

## Project Statement

National Center for Atmospheric Research  
Data Assimilation Initiative

A robust ensemble filter for data assimilation:  
A forecasting system to handle outliers:

Doug Nychka (NCAR contact [nychka@ucar.edu](mailto:nychka@ucar.edu))  
Jeff Anderson  
Chris Snyder  
Kayo Ide (IPAM mentor).

### Overview

Data assimilation (DA) refers to the process of sequentially combining a dynamical model and observations to estimate the state of a system. There are many applications of DA ranging from target tracking to inferring the sources and sinks of air pollution to monitoring a manufacturing process. The most important concept in DA is that the full state (the state vector) of the system is never completely observed. The challenge is make use of partial information from sparse and noisy observations and a dynamical model for the system to estimate the full state vector. This distinction is formalized in the classical description of the Kalman filter with two equations: an observation equation that describes how the observations are related to the state of the system and a state equation that describes how the system evolves over time. This project is to formulate a method to estimate the state in the presence of occasional observations that have significant error (outliers). The introduction of outliers in the observations is not only of practical importance but also destroys methods that are based on observations with a limited error. In statistics any method that is resistant to spurious large observations is termed robust.

### Numerical Weather Prediction

This project will use as a motivating example forecasting the weather based on a sophisticated geophysical model and an network of observing stations. This process is commonly termed numerical weather prediction (NWP). Here the state of the atmosphere is updated whenever observations become available and forecasts are made by stepping the atmospheric model forward in time. The sequential nature of this process is important. Although each observation typically contributes only a small amount of local information about the state vector the accumulation of information over time and the use of a physical model to advance the state from one time point to the next allows a DA method to "learn" the state of the system. There are several features of NWP that will constrain this project. The state vector for the atmosphere and the number of available observations is huge; operational forecasts by the US weather service involve a state vector on the order of  $10^6$  elements and  $10^5$  observations. The equations for the atmosphere are often highly nonlinear and so the conventional linear theory for the Kalman filter does not readily apply. These two aspects have lead to a family of approximate methods known as ensemble filters and this project will adapt the ideas of ensemble filters to handle outliers in the observations.

### Ensemble Filters

An ensemble is sample of state vectors that are useful for estimating the state of the system and describing the uncertainty in the estimate. Typically the mean across the ensemble members is used as a good estimate of the state. An important attribute of the ensemble filter is that the spread in the ensemble members is believed to measure the accuracy of forecasting the state of the system. E.g. if there is a large variance of the ensemble members about their mean then one might infer that the there large uncertainty in estimating the state. A corresponding small variance suggests an accurate forecast. This feature will be important in any method that must decide whether a new observation is consistent with the state of the system or should be rejected as an outlier because it is outside the range reasonable observations. An alternative way to assess whether an observation is an outlier is by comparing it to neighboring observations at the

same time. (This is what is currently done for operational forecasts.) It is possible that a combination of these two criteria may be useful in any final method. Of course the challenge is to make this discrimination given that: the state of system is never known exactly, the observations will always have some error and the ensemble spread may not always be a reliable measure of the forecast accuracy.

#### Project description

Research in NWP uses a range of models from very simple and low dimensional to a complete model for the atmosphere. The strategy is to forge ideas and methods on small problems before migrating them to an operational setting. In this project the team will be charged to develop a robust ensemble filter for a simple dynamical system (Lorenz 40) and then extend the method to a more complicated system that is a simplification of large scale atmospheric flow, a quasi-geostrophic (QG) model. The project will include a visualization component where the team is challenged to illustrate their results using a series of animations and graphics that can be understood by a general mathematical audience.

The project goals are structured so that the team can develop and test algorithms in MATLAB. Although not as efficient as a compiled language, this will allow for rapid prototyping of algorithms, facilitate trouble shooting by NCAR and UCLA mentors and finally become a product that other students and groups can readily study, modify and extend.

#### Project goals:

All of the research will be done using simulation techniques where the true state of the system is known. This is in contrast to the practical NWP problem where it is more complicated to evaluate a DA procedure. Here the dynamical system will be run to produce a "true" sequence of system states. The true states are used to simulate observations and these are used in the DA procedure. Although the DA method never uses the true state directly any measure of how well the method works will of course depend on comparing the estimated state to the truth. Moreover, since there is always a random component in the results enough simulations should be done to reduce the uncertainty in the performance of a DA method and to avoid drawing spurious conclusions.

#### Specific tasks:

- 1) Implement an ensemble Kalman filter (EKF) where each observation is assimilated sequentially. Some back ground and the algorithm is given in Bengtsson, Snyder and Nychka (2003) Sections 1 and 2. (<http://www.cgd.ucar.edu/stats/pub/nychka/manuscripts/bengtsson.pdf>) The NCAR group will help with further references and background material.
  - a) The team should benchmark their code against the results in (BSN2003) for the Lorenz 40 system. (The results for the more complicated mixture and hybrid filters can be ignored.) The NCAR group will confirm the time steps, observation density and error for this testing using Lorenz 40.
  - b) Benchmark the EKF code against a QG model. Specifications for testing with this system will also be supplied by the NCAR group.

The NCAR group will give specifications on reasonable outlier scenarios and model time steps and observation locations. They will also provide MATLAB code for the two dynamical systems.

- 2) a) Adapt the EKF to handle outliers in Lorenz 40 that is devise a robust ensemble Kalman filter (REKF). The team should vary the frequency and size of the outliers and probe when the REKF breaks down. Another measure is how competitive the REKF is when there are `_no_` outliers.
  - b) Migrate the algorithm from the Lorenz 40 context to the QG model and evaluate in the same manner as 2a)

- 3) Present the results of this project to a general mathematically oriented audience. Here are some key points to consider

- a) the inherent nonlinearity and complexity of these dynamical systems used to develop the method.
  - b) the behavior of the EKF under normal circumstances and its breakdown when outliers are introduced.
  - c) the discrimination process that is part of the outlier detection in the REKF and the performance of the REKF.
  - d) what you learned from this project and what you feel is the most productive extension.
- 4) Post to the mentors MATLAB code and scripts with comments that allow another student to reproduce your basic results.