for X

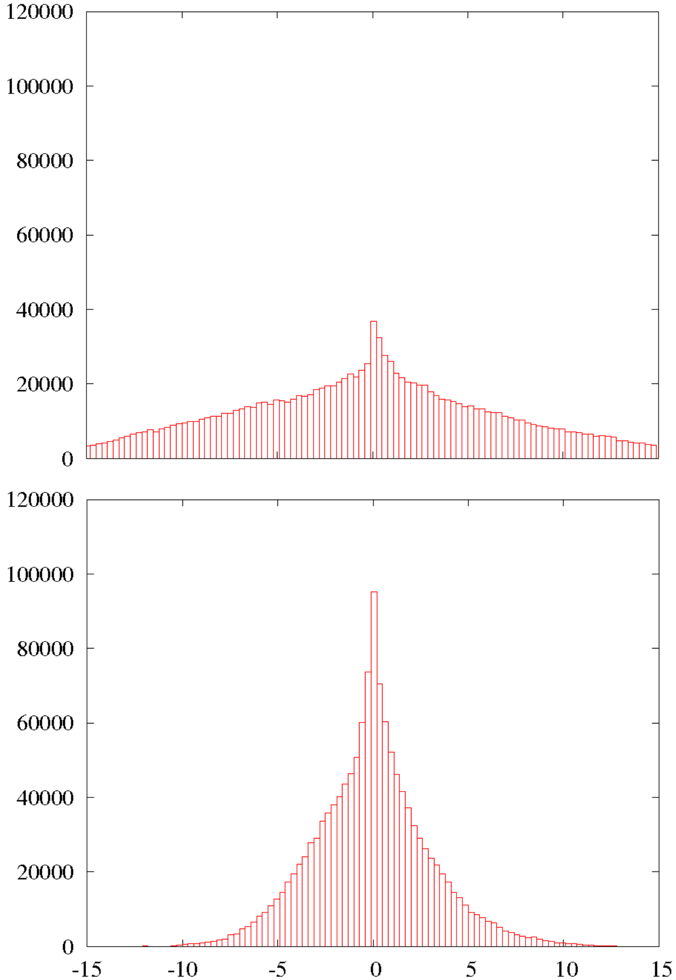
$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k \left(1 + \frac{h^2 c}{b^2} \sum_{j=J(k-1)+1}^{kJ} \frac{1}{\beta_j} \right) + F$$
$$- \frac{hc}{b} \sum_{j=J(k-1)+1}^{kJ} Y_{j0} e^{-c\beta_j t} - hc \sum_{j=J(k-1)+1}^{kJ} \frac{\sigma_j}{\beta_j} \xi_j(t)$$

How to construct the model error

Stochastic mode elimination
for the Lorenz'96 system: results



Error histograms:
upper panel – without SGS
lower panel – stoch. SGS



Outlook

- Implementation and testing of the approach 1 in the COSMO-DE/ICON
- Is it possible to determine the errors from different physical parameterizations separately? (Discretization errors at least.)
- If yes, implementation and testing of the approach 1 for each parameterization separately
- Development of a more consistent approach

Thank you for your attention!

Thanks to Dmitrii Mironov and Bodo Ritter for fruitful discussions!