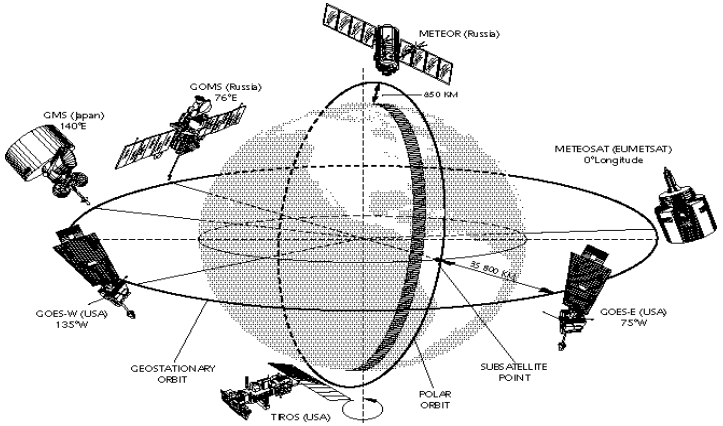
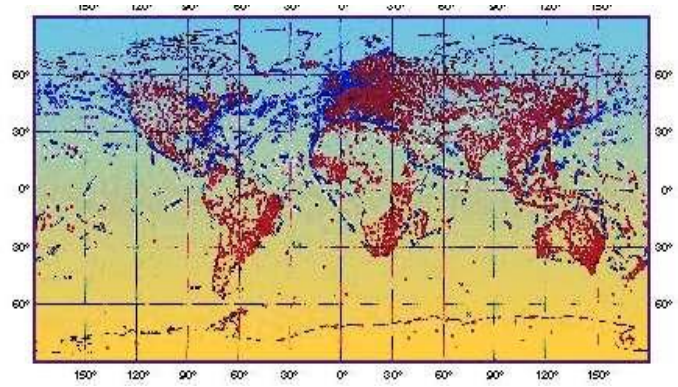


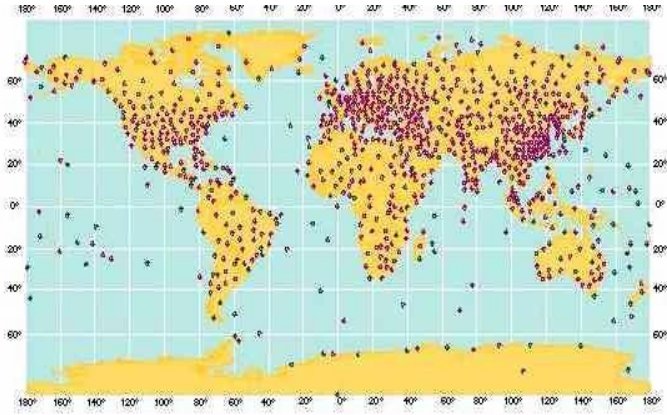
Klimova Ekaterina

Institute of Computational Technologies Russian Academy of Sciences, Siberian Branch

Computing aspects of an environment estimation on the basis of the observational data

klimova@ict.nsc.ru





Introduction

Components of the new forecast system (New Forecast Paradigm):

a. ***Ultimate goal***. The extension of the traditional forecast process: reduction of forecast uncertainty, and also providing forecast uncertainty.
b. ***Forecast process***. In the new paradigm, not only the best estimate of predicted system, but also uncertainty is propagated.
c. ***Observing system***. Estimation of random instrument and representativeness error variance, as well as estimation of systematic errors.
d. ***Data assimilation***. Reduction of analyses error and assessment of uncertainty in the analyses. This information is critical input for the generation of initial ensemble perturbations.
e. ***Numerical modeling***. Reducing systematic and random error related to model formulation. A quantitative assessment and simulation of model related random and systematic errors.
f. ***Ensemble forecasting***. In the new forecast process, ensemble forecasting occupies a central place in the entire process following the observing, data assimilation and numerical modeling components.
g. ***Statistical pre-processing***.

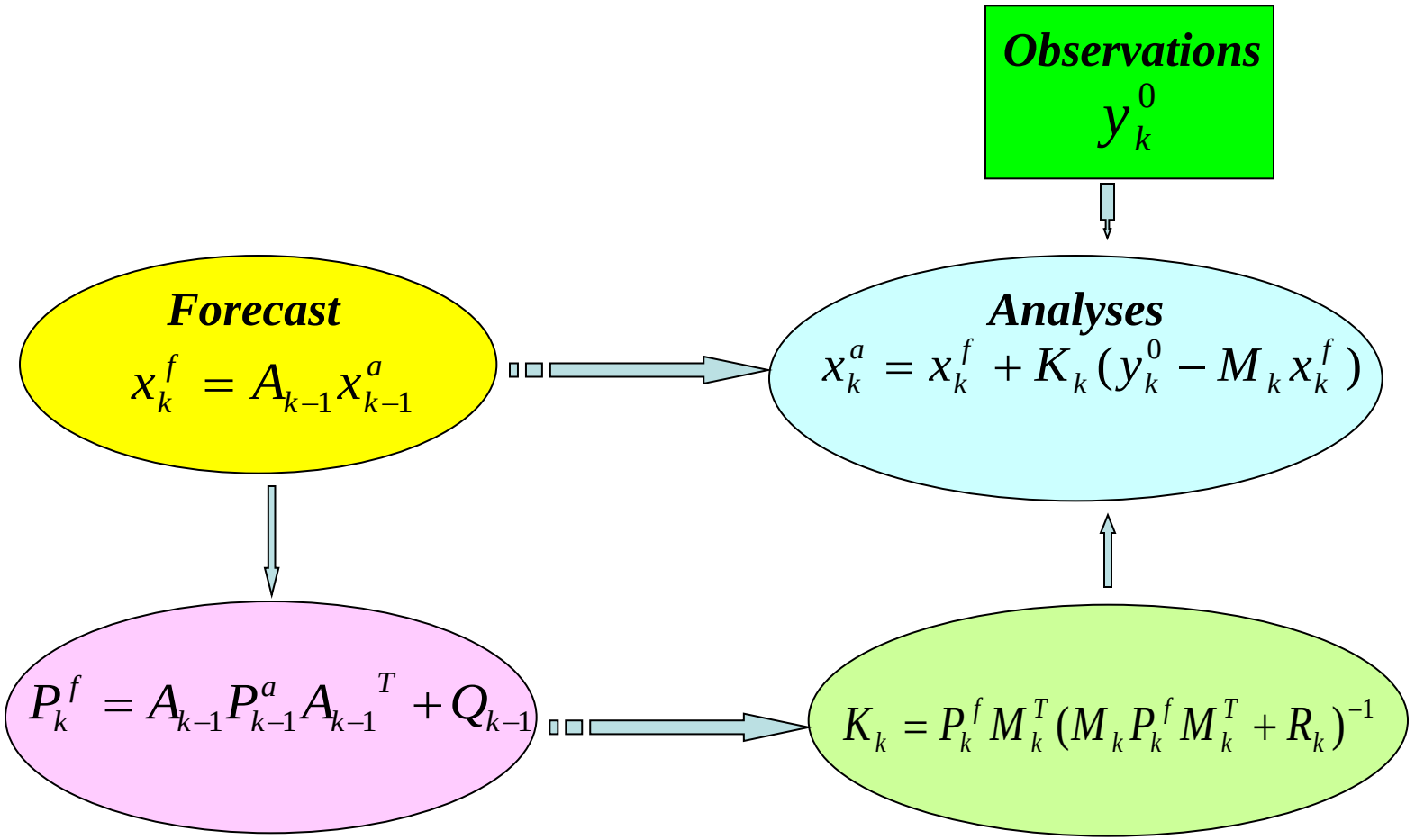
Z.Toth et al. Completing the forecast: assessing and communicating forecast uncertainty. – ECMWF Workshop on Ensemble Prediction 7-9 November 2007.

Introduction

- An Ensemble of Data assimilations (EDA) system was introduced at ECMWF.
- The EDA consists of an ensemble of ten 4D-Var assimilations that differ by perturbing observations, sea surface temperature fields and model physics.
- The main justification for implementing the EDA is that it quantifies analysis uncertainty.
- It can be used to estimate flow-dependent background errors in the deterministic 4D-Var assimilation system.

L.Isaksen et al. Ensemble of Data assimilation at ECMWF. – Technical memorandum N636, December 2010:

Kalman Filter



Ensemble Kalman filter

Ensembles of initial fields and forecasts:

$$x_f^{0(i)} = \hat{x} + \Delta x_0^{(i)}, \quad i = 1, \cdots, N$$

$$x_{k+1}^{f(i)} = A_k(x_k^{a(i)}) + \xi_k^{(i)},$$

Estimation of covariance matrix:

$$P_k^f M_k^T = \frac{1}{N-1}\sum_{i=1}^{N}(x_k^{f(i)} - \overline{x_k^{f(i)}})(M_k x_k^{f(i)} - \overline{M_k x_k^{f(i)}})^T,$$

$$M_k P_k^f M_k^T = \frac{1}{N-1}\sum_{i=1}^{N}(M_k x_k^{f(i)} - \overline{M_k x_k^{f(i)}})(M_k x_k^{f(i)} - \overline{M_k x_k^{f(i)}})^T,$$

Ensemble of «analyses»:

$$K_k = P_k^f M_k^T (M_k P_k^f M_k^T + R_k)^{-1},$$

$$x_k^{a(i)} = x_k^{f(i)} + K_k(y_k^{0(i)} - M_k x_k^{f(i)}),$$

$$y_k^{0(i)} = y_k^0 + r_k^{(i)}$$

$$\overline{\mathrm{x}_{\mathrm{k+1}}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{x}_{\mathrm{k+1}}^{(\mathrm{i})}$$

Ensemble pi-algorithm

The forecast step can be written the following way:

$$\mathbf{x}^{\mathrm{f}}(t_{k+1}) = M(\mathbf{x}^{a}(t_k)) + \boldsymbol{\eta}(t_k),$$

where $\mathbf{x}^{f}(t_{k+1})$ is a vector of forecasted values at moment of time t_{k+1}

$\mathbf{x}^{a}(t_k)$ is the vector of values, obtained after a step of analysis at moment of time t_k, M is a model operator, $\boldsymbol{\eta}(t_k)$ is Gaussian white noise with covariance matrix $\mathbf{Q}_k$.

Ensemble pi-algorithm

The step of analysis is expressed as:

$$\mathbf{x}^a(t_k) = \mathbf{x}^f(t_k) + \mathbf{P}_k^a \mathbf{H}^T \mathbf{R}_k^{-1}(\mathbf{y}_{t_k} - H(\mathbf{x}^f(t_k))),$$

where $\mathbf{P}_k^a$ is an analysis error covariance matrix,

$\mathbf{R}_k$ is an observation error covariance matrix,

H is an operator (generally speaking, nonlinear), transferring values in the grid points to the observations point,

$\mathbf{H}$ - is an linearized operator,

$\mathbf{y}_{t_k}$ is an observations vector at moment of time t_k

Ensemble pi-algorithm

So let present the algorithm in the equivalent way:

$$\mathbf{x}(t_{k+1}) = M(\mathbf{x}(t_k)) + \boldsymbol{\eta}(t_k) + \mathbf{P}^a_{k+1}\mathbf{H}^T\mathbf{R}^{-1}_{k+1}(\mathbf{y}_{t_{k+1}} - H(M(\mathbf{x}(t_k)) + \boldsymbol{\eta}(t_k))),$$

where $\mathbf{P}^a_{k+1} = (\mathbf{I} - \mathbf{KH})\mathbf{P}^f_{k+1}, \mathbf{K} = \mathbf{P}^f_{k+1}\mathbf{H}^T(\mathbf{H}\mathbf{P}^f_{k+1}\mathbf{H}^T + \mathbf{R}_{k+1})^{-1}$,

$\mathbf{P}^f_{k+1}$ is a forecast error covariance matrix.

Written this way the formula unites the steps of analysis and forecast, that allow one to neglect indexes "*a*" and "*f*" further on.

Let true value $\mathbf{x}_t$ suffices the following equation:

$$\mathbf{x}_t(t_{k+1}) = M(\mathbf{x}_t(t_k)),$$
$$\mathbf{x}_t(t_0) = \mathbf{x}_0.$$

The observation data can be expressed as:

$$\mathbf{y}_{t_k} = H(\mathbf{x}_t(t_k)) + \boldsymbol{\varepsilon}^k,$$

where $\boldsymbol{\varepsilon}^k$ is random observation error

with zero-order expectation value and covariance matrix $\boldsymbol{R}_k$.

Ensemble pi-algorithm

Let the estimation error be determined as

$$\mathbf{dx}^{k+1} = \mathbf{x}_t(t_{k+1}) - \mathbf{x}(t_{k+1}).$$

The error suffices the following equation:

$$\mathbf{dx}^{k+1} = M(\mathbf{x}_t(t_k)) - M(\mathbf{x}(t_k)) - \boldsymbol{\eta}(t_k) - \mathbf{P}_{k+1}\mathbf{H}^T\mathbf{R}_{k+1}^{-1}(H(M(\mathbf{x}_t(t_k)) + \boldsymbol{\varepsilon}^{k+1} - H(M(\mathbf{x}(t_k) + \boldsymbol{\eta}(t_k))).$$

If to estimate $\mathbf{P}_{k+1}$ using the formula (Yaglom, 1987)

$$\mathbf{P}_{k+1} = \overline{\mathbf{dx}_n^{k+1}\left(\mathbf{dx}_n^{k+1}\right)^T} \approx \frac{1}{N-1}\sum_{n=1}^{N}\mathbf{dx}_n^{k+1}\left(\mathbf{dx}_n^{k+1}\right)^T,$$

one obtains a version of the ensemble Kalman filter.

Taking this formula for $\mathbf{P}_{k+1}$ into account one obtains a system of equations relative to $\mathbf{dx}_n^{k+1}$

Ensemble pi-algorithm

Now let one consider a modification of the algorithm described above, so it can be applied for forecasting of ensembles. It is known that such a forecast requires setting an ensemble of initial fields $\{\mathbf{x}_n\}$ in such a way that ensemble average $\overline{\mathbf{x}_n}$ is equal to $\mathbf{x}^a$, while covariances $\overline{(\mathbf{x}_n - \overline{\mathbf{x}_n})(\mathbf{x}_n - \overline{\mathbf{x}_n})^T} = \mathbf{P}^a$, where $\mathbf{x}^a$ is the result of the step of analysis of the Kalman filter, $\mathbf{P}^a$ is a analysis error covariance matrix. The following ensemble of initial fields suffices the first condition:

$$\mathbf{x}_n(t_{k+1}) = M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k) + \mathbf{P}_{k+1}\mathbf{H}^T\mathbf{R}_{k+1}^{-1}(\tilde{\mathbf{y}}_{t_{k+1}}^n - H(M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k))),$$

if accepted that:

$$\overline{H(M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k)))} \approx H(M(\overline{\mathbf{x}_n(t_k)})).$$

At this the ensemble average will be the estimation obtained using the Kalman filter, while its deviation from ensemble member is considered as the estimation error. To describe the errors with the formulas of the classic Kalman filter one has to set a perturbed observations ensemble:

$$\tilde{\mathbf{y}}_{t_k}^n = \mathbf{y}_{t_k} + \boldsymbol{\varepsilon}_n^k.$$

Ensemble pi-algorithm

Analyses step:

$$\mathbf{X}_n^{(k+1)T} = \mathbf{F}_2^T + \mathbf{\Pi}_2^T \mathbf{D}^T, \qquad \left(\mathbf{F}_2^T\right)_n = M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k),$$

$$\left(\mathbf{\Pi}_2^T\right)_m^n = \frac{1}{N-1}\left(\mathbf{dx}_m^{k+1}\right)^T \mathbf{H}^T \mathbf{R}_{k+1}^{-1}(\mathbf{y}_{t_{k+1}} + \boldsymbol{\varepsilon}_n^{k+1} - H(M(\mathbf{x}_k^n) + \boldsymbol{\eta}_n(t_k))).$$

Where

$\mathbf{F}$ is a matrix with columns $\{\mathbf{f}_n^k, n = 1,\ldots,N\}$:

$$\mathbf{f}_n^k = M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k) - \overline{M(\mathbf{x}_n(t_k))}$$

$\tilde{\mathbf{F}}$ is a matrix with columns $\{\tilde{\mathbf{f}}_n^k, n = 1,\ldots,N\}$:

$$\tilde{\mathbf{f}}_n^k = H(M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k)) + \boldsymbol{\varepsilon}_n^{k+1} - \overline{H(M(\mathbf{x}_n(t_k)) + \boldsymbol{\eta}_n(t_k))}$$

$$\mathrm{D} = \begin{pmatrix} \mathrm{dx}_1^1 & \cdots & \mathrm{dx}_1^{\mathrm{N}} \\ \vdots & \ddots & \vdots \\ \mathrm{dx}_{\mathrm{K}}^1 & \cdots & \mathrm{dx}_{\mathrm{K}}^{\mathrm{N}} \end{pmatrix}$$

$$\mathbf{D}^T = (\mathbf{I} + \mathbf{\Pi}^T)^{-1}\mathbf{F}^T,$$

$$\mathbf{\Pi}^T = (\mathbf{C} + 0.25\mathbf{I})^{\frac{1}{2}} - 0.5\mathbf{I}.$$

$$\mathbf{C} = \frac{1}{N-1}\mathbf{F}^T\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{HF} + \mathbf{E}) = \mathbf{C}_1 + \mathbf{C}_2.$$

«Local» ensemble pi-algorithm

The formula $P \approx \frac{1}{N-1}\sum_{i=1}^{N} dx^i \left(dx^i\right)^T$ is approximation, so, if dimension of sample N is small covariance function property is not carried out. In a number of works it is offered for repayment "false" covariances on the big distances to use formula $\tilde{P} = P \circ \Phi(\rho)$, where $\Phi(\rho)$ - function from distance between the points, usually looking like $e^{-\alpha\rho^2}$.
It is well known, that $\tilde{P}$ is also covariance matrix.

Let's consider the analyses step:

$$x_a^i = x_f^i + D[(HD)^T R^{-1}(y^i - Hx_f^i)]$$

Rectangular matrix $D(HD)^T$ is a covariance matrix of forecast errors in a grid cells and points of observations. We will multiply the elements of matrix by $\Phi(\rho)$.

«Local» ensemble pi-algorithm

So we have the following formula for m-th grid point:

$$\left(dx^T\right)_m = \left(f^T\right)_m - \frac{1}{N-1}[DD^T H^T \Phi(\rho) R^{-1}(HF+E)]$$

where $\left(dx^T\right)_m = \begin{pmatrix} dx_m^{(1)} \\ \vdots \\ dx_m^{(N)} \end{pmatrix}$ - the ensemble of perturbations.

And, for individual perturbation in m-th grid point

$$\left(dx^T\right)_m = \left(f^T\right)_m - \left(\Pi_m\right)^T \left(dx^T\right)_m$$

$$\left(\Pi_m\right)_i^j = \sum_{i_0} dx_{i_0}^i \left(f_{i_0}^i + \varepsilon_{i_0}^i\right) e^{-\alpha\rho(i_0,m)}$$

$$D^T \equiv [dx_1^T, \cdots, dx_K^T] = F^T - [\Pi_1^T dx_1^T, \cdots, \Pi_K^T dx_K^T],$$

$$F^T \equiv [f_1^T, \cdots, f_K^T]$$

In the case $\Pi_m \equiv \Pi$ we will have the common variant of ensemble π - algorithm.

Numerical realization of «local» algorithm

We will consider iterative variant of analyses step, in one iteration only one observation is used (in the case, when observation errors don't correlate with each other):

$$x_a^{(l)} = x_a^{(l-1)} + Dx_a^{(l)}[(HDx^{(l)})^T R^{-1}(y^i - Hx_a^{(l-1)})],$$

where l- the iteration number.

1) Let's consider for simplicity, that observations are located in grid points. Then, for grid point $k = i_0$:

$$\left(dx^T\right)_k = \left(f^T\right)_k - \left(\Pi_k\right)^T \left(dx^T\right)_k$$

$$\left(\Pi_k\right)_i^j = dx_{i_0}^i r_0^{-2} dx_{i_0}^j$$

$$(I + \Pi_k^T) dx_k^T = f_k^T,$$
$$(I + \Pi_k^T)\Pi_k^T = f_k^T r_0^{-2} dx_{i_0}^T \equiv C_k$$

Π_k for that grid point may be calculated by formulas which have been received earlier (common formulas).

$$\Pi_k = (C_k + 0.25I)^{1/2} - 0.5I.$$

Also "perturbed" observations may be considered.

Numerical realization of «local» algorithm

2) For grid points, not coinciding with i_0:

$$\left(dx^T\right)_k = \left(f^T\right)_k - \left(\Pi_k\right)^T \left(dx^T\right)_k$$

$$\left(\Pi_k\right)_i^j = dx_{i_0}^i r_0^{-2} dx_{i_0}^j e^{-\alpha\rho(i_0,k)} = C_{i_0}^k \equiv C^k e^{-\alpha\rho(i_0,k)} .$$

dx_{i_0} has been calculated on the previous step, so

$$dx_k^T = (I + C_{i_0}^k)^{-1} f_k^T .$$

For calculation of a square root from a matrix in the previous step the own vectors and own values of matrix C^k have been calculated.

$$\lambda(C_{i_0}^k) = \lambda(C^k) e^{-\alpha\rho(i_0,k)}$$

So it is possible to use the own vectors and own values of matrix C^k calculated on the previous step for a finding of a return matrix.

Lorenz model

Lorenz-98 mode: Lorenz E.N., K.A.Emanuel. Optimal sites for supplementary weather obresvations: simulation with a small model. – MWR, 1998, vol.55, p.399-414.

Let's concider J values $x_1,\cdots,x_J$ (J=40).

The model equations are:

$$\frac{dx_j}{dt} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F,\ j = 1,\cdots,J.$$

$$x_{-1} = x_{J-1}, x_{-1} = x_{J-1}, x_{J+1} = x_1$$

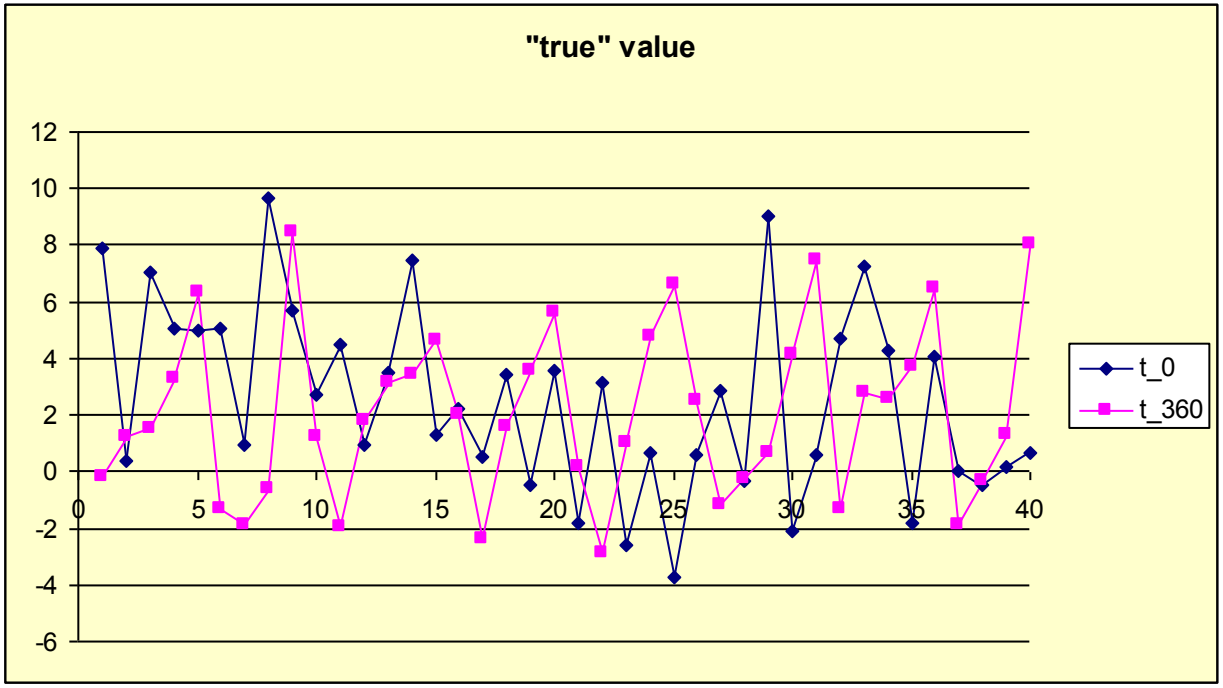
It is one of the elementary possible systems reproduces properties of many atmospheric models. For the solution of system of the equations 4th order Runge-Kutta scheme is used. $\Delta t = 0,05$, that corresponds to 6 hours (t=1 corresponds to 5 days), F=8.

Lorenz model

«Meteorological» properties appears after n 6-hour steps.
In the paper (Lorenz, Emmanuel, 1998) n=7200.
For "true" value modeling the following initial field has been considered:

$$x_0^t \cong N(F/4; F/2)$$

and forecast for n=360 time steps has been made (90 days).



Numerical experiments

Numerical experiments with Lorenz model include the following calculations:

1. x_t - «true» value. The forecast by initial value $x_0^t \cong N(F/4; F/2)$ for 90 days.
2. $x_d(0) = x_t(0) + \Delta, \Delta \sim N(0, s_0)$ - «known» initial value, not equal to "true" value.
3. Ensemble of initial fields $x^i(0) = x_d(0) + \Delta^i, \Delta^i \sim N(0, s_0)$.
4. Observations: $y_0 = Hx_t + \Delta_0, \Delta_0 \sim N(0, r_0)$.
5. Ensemble of observational errors (perturbed observation) for P_a modeling $y_0^{(i)} = y_0 + \Delta_0^i, \Delta_0^i \sim N(0, r_0)$.
6. The observations are modeling at every time step in ¼ part of region.

Numerical experiments

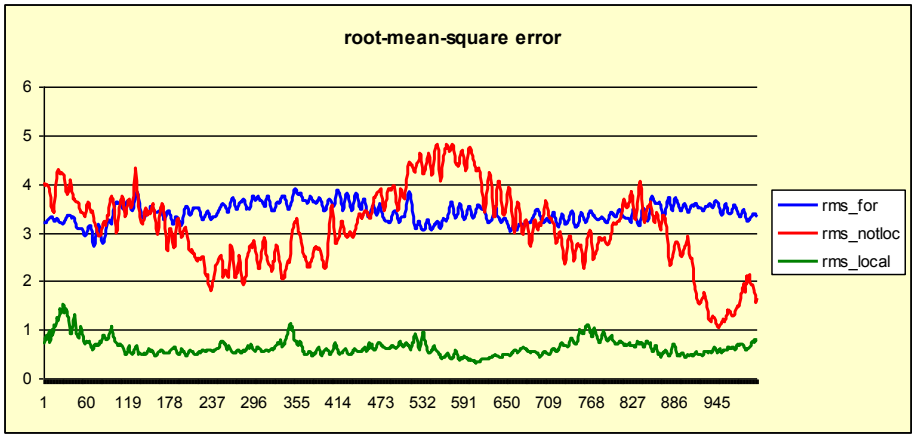
The first series of experiments.

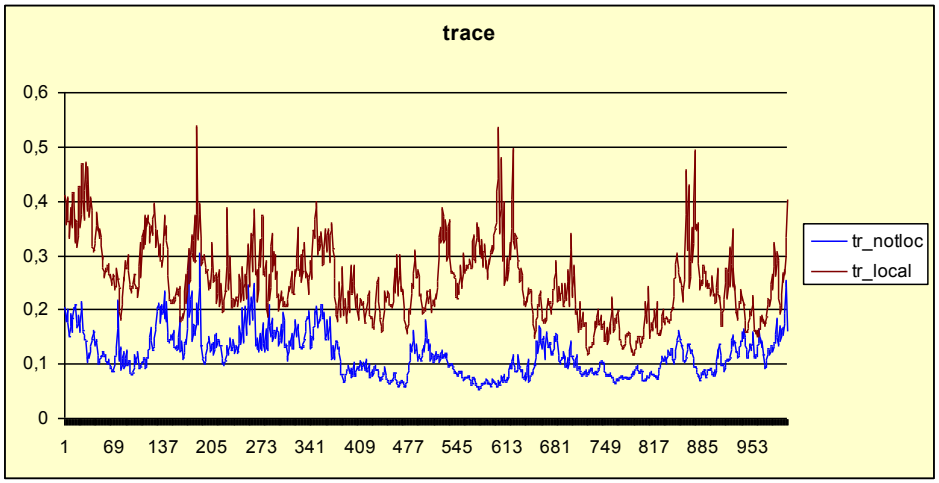
In numerical experiments the following values of parameters have been used:

$r_0 = F/40, s_0 = r_0, N_{ens} = 20, F = 8, F1 = F*0{,}95, N_t = 2000.$

Data assimilation has been made on every time step. Observations have been analised iteratively.

In figures the root-mean square error and trace of covariance matrix for last 1000 time steps are presented





Numerical experiments

The second series of experiments.

In 1st series of experiments it is visible, that algorithm not «diverged», there is no error growth, but after such big time (it is more than year) root-mean-square values and trace values leave on certain asimptotic level ($\mathrm{rms} \approx r_0$).

For this reason the second series of experiments has been organised in a different way.

The following values of parameters have been used

$r_0 = 1, s_0 = r_0, N_{ens} = 20, F = 8, F1 = F * 0{,}95, N_t = 7200.$

The data assimilation has been made for time steps $n = N_t, \cdots, N_t + 200$

Numerical experiments

The adaptive correction of ensembles of the forecast errors was carried out for prevention of fast "divergence" of algorithm. The adaptive algorithm consists in the following:

$$\sigma_k^2 \cong \sum_{i=1}^{N} \left(dh_k^i \right)^2 /(N-1)$$

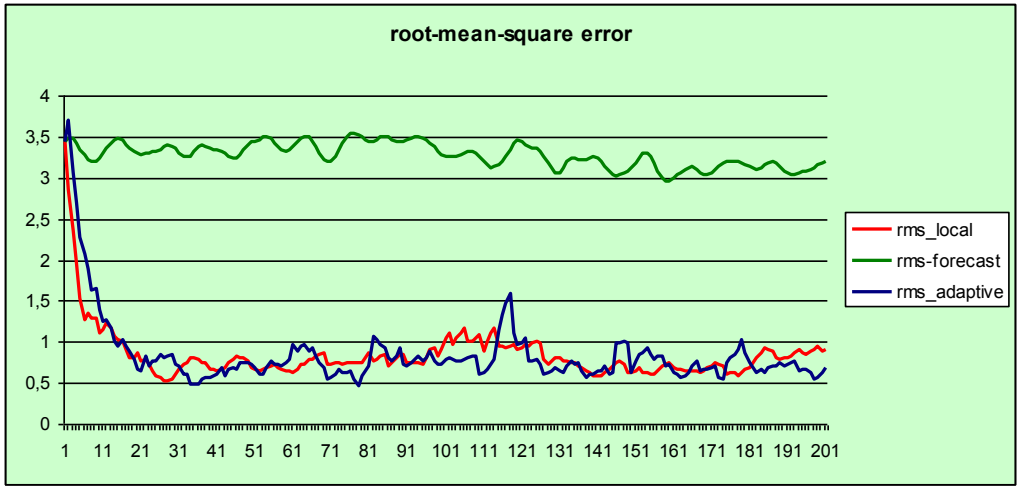
Let's note "residual" by r, $r = y_0 - Hx_f$. It's well known, that $\overline{rr^T} = HP_f H^T + R$. So we will concider, that $\tilde{\sigma}_i^2 = r_i^2 - r_0^2$ is "observations" of variance at observational point. Than correction of variance value in grid point may be made by the formula, which is a variant of simplified analyses:

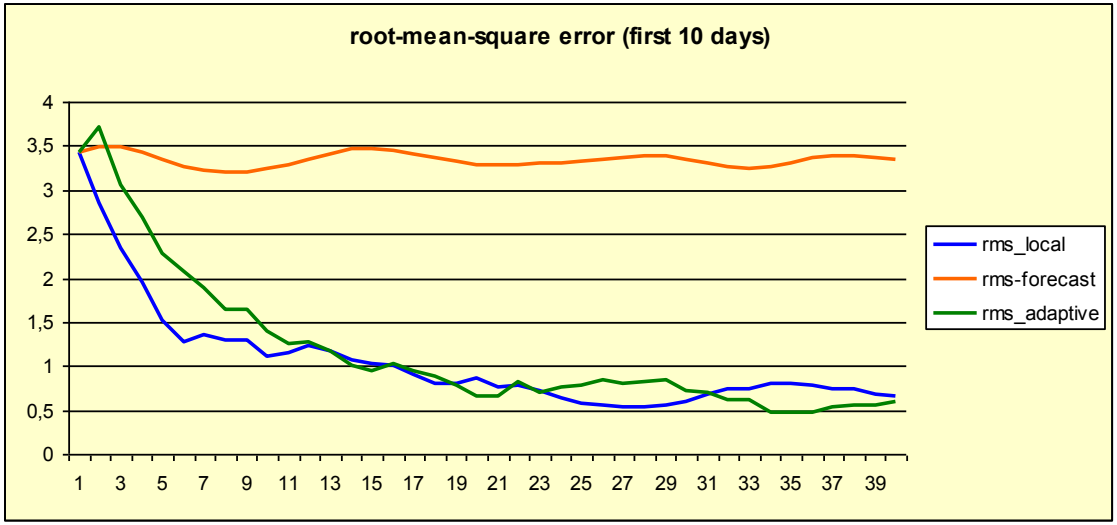
$$\tilde{\sigma}_k^2 = \sigma_k^2 + \sum_{i_0=1}^{I} (\tilde{\sigma}_{i_0}^2 - \sigma_{i_0}^2) e^{-\alpha \rho^2(k,i_0)} / \sum_{i_0=1}^{I} e^{-\alpha \rho^2(k,i_0)} .$$

After that the correction of ensemble of the forecast errors can be made as follows $dh_k^i = dh_k^i \left(\frac{\tilde{\sigma}_k^2}{\sigma_k^2} \right)^{1/2}$

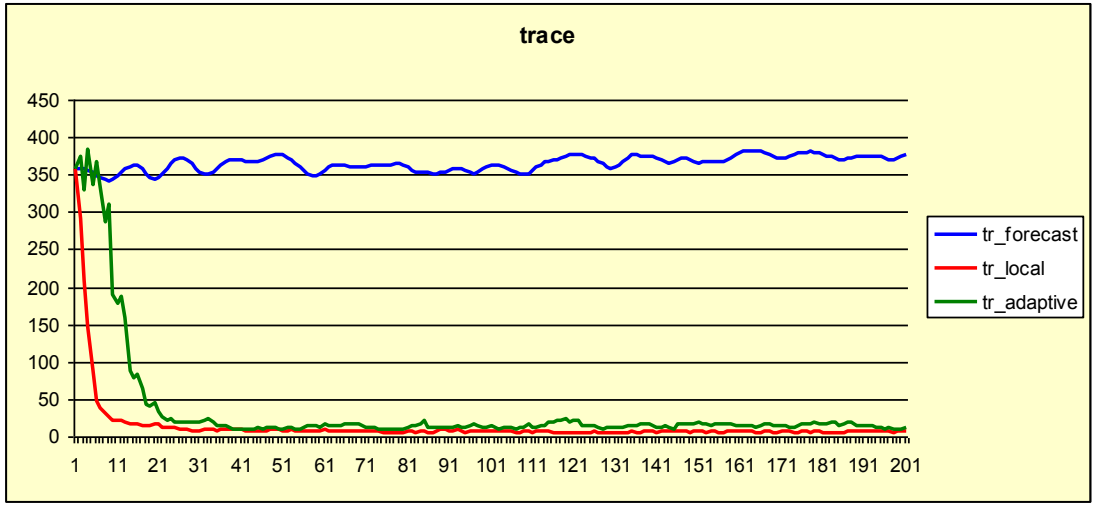
and, accordingly, the matrix P_f will change.

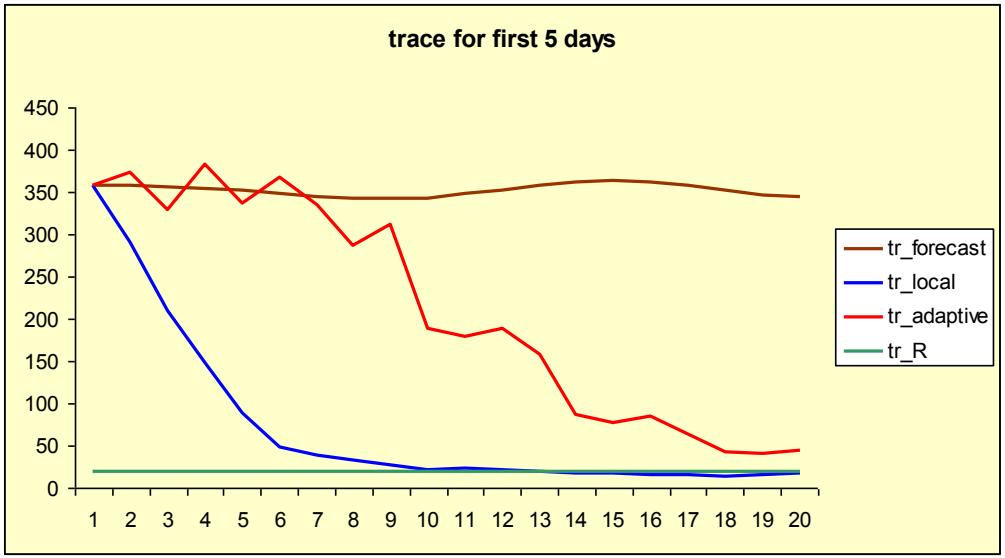
Numerical experiments





Numerical experiments





Final conclusions:

1) In the new forecast process, ensemble forecasting occupies a central place in the entire process following the observing, data assimilation and numerical modeling components.

2) The major part of the modern forecast system is data assimilation.

3) Algorithms of the data assimilation based on the dynamical-stochastic approach, allows to solve a problem of generation perturbations corresponding to the analysis error.

4) The major problems are:

- Small number of ensemble members;
- The theoretical justification of data assimilation algorithms in case of nonlinear operators of forecast and observations;
- The estimation of random errors of model and observations;
- «Adaptive observations».

Thank you for attention!

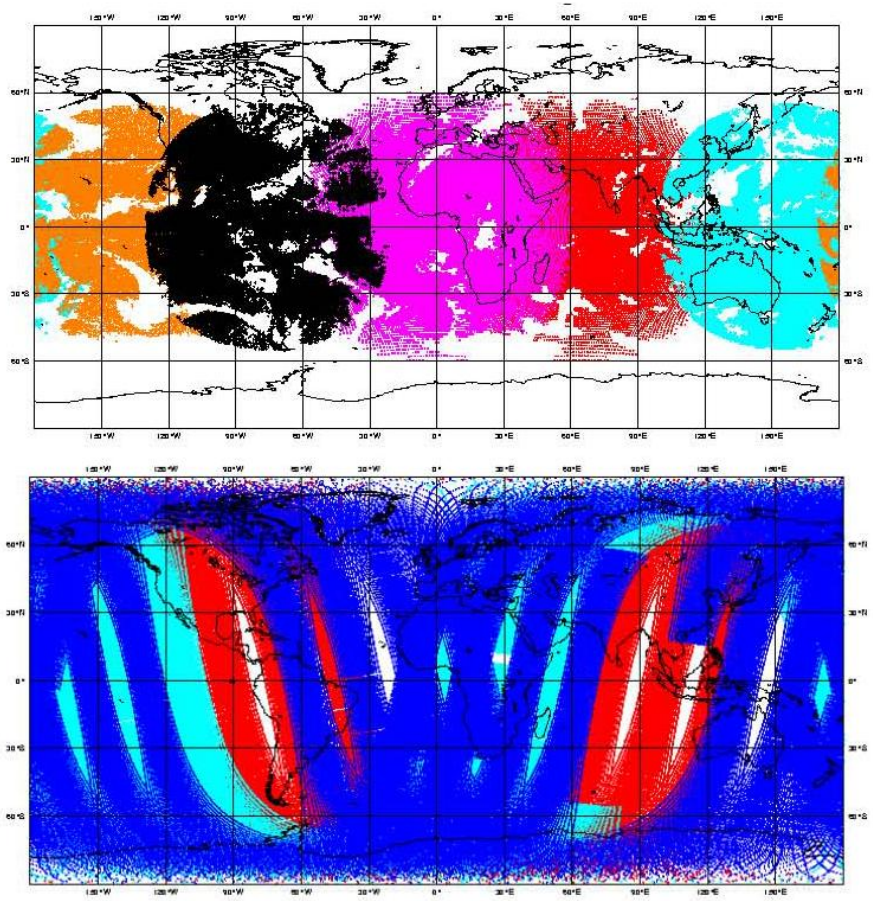


Figure 1: Typical data coverage provided by the Geostationary constellation (top): GOES-W/E (orange/black), Meteosat-7/5 (pink/red) and GMS-5/GOES-9 (cyan). Bottom plot displays the LEO constellation from the NOAA satellites (NOAA-15 in red, NOAA-16 in cyan, NOAA-17 in blue).