Using Data Assimilation for Systematic Model Improvement

Matthew Lang
Laboratoire des Sciences du Climat et de l'Environnement (LSCE)

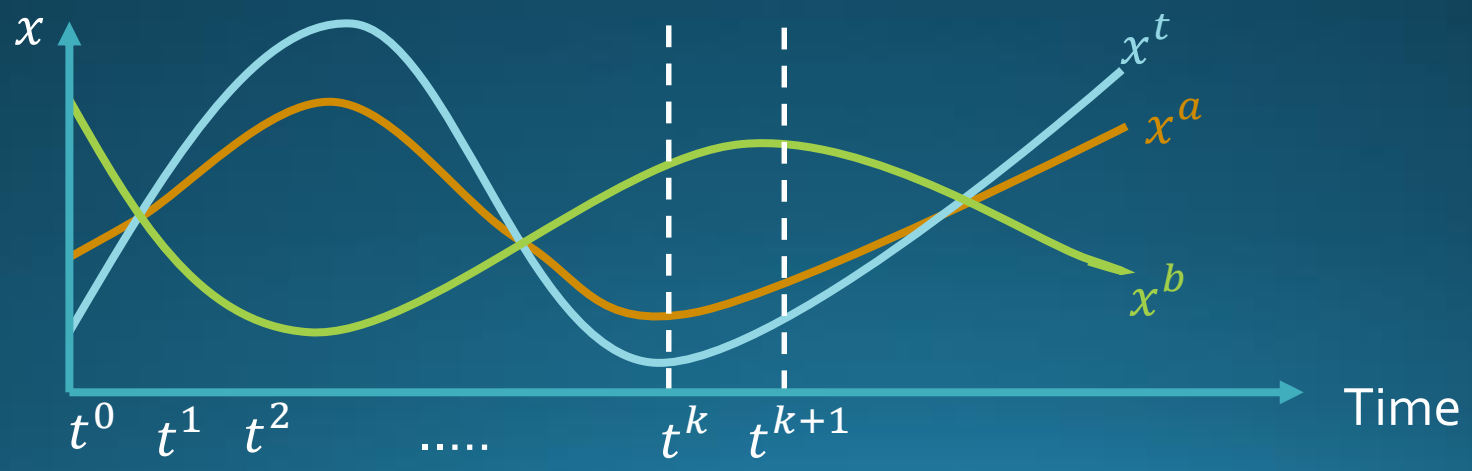
Peter Jan van Leeuwen, Phil Browne
University of Reading

Introduction

- All models of physical processes contain errors.
- Errors can be due to:
  - A lack of scientific understanding
  - A lack of computing power
  - eg. Sub-grid turbulence, radiation
- Parameterisations are simplified functions used in numerical models to account for these errors.
- Improving the errors in these parameterisations is currently done in an ad-hoc fashion, eg. via parameterisation tuning
- We propose a systematic method for estimating these parameterisation errors, through the use of state estimation.

Our method

- **Aim**: Estimate how the true state, x^t,evolves
- Background state, x^b, is generated using the prior model



Our method

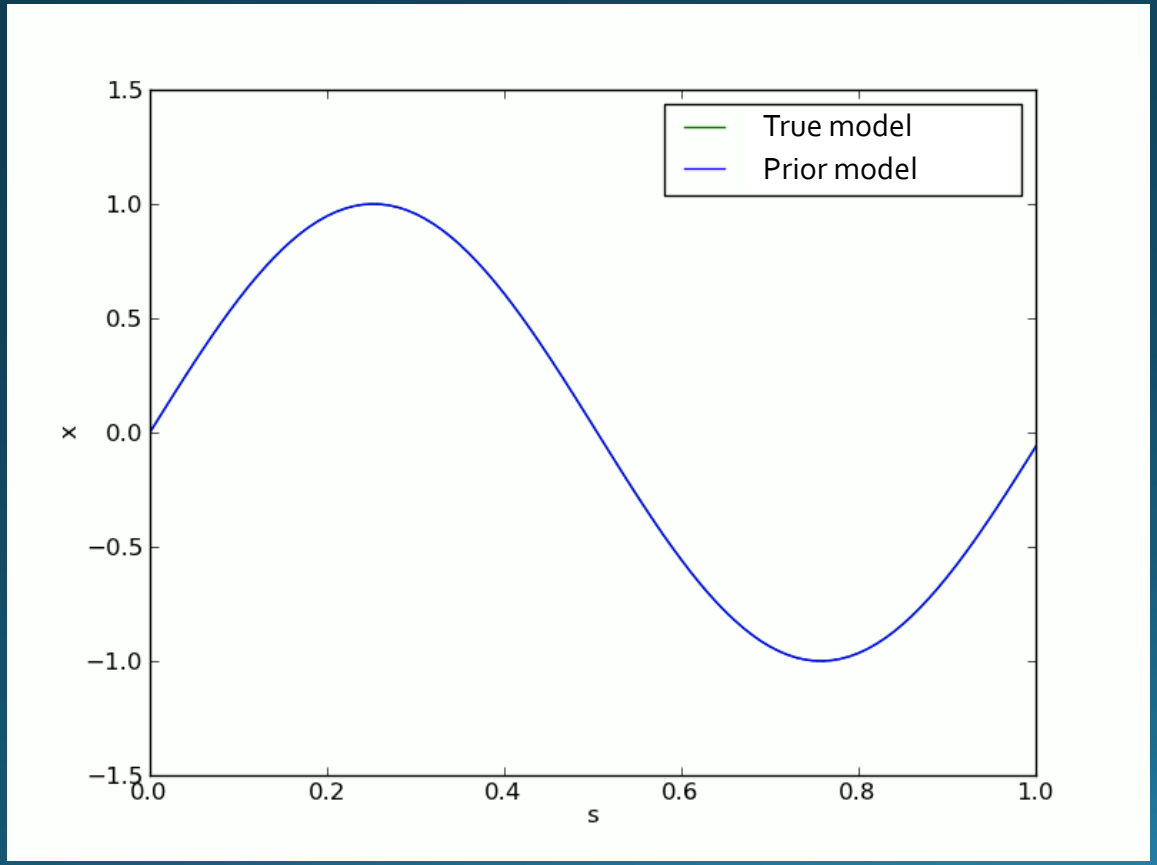
- Zooming into a single time step of the DA trajectory

Our method

- We create a new variable $\tilde{x}_{k+1}$.
  - Is the analysis state evolved one time step using the prior model
  - Represents the analysis forecast
- Compute the differences between x^a and $\tilde{x}$ over the whole domain at all timesteps

Motivating example

- This method is now applied to estimated the functional difference between a linear advection scheme and a nonlinear advection scheme over same domain.



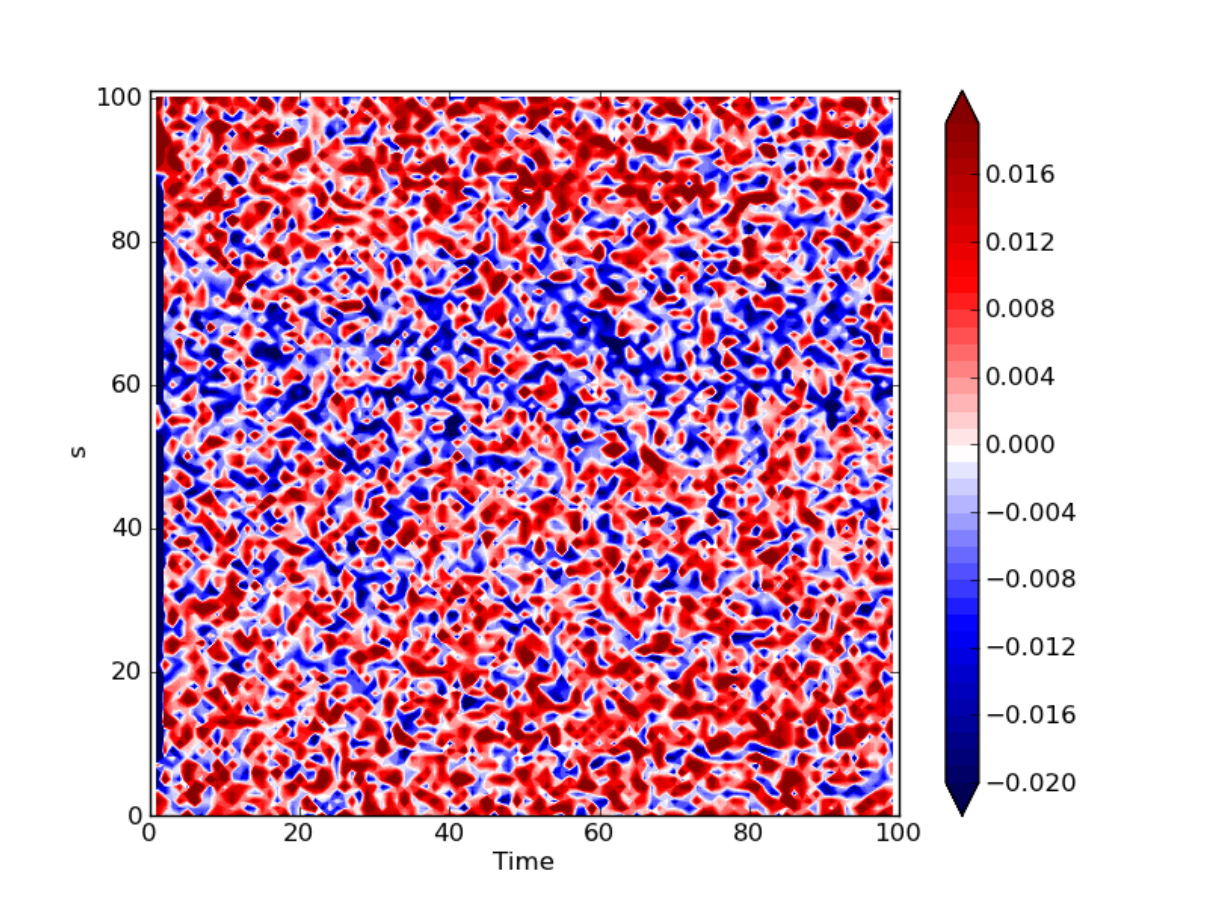
True model

$$\frac{\partial x}{\partial t} + (x+3)\frac{\partial x}{\partial s} = \xi_q$$

Prior model

$$\frac{\partial x}{\partial t} + 4.5\frac{\partial x}{\partial s} = \xi_q$$

Generating $x^a - \tilde{x}$



- The mean (over the ensemble) of $x^a - \tilde{x}$ plotted for each gridpoint and timestep generated by the 40 ensemble member EnKF using the prior model, with observations at all timesteps.

Extracting the model error structure

- To extract the model error present, it is required that the structure in $\overline{x^a - \tilde{x}}$ be determined

- To do this, a test function, $g(x)$, using any prior knowledge available, needs to be defined. This is denoted with the form:

$$g(x) = f_0(x, \alpha_0) + \cdots + f_n(x, \alpha_n)$$

- Very important this is specified well
  - Our method only searches in span of test function

Analysing test function

- Once an adequate test function is determined, the test function is split into separate sub-functions:

$$g_0(x) = f_0(x, \alpha_0)$$
$$g_1(x) = f_0(x, \alpha_0) + f_1(x, \alpha_1)$$
$$\vdots$$
$$g_n(x) = f_0(x, \alpha_0) + f_1(x, \alpha_1) + \cdots + f_n(x, \alpha_n)$$

- A regression analysis method is used to estimate optimal parameters, α_i, with uncertainties, for each of the $g_i(x)$ $(i = 0, 1, \ldots, n)$ that best fits the estimated model error, $\overline{x^a - \tilde{x}}$.

The Bayesian Information Criterion

- To assess the quality of the terms, the Bayesian Information Criterion (BIC) values are computed for each $g_i(x)$, where

$$BIC = \mathrm{k}\ \log N - 2\log \mathcal{L}$$

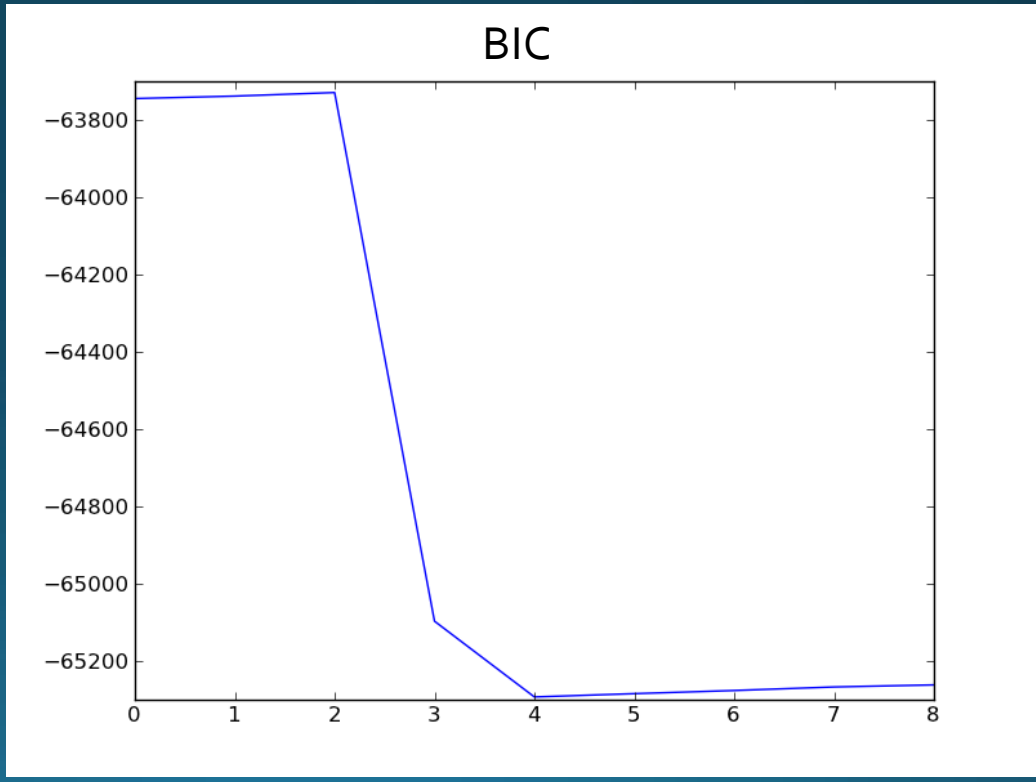
k = Number of terms, N=Number of fitting points in the regression, $\mathcal{L}$ is maximised likelihood function

- The BIC represents the trade-off between how well the sub-function fits the $\overline{x^a - \tilde{x}}$ field and the complexity of the model
- Smaller values of BIC indicate 'better' sub-functions
- The greatest decreases in BIC correspond to the terms with the 'most' information about the structure of the model error

BIC of Functional Estimates obtained

$$g(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 \frac{\partial x}{\partial s} + \alpha_4 x \frac{\partial x}{\partial s} + \alpha_5 x^2 \frac{\partial x}{\partial s} + \alpha_6 \frac{\partial^2 x}{\partial s^2} + \alpha_7 x \frac{\partial^2 x}{\partial s^2} + \alpha_8 x^2 \frac{\partial^2 x}{\partial s^2}$$

- Most information is added by terms 3 and 4 which represent the $\frac{\partial x}{\partial s}$ and $x \frac{\partial x}{\partial s}$ terms, respectively.
- Test function is reordered based on greatest decreases in BIC and regression is performed again

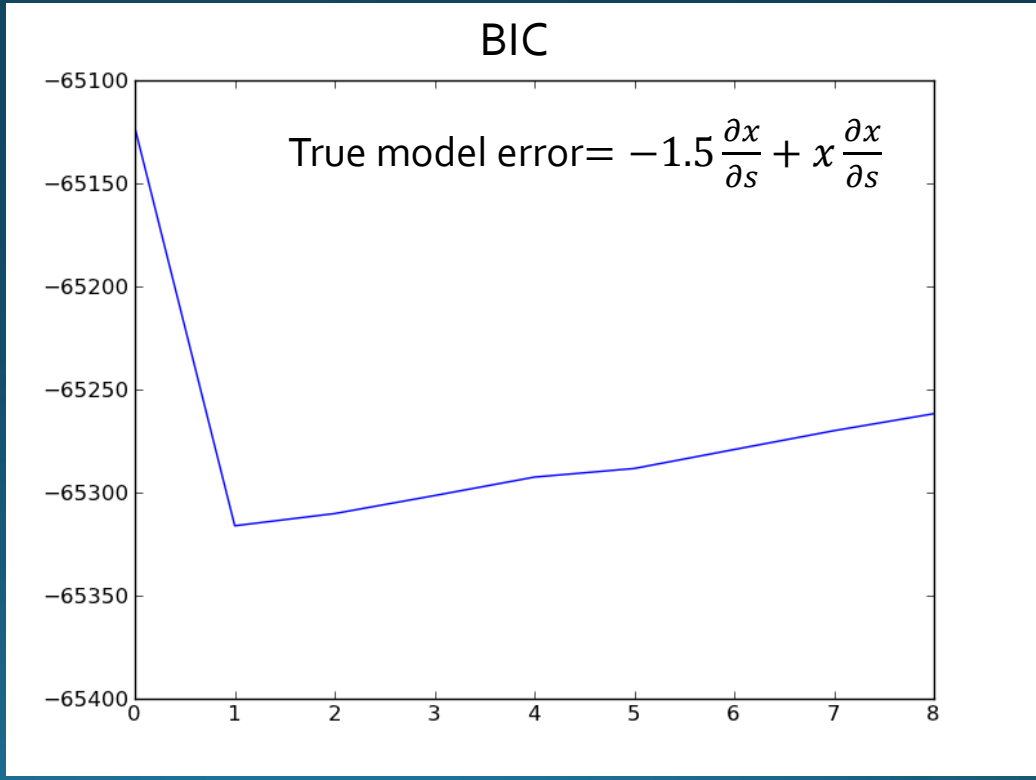


Reordered Functional Estimates obtained

$$h(x) = \alpha_0 \frac{\partial x}{\partial s} + \alpha_1 x \frac{\partial x}{\partial s} + \alpha_2 + \alpha_3 x \frac{\partial^2 x}{\partial s^2} + \alpha_4 x + \alpha_5 \frac{\partial^2 x}{\partial s^2} + \alpha_6 x^2 \frac{\partial x}{\partial s} + \alpha_7 x^2 \frac{\partial^2 x}{\partial s^2} + \alpha_8 x^2$$

- Minimum occurs after the addition of Term 1, corresponding to $x \frac{\partial x}{\partial s}$ term
- Optimal coefficients obtained by applying least squares to $\overline{x^a - \tilde{x}}$
- Optimal functional form of model error:

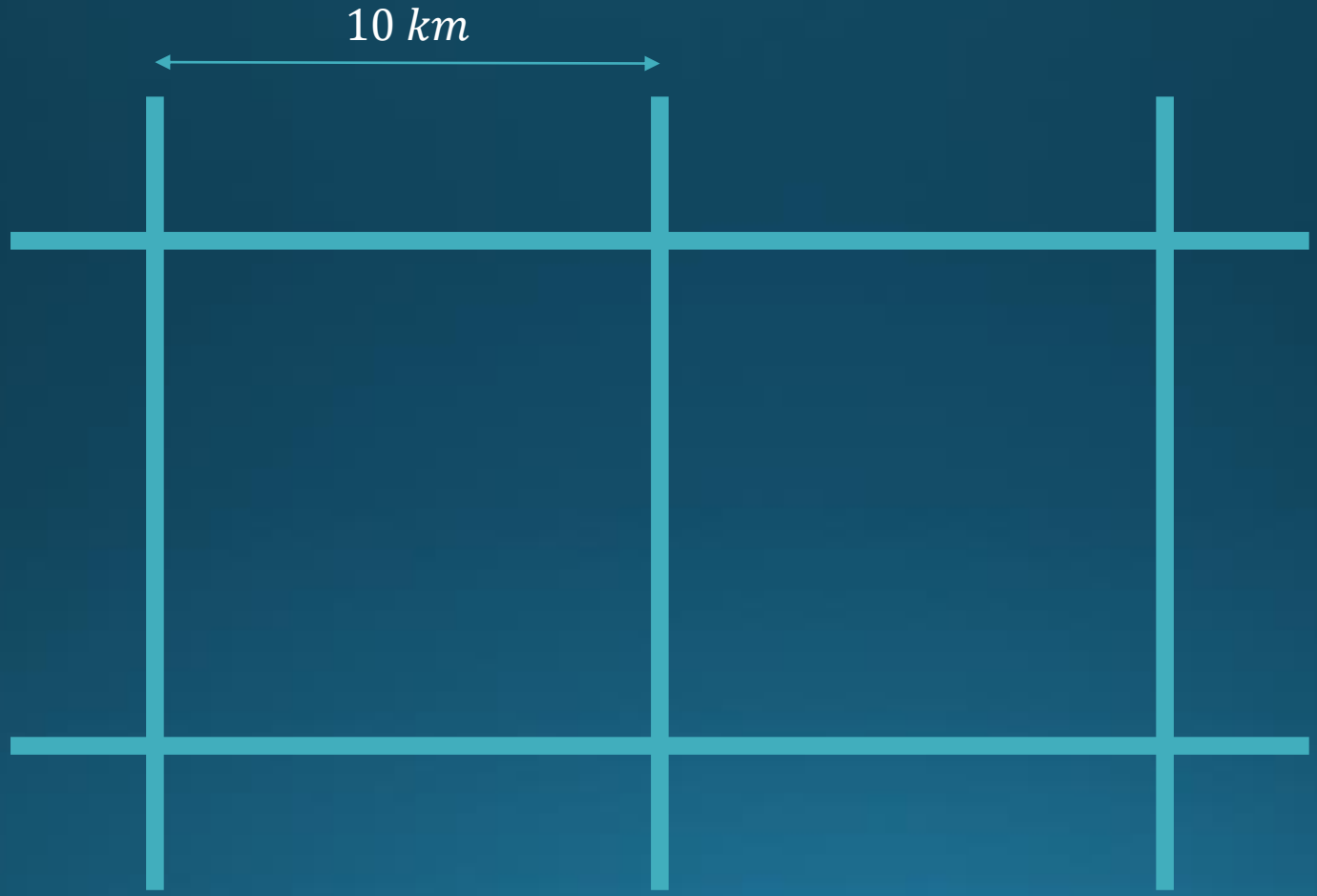
$$(-1.51 \pm 0.04) \frac{\partial x}{\partial s} + (1.03 \pm 0.07) x \frac{\partial x}{\partial s}$$



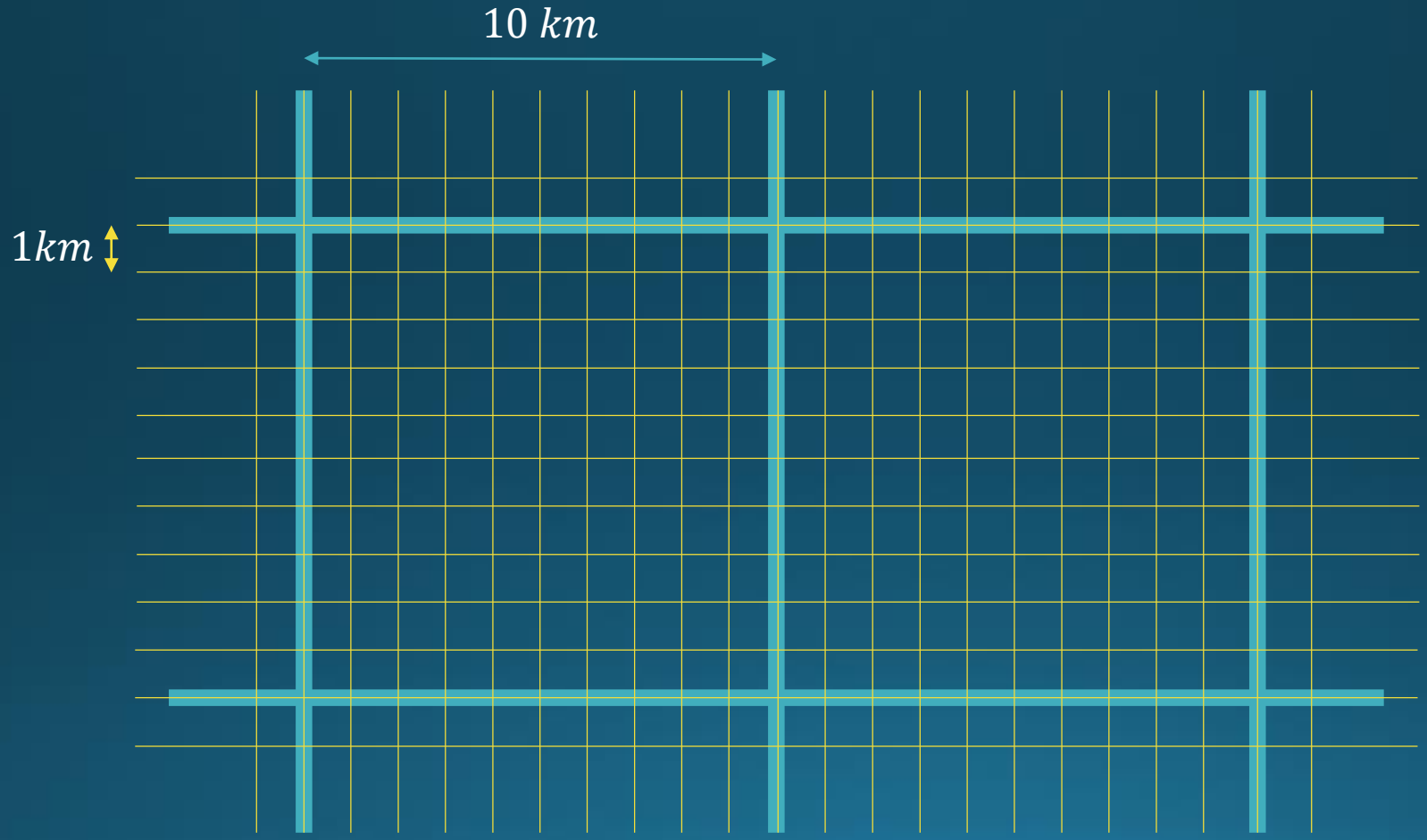
Discussion and future work

- Quality of model error parameterisations are heavily dependent upon the quality of the data assimilation method used.
- Works best when data assimilation updates the state at all gridpoints in the domain and at all timesteps.
- A high-resolution model can be used to ensure that all points of a lower resolution model are updated

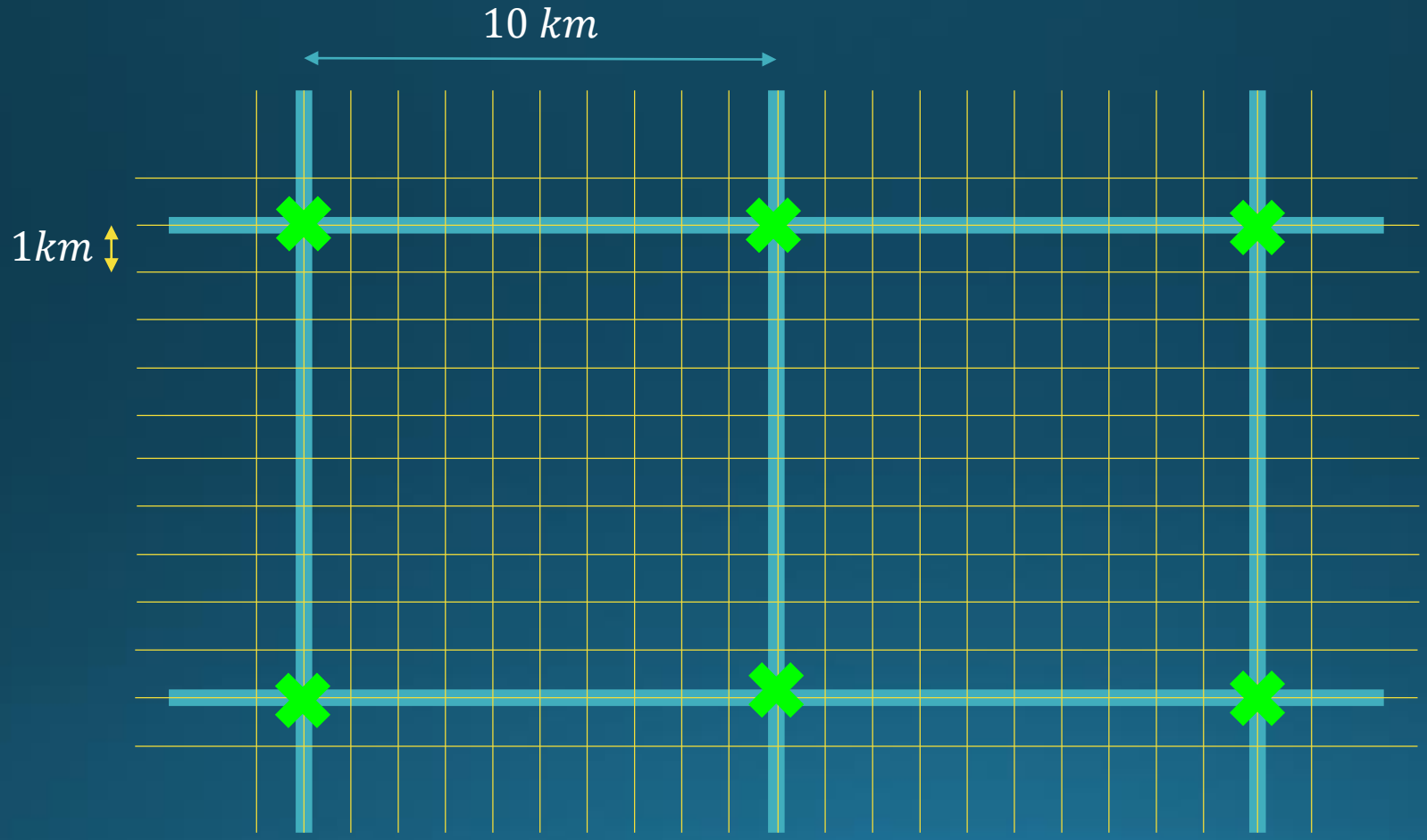
Use High Resolution models to improve lower resolution models



Use High Resolution models to improve lower resolution models



Use High Resolution models to improve lower resolution models



Remarks and Conclusions

- Parameterisation Estimation Method is shown to work well for the advection model
- Picks out the optimal functional form of the model error and generates uncertainties that contain the true model error
- The method is very dependent upon how well the DA scheme used performs
  - If DA is worse, error estimates are larger in a consistent manner.
- Future work is looking into applying this to higher resolution models to improve the parameterisations of lower resolution models.

Any questions?

Comparison with state augmentation

- State augmentation is used to estimate the model error between linear and nonlinear advection models with same uncertainty in relevant terms

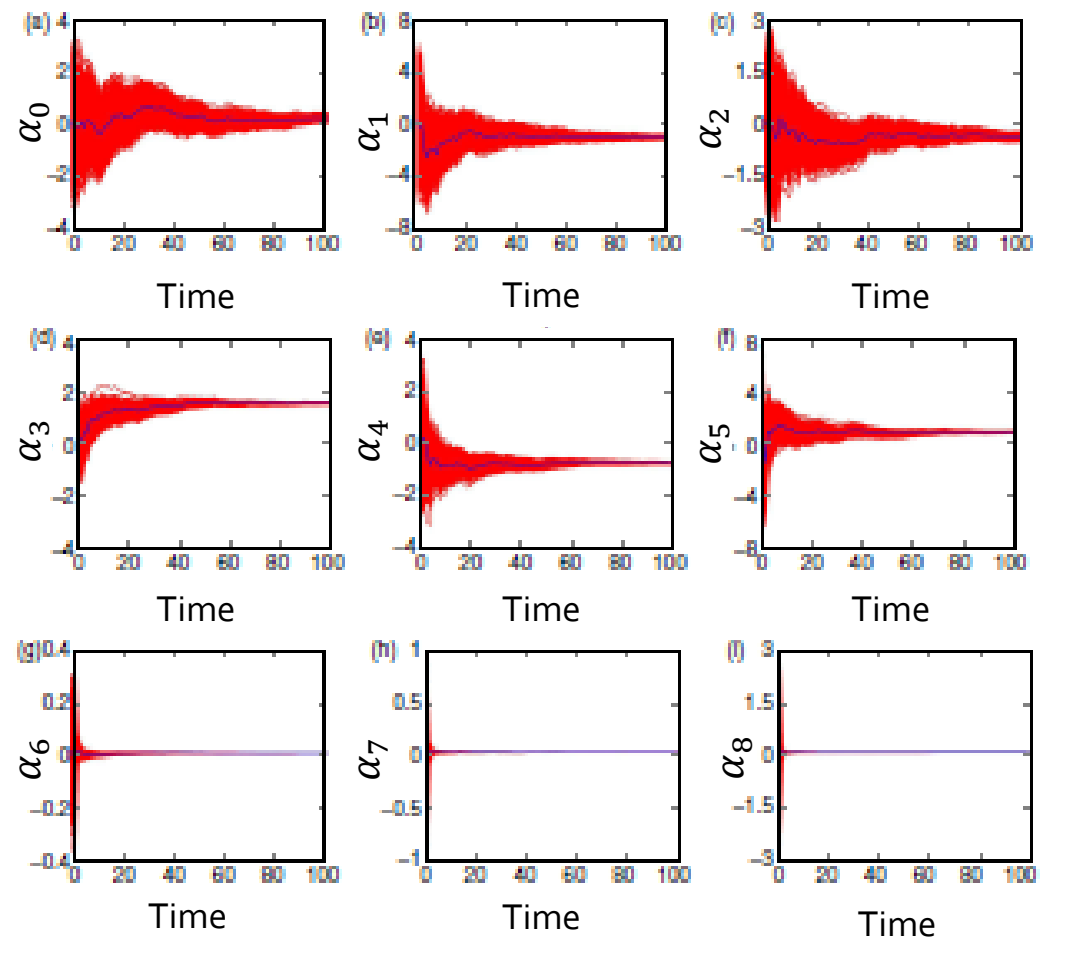
- Prior augmented model is:

$$\frac{\partial x}{\partial t} + 4.5\frac{\partial x}{\partial s} + \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 \frac{\partial x}{\partial s} + \alpha_4 x \frac{\partial x}{\partial s} + \alpha_5 x^2 \frac{\partial x}{\partial s}$$
$$+\alpha_6 \frac{\partial^2 x}{\partial s^2} + \alpha_7 x \frac{\partial^2 x}{\partial s^2} + \alpha_8 x^2 \frac{\partial^2 x}{\partial s^2} = \xi_q$$

where the α_i are parameters to be estimated by state augmentation, all initialised at 0, with initial variances specified to include true model error.

- 2000-member EnKF is used to adequately estimate augmented forecast error covariance matrix without P^f -localisation.

Comparison with state augmentation



Parameter Ensemble

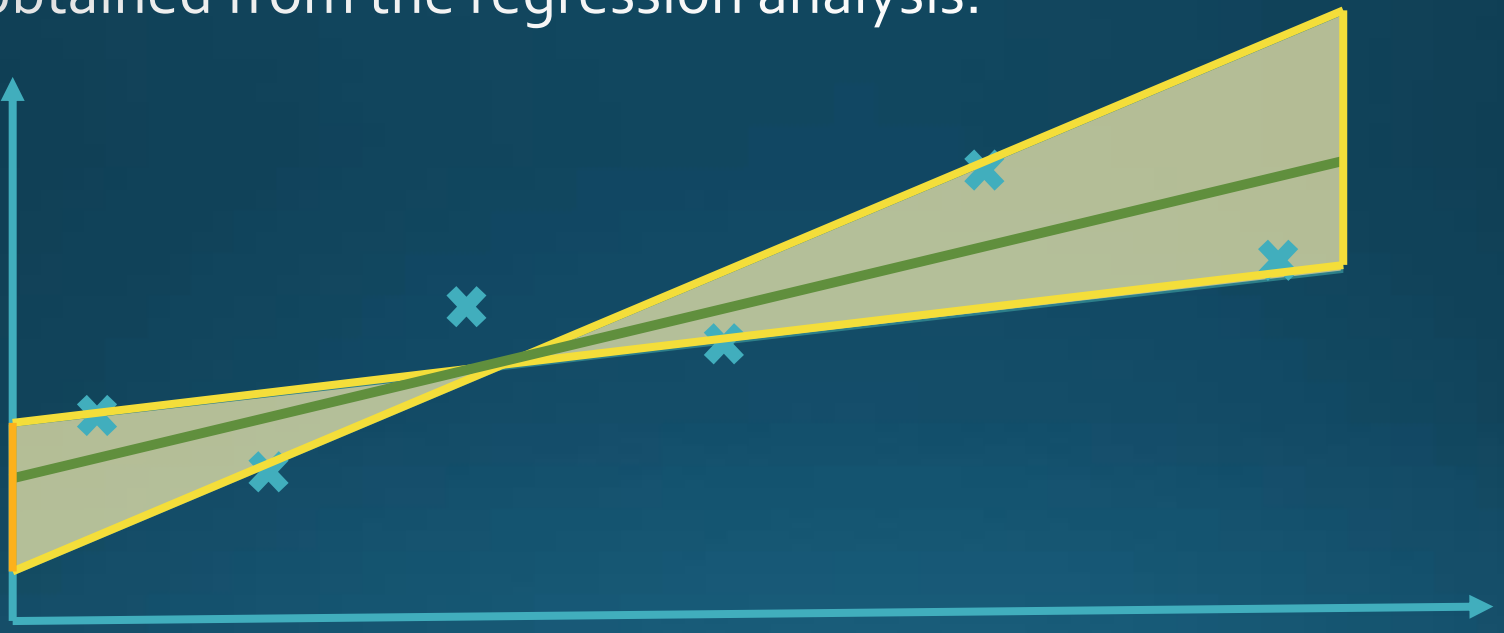
Parameter Ensemble Mean

Image Source:
Lang et al. (2016), Tellus A

$$\frac{\partial x}{\partial t} + 0.11 - 1.27x - 1.19x^2 + 5.50\frac{\partial x}{\partial s} - 0.92x\frac{\partial x}{\partial s} + 0.66x^2\frac{\partial x}{\partial s} - 0.001\frac{\partial^2 x}{\partial s^2}$$
$$-0.003x\frac{\partial^2 x}{\partial s^2} + 0.001\mathrm{x}^2\frac{\partial^2 x}{\partial s^2} = \xi_q$$

Estimated coefficient uncertainty

- Uncertainties in the calculated coefficients can be obtained from the regression analysis.



- For ensemble-based methods, this can be applied to each ensemble member's $x_m^a - \widetilde{x_m}$ to incorporate further uncertainty from the ensemble.

Estimated parameterisation uncertainty

- Uncertainties in the calculated coefficients can be obtained from the regression analysis.
- For example, for linear least-squares, the coefficient error covariance matrix is obtained from:

$$Cov(\hat{\beta}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma_y^2$$

where $\boldsymbol{X}^T = (\mathrm{g}_0\ (x_0), g_1(x_0), \dots, g_n(x_0), g_0(x_1), \dots)$ and

$$\sigma_y^2 = \frac{1}{NT-1} \sum_{i=1}^{NT} [((x^a - \tilde{x}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}})((x^a - \tilde{x}) - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T]$$

- For ensemble-based methods, this can be applied to each ensemble member's $x_m^a - \widetilde{x_m}$ to incorporate further uncertainty from the ensemble.