A data assimilation tutorial based on the Lorenz-95 system

Martin Leutbecher

Training Course 2010

Technical setup

- open a terminal window and execute `/scratch/rd/nel/public/kf_install`
- open a **new terminal** and execute the command `tutorial`
- click on KFX in the menu bar of the *scilab console* window
- click OK in first menu and click OK in second menu

Introduction

Why a tutorial with the Lorenz-95 system?

- The L95 system is a simple dynamical system that exhibits chaotic dynamics in a spatially extended domain.
- Appears frequently in publications on data assimilation methods where an idealized setting is desired.
- It could help you to understand, repeat and extend some of the published studies.
- The system is small enough to run an extended Kalman filter for many assimilation cycles on a personal computer. Yet, it is not too small to be completely trivial. However, beware! The conclusions may not apply to a particular real data assimilation problem . . .

What to take home from this tutorial?

- The scilab code for more in depth studies after the Training Course
- Some experience on how a Kalman Filter works for this simple system
- Some idea on the potential usefulness of flow-dependent background error covariances

Outline

1. The Lorenz-95 model and its dynamics
2. Extended Kalman filter
3. Assimilation experiments
   - Brief user guide
   - Single-observation experiments with the KF
   - KF divergence and the role of Q
   - Analysis uncertainty and observation error
   - Misspecification of R
   - Misspecification of bg-err variance
   - Comparison of 3 assimilation schemes
   - Long-window 4D-Var and KF
4. Outlook

The Lorenz-95 system

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F$$

with $i = 1, 2 \ldots N$, cyclic boundary conditions and a forcing of $F = 8$. Here, $N = 8, 12, 16$ or 40.

- nonlinear system introduced by E. Lorenz in 1995
- allegorical for dynamics of “weather” at a fixed latitude
- similar error growth characteristics as full NWP models if a time unit $\Delta t = 1$ is associated with 5 days
- “eastward” group velocities (increasing i)
- numerical integration with 4th order Runge-Kutta , $\Delta t = 0.025$

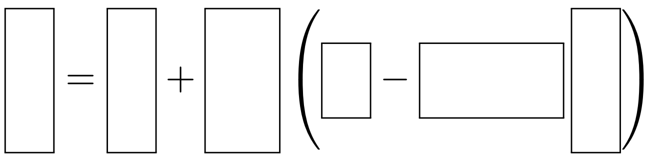
Illustration of the chaotic dynamics

L95

- growth of errors with forecast range (error growth curve and plumes). Quasi-exponential phase . . . saturation.
- state dependence of predictability: try different unperturbed initial conditions
- "eastward" (that is to higher indices) propagation of perturbations (Hovmueller plot of ensemble spread and plumes at different locations). Use localized perturbations.

Linear state estimation

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$$



menu	assimilation scheme	gain matrix $\mathbf{K}$
DI	direct insertion	$\mathbf{H}^{\mathrm{T}}$
OI	optimum interpolation	$\mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}})^{-1}$
KF	extended Kalman filter	$\mathbf{P}^f\mathbf{H}^{\mathrm{T}}(\mathbf{R}+\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}})^{-1}$

$\mathbf{x}^a$ and $\mathbf{x}^b$	n-by-1	state vector of analysis and background
$\mathbf{y}$	m-by-1	vector of observed values
$\mathbf{H}$	m-by-n	observation operator
$\mathbf{K}$	n-by-m	gain matrix
$\mathbf{P}^f$	n-by-n	predicted background error covariances
$\mathbf{B}$	n-by-n	static background error covariances
$\mathbf{R}$	m-by-m	observation error (co-)variances

extended Kalman filter equations

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b) \tag{1}$$

$$(\mathbf{P}^a)^{-1} = (\mathbf{P}^f)^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H} \tag{2}$$

$$\mathbf{x}^b = M(\mathbf{x}^a) \tag{3}$$

$$\mathbf{M} = \mathbf{M}(\mathbf{x}^a) \tag{4}$$

$$\mathbf{P}^f = \mathbf{M}\,\mathbf{P}^a\,\mathbf{M}^{\mathrm{T}} + \mathbf{Q} \tag{5}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^{\mathrm{T}}(\mathbf{R} + \mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}})^{-1} \tag{6}$$

where function M corresponds to the integration of the nonlinear model from one observation time to the next. Further, matrix $\mathbf{P}^a$ denotes the analysis error covariances, $\mathbf{Q}$ the estimate of the model error covariances (here static σ_q^2 times identity) and $\mathbf{M}$ the propagator of the tangent-linear model. The propagator $\mathbf{M}$ is linearised around a forecast started from the analysed state $\mathbf{x}^a$.

The equations are solved numerically in the succession (1)–(6).

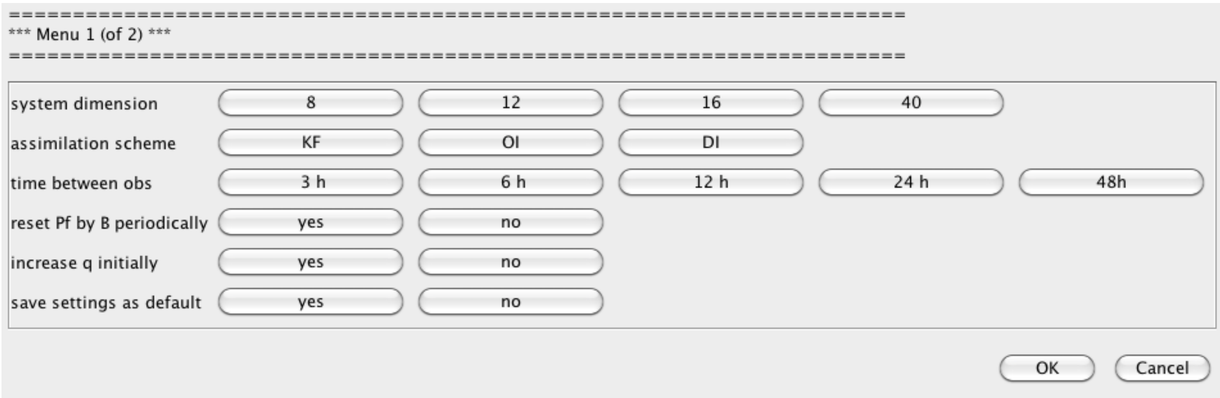
Assimilation experiments

- Perfect model scenario: One long integration of the Lorenz-95 system is the truth. The forecast model is identical to the model defining the truth.
- Observations are generated by taking data from the truth trajectory and adding noise in order to represent observation errors. The observation errors are uncorrelated (in time), unbiased and normally distributed with standard deviation σ_o.

How to run the assimilation experiments

Menu 1

click on KFX



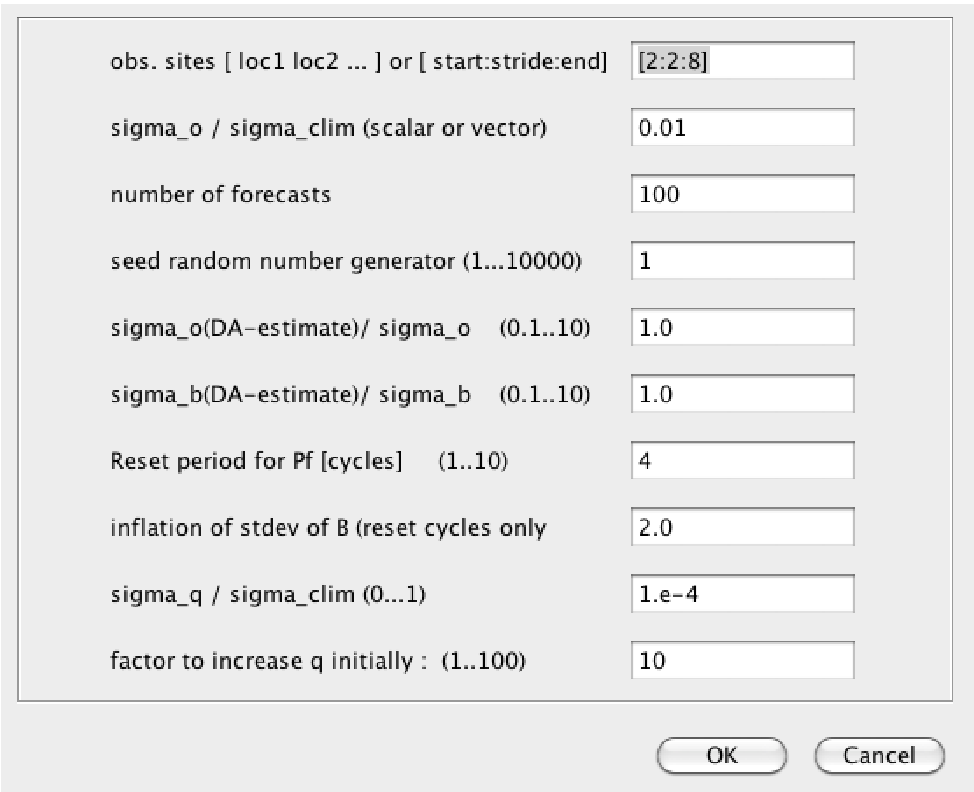
===
*** Menu 1 (of 2) ***
===

system dimension	8	12	16	40	
assimilation scheme	KF	OI	DI		
time between obs	3 h	6 h	12 h	24 h	48h
reset Pf by B periodically	yes	no			
increase q initially	yes	no			
save settings as default	yes	no			

OK Cancel

How to run the assimilation experiments

Menu 2



obs. sites [loc1 loc2 ...] or [start:stride:end]	[2:2:8]
sigma_o / sigma_clim (scalar or vector)	0.01
number of forecasts	100
seed random number generator (1...10000)	1
sigma_o(DA-estimate)/ sigma_o (0.1..10)	1.0
sigma_b(DA-estimate)/ sigma_b (0.1..10)	1.0
Reset period for Pf [cycles] (1..10)	4
inflation of stdev of B (reset cycles only	2.0
sigma_q / sigma_clim (0...1)	1.e-4
factor to increase q initially : (1..100)	10

OK Cancel

How to run the assimilation experiments

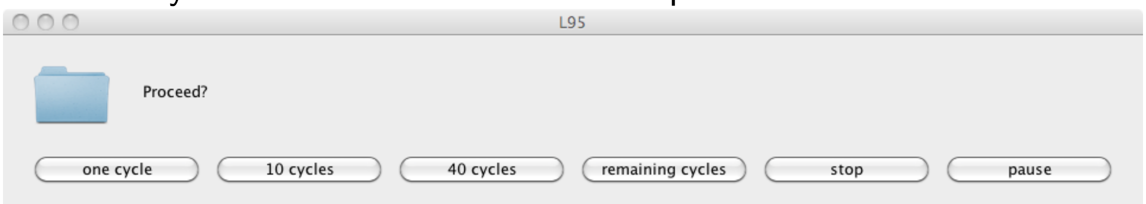
Output

- Each time $\boxed{\text{KFX}}$ is executed, a unique experiment identifier of the form `AAnn_xxx` is generated, where
  - `AA` is KF, OI or DI, i.e. the assimilation scheme,
  - `nn` is the dimension of the system (08, 12, 16, 40),
  - `xxx` is a counter that is incremented by one each time this configuration is run.
- Output is written to directory `$SCRATCH/L95data/<AAnn_xxx>`
  - The input to the two menus is saved in `kfx.menu1/2.defaults`
  - The summary information appearing in the scilab console is saved in `summary.txt`
  - Some plots are saved in pdf-files.
  - For instance, `KF08_014/pf0123.pdf` contains a plot of $\mathbf{P}^f$ of Experiment KF08_014 in assimilation cycle 123.
  - If the experiment uses the KF for assimilation (without resetting $\mathbf{P}^f$), the average background error covariance matrix is saved in `pf_mean.r8`.

Kalman filter, single obs experiments — exercises 1

Use dim=8; 1 obs every 6 h; $\sigma_o = 0.01\sigma_{\rm clim}$; $\sigma_q = 10^{-4}\sigma_{\rm clim}$

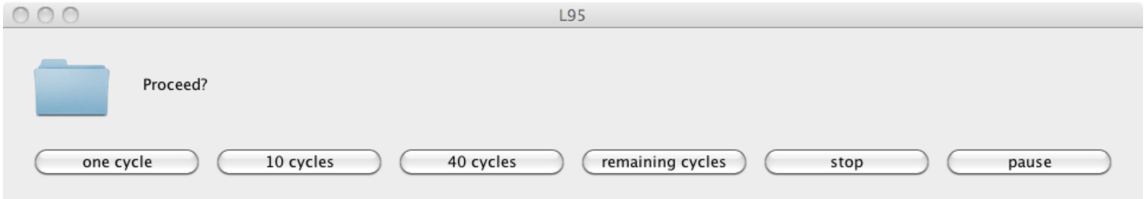
- Run one cycle at a time and look at the plots



  - truth, observed value y, background $\mathbf{x}^b$ and analysis $\mathbf{x}^a$
  - analysis increment $\mathbf{x}^a - \mathbf{x}^b = \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$
  - error covariances $\mathbf{P}^a$ and $\mathbf{P}^f$
- When are there cases with a very small increment?
- Is the analysis always more accurate than the background?
- Observe the variations in the shape of the analysis increment.
- Can you relate the structure of the analysis increment to $\mathbf{P}^f$?
- Compare the diagonal elements of $\mathbf{P}^a$ and $\mathbf{P}^f$.
- How close draws the analysis to observed value and background? Why does this vary considerably between cycles?

Kalman filter, single obs experiments — exercises 2

- Run "remaining cycles" and look at the plots



  - evolution of the bg and an-error covariances
  - time series plots (type q or a in the scilab console)
  - the global rms errors (table in scilab console)
- Does the predicted analysis (background) error variance match the actual RMS error of analysis (background)?

Kalman filter divergence

Use dim=8; 1 obs every 6 h; $\sigma_o = 0.01\sigma_{\rm clim}$;

- What happens as σ_q approaches zero?
- What is the smallest σ_q for which the KF works?
- Why is $\mathbf{Q}$ required in the perfect model scenario?

Use dim=12, 16, 40; obs every 6 h; $\sigma_o = 0.01\sigma_{\rm clim}$; $\sigma_q = 10^{-3}\sigma_{\rm clim}$

- How many observations are required to avoid the divergence of the KF for the 12, 16 and 40-dimensional system?
- Are some some spatial distributions of n observations more favourable for producing a good analysis than others?

Analysis uncertainty and observation error

Use dim=40; 20 obs ([2 : 2 : 40]) every 6 h; assimilate with KF

- Start with observation error stdev $\sigma_o = 0.1\sigma_{\mathrm{clim}}$.
- A suitable model error term for this σ_o is $\sigma_q = 10^{-3}$.
- Run all cycles and note RMS error of analysis.
- Now repeat with smaller values of obs. error stdev e.g. $\sigma_o/\sigma_{\mathrm{clim}} = 0.05, 0.01, 0.001$. Adapt σ_q to get the lowest possible analysis error.
- How does the RMS error of the analysis depend on σ_o?
- Would you expect a similar behaviour for a real forecast system?

Misspecification of observation error variance

- What happens if the assumed $\tilde{\sigma}_o$ in the KF is larger or smaller than the actual σ_o?
- Use, e.g. dim=40; 20 obs ([2 : 2 : 40]) every 6 h; $\sigma_o = 0.01\sigma_{\mathrm{clim}}$
- Now, inflate the observation error standard deviation used in the KF by setting the factor `sigma_o(DA-estimate)/ sigma_o` to a value larger than 1.
- How accurate is the analysis compared to the experiment with correct observation error standard deviation in the KF?
- What happens to $\mathbf{P}^f$ and σ_b?
- You may be able to understand the result by looking at the KF equations assuming that $\mathbf{Q}$ is scaled by the same factor as $\mathbf{R}$.
- Repeat with deflated observation error standard deviation.

Misspecification of background error variance

- What happens if the background error variances are inflated or deflated?
- Use different values for `sigma_b(DA-estimate)/ sigma_b`. Try moderate values between 0.9 and 1.1.
- Note that the modified $\mathbf{P}^f$ is cycled.
- What happens to the accuracy of the analysis?

KF, Optimum Interpolation and Direct Insertion

Dense observation network

- Compare the three different assimilation schemes for the same observing network. Note the analysis error for each assimilation experiment.
- Use KF, dim=40, obs at every location every 6 h, $\sigma_o/\sigma_{\mathrm{clim}} = 0.01$ as reference. Small value of σ_q.
- The time-average background error covariance matrix $\overline{\mathbf{P}^f}$ is saved. In subsequent runs of the OI scheme, this matrix is read in as the *static* background error covariance matrix $\mathbf{B}$.
- The actual background error variance obtained with OI is different from the average variance of the KF. This can be accounted for by inflating $\mathbf{B}$ by setting `sigma_b(DA-estimate)/sigma_b` to a suitable value larger than 1.
- Which inflation factor yields the most accurate analysis with the OI scheme?
- How different are the analysis errors of KF, OI and DI?

KF, Optimum Interpolation and Direct Insertion

Sparser observation networks

- Repeat the comparison of the three different assimilation schemes for different observing networks:
  1. moderately sparse network: one obs. every 2nd location: $[2:2:40]$
  2. fairly sparse network: one obs. every 4th location: $[4:4:40]$
- Use dim=40, obs every 6 h, $\sigma_o/\sigma_{\text{clim}} = 0.01$ as reference. Small value of σ_q.
- As the time-averaged background error covariance matrix is saved, the KF experiment needs to be run first.
- Tune again the variance of $\mathbf{B}$ in the OI scheme by setting `sigma_b(DA-estimate)/sigma_b` to a suitable value larger than 1.
- Which inflation factor yields the most accurate analysis with the OI scheme?
- How different are the analysis errors of KF, OI and DI now for the sparser networks?

Long-window 4D-Var and Kalman filter

Time scale for the spin-up of $\mathbf{P}^f$

Use: dim=8; 4 obs every 2nd site $[2:2:8]$ and every 6 h; $\sigma_o/\sigma_{\text{clim}} = 0.01$

- Run the KF without resetting $\mathbf{P}^f$. (This generates $\mathbf{B} = \overline{\mathbf{P}^f}$.)
- Now specify in Menu 1, that you would like to reset $\mathbf{P}^f$. Set the period (in number of assimilation cycles) for resetting $\mathbf{P}^f$ in Menu 2. Try, e.g. 2, 4, 8, 20, 40. (Hint: you may need to inflate $\mathbf{B}$.)
- Run up to cycle 161 (proceed 40 cycles). Compare covariance matrices stored as pdf file in experiment directory.
- How many cycles are required to spin up the system and obtain a good estimate of $\mathbf{P}^f$?
- How does the RMS analysis error depend on the reset period?

Outlook

get your own copy of the scilab-code:

```
cd /home/unix/presentations/TC_NWP2010/DA/leutbecher
mail -a DA-TC.tar.gz   my_name@nwp.org
SUBJECT: your subject
email text here
finish and send with
CTRL-D
```

∗ go through the code and try to understand it
∗ extend the programmes yourself and investigate

- model error
- more complicated observing networks, e.g. temporally varying observation coverage
- biased observations and bias estimation
- observational quality control
- variational data assimilation (tangent-linear and adjoint have already been coded, see `kf.sci`)
- comparison of different techniques for background error covariance estimation
- ensemble data assimilation algorithms

References

Fisher, M., M. Leutbecher, and G. A. Kelly, 2005: On the equivalence between Kalman smoothing and weak-constraint four-dimensional variational data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3235–3246.

Leutbecher, M., R. Buizza, and L. Isaksen, 2007: Ensemble forecasting and flow-dependent estimates of initial uncertainty. In *Flow-dependent aspects of data assimilation*, Workshop Proceedings, ECMWF, Shinfield Park, Reading, UK, 185–201. Available at www.ecmwf.int/publications/library/ecpublications/_pdf/workshop/2007/Data_assimilation.

Lorenz, E. N., 1995: Predictability: A problem partly solved. In *Seminar on Predictability*, volume Vol. I, ECMWF, Reading, UK, 1–18. Available at http://www.ecmwf.int/publications/library/ecpublications/_pdf/seminar/1995.

Lorenz, E. N., 2005: Designing chaotic models. *J. Atmos. Sci.*, **62**, 1574–1587.

Lorenz, E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399–414.

Wilks, D. S., 2005: Effects of stochastic parametrizations in the Lorenz '96 system. *Quart. J. Roy. Meteor. Soc.*, **131**, 389–407.

Tangent-linear and adjoint model

The propagator $\mathbf{M}$ is obtained by multiplying the single-time-step propagators for the 4-th order Runge-Kutta scheme:

$$\mathbf{M} = \mathbf{M}_\ell \, \mathbf{M}_{\ell-1} \ldots \mathbf{M}_2 \, \mathbf{M}_1. \tag{7}$$

The single time step propagator for the Runge-Kutta scheme is given by

$$\begin{aligned}\mathbf{M}_j = \mathbf{I} + \frac{1}{6}\Big[&(\mathbf{I} + (\mathbf{I} + (\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A}_3)\tfrac{1}{2}\Delta t\mathbf{A}_2)\Delta t\mathbf{A}_1)\Delta t\mathbf{A}_0 \\ &+ (2\mathbf{I} + (\mathbf{I} + \tfrac{1}{2}\Delta t\mathbf{A}_3)\Delta t\mathbf{A}_2)\Delta t\mathbf{A}_1 + (2\mathbf{I} + \Delta t\mathbf{A}_3)\Delta t\mathbf{A}_2 + \Delta t\mathbf{A}_3\Big]\end{aligned} \tag{8}$$

Here $\mathbf{I}$ denotes the identity matrix and $\mathbf{A}_0$, $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{A}_3$ the Jacobi matrices of the right-hand-side of the ODE

$$dx_i/dt = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F \tag{9}$$

at $\mathbf{x}$, $\mathbf{x} + \frac{1}{2}\mathbf{q}_1$, $\mathbf{x} + \frac{1}{2}\mathbf{q}_2$ and $\mathbf{x} + \mathbf{q}_3$, respectively (standard Runge-Kutta notation).