Hybrid Data Assimilation with Ensemble Covariance Matrices for Forecasting Air Pollution Data

— Background Report —

Edward Lim
eyl215@imperial.ac.uk

Supervisor: Dr. Rossella Arcucci
Course: CO541, Imperial College London

7th June, 2019

1 Overview

In recent years, global warming and deteriorating outdoor air quality has resulted in excessive energy consumption for cooling, air-conditioning and burning of fossil fuels [1]. In the wake of this, the emissions of greenhouse gases, pollutants and heat have given rise to heat islands, air pollution and unhealthy air conditions particularly in urban areas. One of the methods of combatting this is with proper urban planning such that the natural ventilation of buildings can provide a sustainable way to cool indoor environments, manage building energy consumption and control the flow of pollutants in the air.

Hence, the Managing Air for Green Inner Cities (MAGIC) project was established to develop an advanced computaitonal system that can preduct the airflow and air quality for an urban area to assist with urban planning. This goal is to optimise the natural ventilation in buildings to reduce energy usage and greenhouse gas emission using these airflow models.

This project will be focussed on the model used to predict pollutant concentration discussed in [1]. In particular, the project will explore hybrid extensions of the data assimilation (DA) methods used in [2]. This background report will cover the field of data assimilation, the specific DA methods attempted previously in the forecasting air pollution based on Fluidity models and how new state-of-the-art methods such as hybrid variational-ensemble filters can improve performance for this application. The main goal of this project is to implement these changes to the data assimilation and benchmark its' performance.

2 Background Information

As this project is heavily based on DA methods, this section will cover the motivations and mathematical derivations for variational DA methods. Sequential DA methods (ie. Kalman Filters) will not be detailed in this report. Building on this, hybrid DA techniques which involve the use of ensembles are described, along with their unique advantages and disadvantages.

2.1 Variational Data Assimilation

Data assimilation is a means of dealing with uncertainty and error with forecast models. Forecast models for inherently chaotic systems such as airflow prediction in our case are based on imperfect representations of physical processes. As a result, the model will accumulate errors until the forecast is no longer valid [3]. Data assimilation allows observations to influence the model to make a more informed prediction [4].

Most DA methods are based on probabilistic theories and Bayes Theorem [5]. In this sense, the goal of DA is to gain information about the posterior probability density function. Observations influence the posterior by matching predictions found from the model state via observation operators to real measured observations [6].

2.1.1 Full-Form Cost Function

Variational DA [7] works by finding an analysed state that minimizes a cost function to fit

a) initial conditions of the background

b) model observations to actual observations over a time window

Let the n-element vectors $\boldsymbol{x}_k$ and $\boldsymbol{x}_0^b$ represent the state of the model at a time k and the background state at time 0 respectively. The dynamics are described by the state equation:

$$\boldsymbol{x}_{k+1} = \boldsymbol{M}_{k+1}(\boldsymbol{x}_k) + \boldsymbol{\eta}_{k+1} \tag{1}$$

Where $\boldsymbol{M}_{k+1}$ is the model operator that propagates the state from time k to time $k+1$ and $\boldsymbol{\eta}_{k+1}$ is the model error at time $k+1$. Additionally, suppose there are a set of measurements, $\boldsymbol{y}_k$ which are the measurements made at a time k and their modelled counterparts, $\boldsymbol{y}_k^x$ which are related to x_k using the observation operator $\boldsymbol{H}_k$

$$\boldsymbol{y}_k^x = \boldsymbol{H}_k(\boldsymbol{x}_k) \tag{2}$$

These state equations form the basis of the 4D-Var problem (the term 4D stemming from the temporal component of the problem) which has a cost function that is the functional of the $K+1$ states in $\underline{\boldsymbol{x}} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_K\}$ and varying $\underline{\boldsymbol{x}}$, the underbar

indicating a set over a time K to minimize J [8]:

$$\begin{aligned}\boldsymbol{J}(\underline{\boldsymbol{x}}) =& \frac{1}{2}\Big(\boldsymbol{x}_0 - \boldsymbol{x}_0^b\Big)^T \boldsymbol{B}_0^{-1}\Big(\boldsymbol{x}_0 - \boldsymbol{x}_0^b\Big) \\ &+ \frac{1}{2}\sum_{k=0}^{K}\Big(\boldsymbol{H}_k(\boldsymbol{x}_k) - \boldsymbol{y}_k\Big)^T \boldsymbol{R}_k^{-1}\Big(\boldsymbol{H}_k(\boldsymbol{x}_k) - \boldsymbol{y}_k\Big) \\ &+ \frac{1}{2}\sum_{k=0}^{K-1}\Big(\boldsymbol{x}_{k+1} - \boldsymbol{M}_{k+1}(\boldsymbol{x}_k)\Big)^T \boldsymbol{Q}_{k+1}^{-1}\Big(\boldsymbol{x}_{k+1} - \boldsymbol{M}_{k+1}(\boldsymbol{x}_k)\Big)\end{aligned} \tag{3}$$

The intepretation of this cost function is:

- The first term (background term, J_b) is the difference of the analysed state $\boldsymbol{x}_0$ with the background state $\boldsymbol{x}_0^b$. This term is calculated in the L^2 regularisation with the background-error covariance matrix, $\boldsymbol{B}_0$
- The second term (observation term, J_o) is the difference of the set of observations $\boldsymbol{y}_k$ with the modelled counterparts $\boldsymbol{y}_k^x$. This term is calculated in the L^2 regularisation with the observation-error covariance matrix, $\boldsymbol{R}_k$
- The third term (model error term, J_Q) is the difference between the state at time $k+1$, $\boldsymbol{x}_{k+1}$ with the model output of the previous timestep, $\boldsymbol{M}_{k+1}(\boldsymbol{x}_k)$. This term is calculated in the L^2 regularisation with the model-error covariance matrix, $\boldsymbol{Q}_{k+1}$

The calculation of this cost function is described as an inverse problem, which can be solved if the forward problems ($\boldsymbol{M}_{k+1}$ and $\boldsymbol{H}_k$) can be solved followed by minimizing $\boldsymbol{J}(\underline{\boldsymbol{x}})$. This version of the cost function is termed *weak-constraint* [9] meaning that the model $\boldsymbol{M}_{k+1}$ is not perfect and the model-error is accounted for in the cost function with $\boldsymbol{J}_Q$.

A common strategy for approximation is to set $\boldsymbol{Q}_{k+1} = 0$ which essentially makes the states, $\underline{\boldsymbol{x}}$ follow the model exactly. This is called *strong-constraint* 4D-Var [10]. Strong-constraint variational DA only needs the state at $k = 0$ to be determined as the subsequent states follow from the model equations (Equation 1 with $\eta = 0$) directly.

The full-form cost function has a quadratic form if $\boldsymbol{M}_{k+1}$ and $\boldsymbol{H}_k$ are linear. In most real applications this is not the case and non-quadratic cost functions are difficult to minimize. One method to overcome this is by linearizing $\boldsymbol{M}_{k+1}$ and $\boldsymbol{H}_k$ about a guess or reference state, $\boldsymbol{x}_k^{(0)}$. This results in an incremental cost function

2.1.2 Incremental Cost Function

In most cases of DA, the model and observation operators are linearized about the background state, meaining that $\boldsymbol{x}_k^{(0)} = \boldsymbol{x}_k^b$ where $k = \{0, 1, 2, ...K\}$ [11]. This 4D state is found by integrating equation 1 with the assumption $\boldsymbol{\eta} = 0$ to give:

$$\boldsymbol{x}_{k+1}^b = \boldsymbol{M}_{k+1}(\boldsymbol{x}_k^b) \tag{4}$$

This can be perturbed to give the general state:

$$\boldsymbol{x}_k = \boldsymbol{x}_k^b + \delta\boldsymbol{x}_k \tag{5}$$

$$\boldsymbol{\eta}_k = \delta\boldsymbol{\eta}_k \tag{6}$$

This perturbation is propagated through the use of a tangent linear model:

$$\delta\boldsymbol{x}_{k+1} = \boldsymbol{M}'_{k+1}\delta\boldsymbol{x}_k \tag{7}$$

where $\boldsymbol{M}'_k$ is an appropriately simplified version of the Jacobian operator $\boldsymbol{M}_k$, ie. $\delta\boldsymbol{x}_k/\delta\boldsymbol{x}_{k-1}$. Using this linearization we can approximate the perturbation, $\delta\boldsymbol{x}_k$ with equations 5 and 1:

$$\delta\boldsymbol{x}_k = \boldsymbol{M}'_{0\to k}\delta\boldsymbol{x}_0 + \sum_{\tau=1}^{k}\boldsymbol{M}'_{\tau\to k}\delta\boldsymbol{\eta}_\tau \tag{8}$$

where $\boldsymbol{M}'_{0\to k}$ indicates the composition of the tangent linear model in the form, $\boldsymbol{M}'_0\boldsymbol{M}'_1\boldsymbol{M}'_2...\boldsymbol{M}'_k$. This linearization can be done similarly with the observation operator, $\boldsymbol{H}_k$ by approximating an $\boldsymbol{H}'_k = \delta\boldsymbol{y}_k^x/\delta\boldsymbol{x}_k$:

$$\boldsymbol{H}_k(\boldsymbol{x}_k) = \boldsymbol{H}_k(\boldsymbol{x}_k^{(0)}) + \boldsymbol{H}'_k(\delta\boldsymbol{x}_k) \tag{9}$$

In this case, $\boldsymbol{x}_k^{(0)} = \boldsymbol{x}_k^b$ as mentioned previously. This will result in the cost function (3) becoming:

$$\begin{aligned}\boldsymbol{J}(\delta\boldsymbol{x}_0, \delta\underline{\boldsymbol{\eta}}) =& \frac{1}{2}\Big(\delta\boldsymbol{x}_0\Big)^T\boldsymbol{B}_0^{-1}\Big(\delta\boldsymbol{x}_0\Big) \\ &+ \frac{1}{2}\sum_{k=0}^{K}\Big(\delta\boldsymbol{d}_k\Big)^T\boldsymbol{R}_k^{-1}\Big(\delta\boldsymbol{d}_k\Big) \\ &+ \frac{1}{2}\sum_{k=1}^{K}\Big(\delta\boldsymbol{\eta}_k\Big)^T\boldsymbol{Q}_k^{-1}\Big(\delta\boldsymbol{\eta}_k\Big)\end{aligned} \tag{10}$$

where $\delta\boldsymbol{x}_k$ satisfies equation 5, $\delta\boldsymbol{d}_k = \boldsymbol{d}_k - \boldsymbol{H}'_k(\boldsymbol{M}'_{0\to k}\delta\boldsymbol{x}_0 + \sum_{\tau=1}^{k}\boldsymbol{M}'_{\tau\to k}\delta\boldsymbol{\eta}_\tau)$ and $\boldsymbol{d}_k = \boldsymbol{y}_k - \boldsymbol{H}_k(\boldsymbol{x}_k^b)$. This cost function is quadratic in the inital perturbation, $\delta\boldsymbol{x}_0$ and through the use of the TLM (Equation 7) ensures that each update remains quadratic.

The derivatives of the cost function (Equation 10) can then be found directly:

$$\nabla_{\delta\boldsymbol{x}}\boldsymbol{J} = \boldsymbol{B}_0^{-1}\delta\boldsymbol{x}_0 - \sum_{k=0}^{K}\boldsymbol{M}'^T_{0\to k}\boldsymbol{H}'^T_k\boldsymbol{R}_k^{-1}\delta\boldsymbol{d}_k \tag{11}$$

$$\nabla_{\delta\boldsymbol{\eta}_\tau}\boldsymbol{J} = -\sum_{k=\tau}^{K}\boldsymbol{M}'^T_{\tau\to k}\boldsymbol{H}'^T_k\boldsymbol{R}_k^{-1}\delta\boldsymbol{d}_k + \boldsymbol{Q}_\tau^{-1}\delta\boldsymbol{\eta}_{\boldsymbol{\tau}} \tag{12}$$

Note that the gradient of the cost functions require the integration of the forward linear model from earlier times and adjoint models from future times. However, the flexibility of choice in the linear operators $\boldsymbol{M}'_k$ and $\boldsymbol{H}'_k$ to balance between ease of implementation and fidelity makes the incremental approach the most widely used in modern DA applications despite convergence issues [12].

2.1.3 Control Variable Transforms

The cost functions of both forms (Equations 3 and 10) requires explicit knowledge of $\boldsymbol{B}_0$, $\boldsymbol{R}_k$, $\boldsymbol{Q}_k$. Computation of these matrices is not feasible due to their size. Hence, a method of control variable transforms (CVTs) are introduced to represent $\boldsymbol{B}_0$ and $\boldsymbol{Q}_k$ without needing to know them explicitly. Instead of minimizing the cost function around $(\delta\boldsymbol{x}_0, \delta\underline{\boldsymbol{\eta}})$, a new set of 'control' variables are introduced as an alternative means. Many studies have investigated the appropriate selection of control variables [13] and [14].

$$\delta\boldsymbol{x}_0 = \boldsymbol{U}\delta\boldsymbol{\chi} \tag{13}$$

$$\delta\boldsymbol{\eta}_k = \boldsymbol{V}_k\delta\boldsymbol{\nu}_k \tag{14}$$

Here, $\delta\boldsymbol{\chi}$ and $\delta\boldsymbol{\nu}_k$ are the 'control' variables and $\boldsymbol{U}$ and $\boldsymbol{V}_k$ are the CVTs. The aim of this method is to have control variables that have error covariances $\boldsymbol{I}$. This can be accomplished by choosing $\boldsymbol{U}$ and $\boldsymbol{V}_k$ such that $\boldsymbol{U}^T\boldsymbol{B}_0^{-1}\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}_k^T\boldsymbol{Q}_k^{-1}\boldsymbol{V}_k = \boldsymbol{I}$. Setting these gives the cost function:

$$\begin{aligned}\boldsymbol{J}(\delta\boldsymbol{\chi}, \delta\underline{\boldsymbol{\nu}}) = &\frac{1}{2}\delta\boldsymbol{\chi}^T\boldsymbol{I}\delta\boldsymbol{\chi} \\ &+ \frac{1}{2}\sum_{k=0}^{K}\delta\boldsymbol{d}_k^T\boldsymbol{R}_k^{-1}\delta\boldsymbol{d}_k \\ &+ \frac{1}{2}\sum_{k=1}^{K}\delta\boldsymbol{\nu}_k^T\boldsymbol{I}\delta\boldsymbol{\nu}_k\end{aligned} \tag{15}$$

where $\delta\boldsymbol{d}_k = \boldsymbol{d}_k - \boldsymbol{H}_k'(\boldsymbol{M}_{0\to k}'\boldsymbol{U}\delta\boldsymbol{\chi} + \sum_{\tau=1}^{k}\boldsymbol{M}_{\tau\to k}'\boldsymbol{V}_\tau\delta\boldsymbol{\nu}_\tau)$. The gradients in the control space are found from equations 11 and 12 and the chain rule:

$$\nabla_{\delta\boldsymbol{\chi}}\boldsymbol{J} = \boldsymbol{U}^T\nabla_{\delta\boldsymbol{x}}\boldsymbol{J} \tag{16}$$

$$\nabla_{\delta\boldsymbol{\nu}_\tau}\boldsymbol{J} = \boldsymbol{V}_\tau^T\nabla_{\delta\boldsymbol{\eta}_\tau}\boldsymbol{J} \tag{17}$$

This cost function (equation 15) is easier to evaluate and better conditioned problem than equation 10 [15]. For real applications, the CVTs $\boldsymbol{U}$ and $\boldsymbol{V}_k$ are modelled by making assumptions on the relationship of the errors [13].

Furthermore, the convergence of the minimization algorithm depends on the conditioning of the Hessian of the cost function. 4D-Var is generally ill-conditioned [16] and the the Hessian (ignoring model error) is given by:

$$\boldsymbol{J}'' = \boldsymbol{B}_0^{-1} + \boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H} \tag{18}$$

One of the more well-known preconditioning techniques [17] involves the introduction of $\boldsymbol{L}$ as the CVT $\boldsymbol{U}$:

$$\delta\boldsymbol{x} = \boldsymbol{L}\delta\boldsymbol{\chi}, \quad \boldsymbol{B}_0^{-1} = \boldsymbol{L}\boldsymbol{L}^T \tag{19}$$

where $\boldsymbol{L}$ is a simple matrix. The resultant Hessian is then:

$$\boldsymbol{J}'' = \boldsymbol{I} + \boldsymbol{L}^T\boldsymbol{H}^T\boldsymbol{R}^{-1}\boldsymbol{H}\boldsymbol{L} \tag{20}$$

which is in general much more well-conditioned and exhibits strong improvement in convergence rates.

2.1.4 Limitations

This project is a direct extension of [1] and [2]. Previous work utilised Truncated Singular Value Decomposition (TSVD) as a form of preconditioning of the background covariance matrix with great success. However, the method assumes a static background covariance matrix, $\boldsymbol{B}_0$ which does not allow the flow-dependence of errors. In reality, background error statistics change in time and hence the error increases with time, resulting in divergence. Additionally, 4D-Var requires the use of tangent linear models and adjoint calculations. The adjoint model in particular is difficult to determine in chaotic systems like fluid models and weather forecasting and are based upon the idea of linearizing the models. As most real-life systems are non-linear in nature, error arises and accumulates in the predictions. This project will attempt to leverage the use of ensembles with 4D-Var DA in a hybrid method. Ensemble methods allow for flow-dependent error to be represented in DA overcoming the issue of static error covariance matrix. [6] as well as circumvent the need for an explicit tangent linear and adjoint model. This is detailed further in the following section.

2.2 Hybrid Data Assimilation

Hybrid DA refers to a system where two different DA methods run concurrently and exchange information about the errors to enrich the forecast predictions. In practice, hybrid DA combines static and predetermined prior error statistics- in our case, 4D-Var and ensemble methods. Therefore, it is more correct to say the error covariances are hybridized.

There are many flavours of hybrid DA that use a combination of different techniques, but for the applications of predicting air pollution, the 4DEnVar approach is focused on. This is due to its' ability to allow for flow-dependant error and avoidance of the tangent linear and adjoint model which is not very fitting for forecasting airflow as it is inherently a nonlinear process.

2.2.1 4DEnVar Data Assimilation

Consider a population of N background forecasts where $N \ll n$. The true background error covariance matrix, $\boldsymbol{P}_0^b$ can be approximated by this population:

$$\boldsymbol{P}_0^b \approx \boldsymbol{P}_{(N)}^b = \frac{1}{N-1}\sum_{i=1}^{N} \delta\boldsymbol{x}_{(i)}^b \delta\boldsymbol{x}_{(i)}^{b\,T} = \boldsymbol{X}^b \boldsymbol{X}^{bT} \tag{21}$$

where $\delta\boldsymbol{x}_{(i)}^b = \boldsymbol{x}_{(i)}^b - \overline{\boldsymbol{x}^b}$ where the overbar indicates the sample average. $\boldsymbol{P}_0^b$ is the true matrix, approximated in 4D-Var with $\boldsymbol{B}_0$ and by $\boldsymbol{P}_{(N)}^b$ in ensemble methods. Note that $\boldsymbol{P}_{(N)}^b$ is approximate due to sampling errors as N is finite. A linear mixture of $\boldsymbol{B}_0$ and $\boldsymbol{P}_{(N)}^b$ would allow for flow-dependant errors in a variational DA context, but the tangent linear model and adjoint model still need to be specified.

To avoid the use of the adjoint, the variational optimisation is performed in the reduced basis of the ensemble of perturbations. The idea is to find a solution in the affine subspace $\overline{\boldsymbol{x^b}}$ with $\boldsymbol{\chi}$ as the control variable and the ensemble $\boldsymbol{X}^b$ as the CVT:

$$\delta \boldsymbol{x} = \boldsymbol{X}^b \boldsymbol{\chi} \tag{22}$$

The goal is to then find the solution for the reduced-order cost function (with the model-error term excluded):

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{\chi}) = &\frac{1}{2}\boldsymbol{\chi}^T \boldsymbol{I} \boldsymbol{\chi} \\ &+ \frac{1}{2}\sum_{k=0}^{K} \delta \boldsymbol{d}_k^T \boldsymbol{R}_k^{-1} \delta \boldsymbol{d}_k \end{aligned} \tag{23}$$

where $\delta \boldsymbol{d}_k = \boldsymbol{d}_k - \boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b \boldsymbol{\chi}$ and $\boldsymbol{d_k} = \boldsymbol{y}_k - \boldsymbol{H}_k(\boldsymbol{M}_{0\to k}(\overline{\boldsymbol{x^b}}))$. $\boldsymbol{H}_k$ and $\boldsymbol{M}_{0\to k}$ indicate the non-linear observation and model operator respectively with the inverted comma versions indicating linearization about the first guess. The solution that gives the minimum of the cost function is given by:

$$\boldsymbol{\chi}^a = \Big[\boldsymbol{I} + \sum_{k=0}^{K} (\boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b)^T \boldsymbol{R}_k^{-1} (\boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b)\Big] \sum_{k=0}^{K} (\boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b)^T \boldsymbol{R}_k^{-1} \boldsymbol{d}_k \tag{24}$$

The key here is that $\boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b$ and its' transpose can be estimated with a single ensemble forecast using the equation:

$$\boldsymbol{H}'_k \boldsymbol{M}'_{0\to k} \boldsymbol{X}^b \approx \frac{1}{\epsilon} \boldsymbol{H}_k(\boldsymbol{M}_{0\to k}(\overline{\boldsymbol{x^b}} \boldsymbol{1}^T + \epsilon \boldsymbol{X}^b)) \Big(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^T}{m}\Big) \tag{25}$$

where $0 < \epsilon \ll 1$ is a scaling parameter to compute finite differences. $\epsilon = 1$ if the ensemble spread is small enough. From equation 25 we can avoid the use of the linear approximation models, $\boldsymbol{H}'_k, \boldsymbol{M}'_{0\to k}$ with the use of ensembles and maintian the non-linear representation of the actual observation and model operators, $\boldsymbol{H}_k, \boldsymbol{M}_{0\to k}$.

2.2.2 Localisation

The drawback of performing 4D-Var in the ensemble space is that it tends to be rank-deficient for high-dimensional models as $N \ll n$. This implies large sampling errors. Localisation is a method of filtering distant observations that are weakly correlated. Localisation hinges on the idea that two distant points are independent for short time scales [18]. By localising the ensemble building process, we can reduce the effect of sampling errors.

Consider a single field and let the sample covariance between points p and q at positions $\boldsymbol{r}_p$ and $\boldsymbol{r}_q$ respectively be the matrix element $[\boldsymbol{P}^b_{(N)}]_{pq}$. This is multiplied with a moderation function $c(\boldsymbol{r}_p, \boldsymbol{r}_q)$ which goes to zero when $|\boldsymbol{r}_p - \boldsymbol{r}_q| \to \infty$ [19] and unity when $\boldsymbol{r}_p = \boldsymbol{r}_q$. The localized covariance function is then:

$$c(\boldsymbol{r}_p, \boldsymbol{r}_q)\big[\boldsymbol{P}^b_{(N)}\big]_{pq} = \frac{c(\boldsymbol{r}_p, \boldsymbol{r}_q)}{N-1} \sum_{i=1}^{N} \delta \boldsymbol{x}^b_{(i)}(\boldsymbol{r}_p) \delta \boldsymbol{x}^b_{(i)}(\boldsymbol{r}_q) \tag{26}$$

where $\delta \boldsymbol{x}^b_{(i)}(\boldsymbol{r}_p)$ is the ith background perturbation at $\boldsymbol{r}_p$. A matrix form, defined with an n x n localization matrix $\boldsymbol{C}$ where $\boldsymbol{C}_{pq} = c(\boldsymbol{r}_p, \boldsymbol{r}_q)$ can be described with the eqaution:

$$\hat{\boldsymbol{P}}^b_{(N)} = \boldsymbol{C} \circ \boldsymbol{P}^b_{(N)} \tag{27}$$

where the $\circ$ operator is the element by element product. This localisation can be generalized to multivariate situations [20]. This form of localization using the element by element product can be used in 4DEnVar DA without the need of explicitly knowing the covariance matrices.

3 Progress & Plan

3.1 Progress

Majority of the work up to this date has been reading up and doing research on Data Assimilation methods as well as the state-of-the-art approaches to Data Assimilation in high dimensional problems. The main material for the background reading is *Data Assimilation: Methods, Algorithms, and Applications* [16]. Following that, focus was moved to understanding the work done in the past in the field of Data Assimilation for forecasting airflow and air pollution and the unique problems it poses.([2] and [1]) Emphasis was placed on hybrid DA methods that combine the use of variational DA, ensemble methods and the Kalman Filter as suggested by Prof. Rossella Arcucci (supervisor for this project).

3.2 Plan

A rough timeline of the project is outlined in table 1. An implementation of existing working models is introduced after background research to facilitate practical experience with DA. The focus will then shift to experimenting with hybrid DA.

Table 1: Timeline

Task	**Period**
Study fundamental data assimilation theory	Early June
Establish ensemble methods to research and implement	Early June
Implement known working DA models based on previous work	Late June
Extend/Evovle the DA system to incorporate Ensemble Methods	Late June
Build in localization for DA	Early July
Complete initial tests and benchmark results with previous work	Early July
Make adjustments and improvements to DA methods based on results	Late July
Finalize model, collect results and assess performance	Early August
Complete first draft of final report	Early August
Make edits and finalise report	Late August

References

[1] Jiyun Song, S. Fan, W. Lin, L. Mottet, H. Woodward, M. Davies Wykes, R. Arcucci, D. Xiao, J.-E. Debay, H. ApSimon, E. Aristodemou, D. Birch, M. Carpentieri, F. Fang, M. Herzog, G. R. Hunt, R. L. Jones, C. Pain, D. Pavlidis, A. G. Robins, C. A. Short, and P. F. Linden. Natural ventilation in cities: the implications of fluid mechanics. *Building Research & Information*, 46(8):809–828, 2018.

[2] Rossella Arcucci, Laetitia Mottet, Christopher Pain, and Yi-Ke Guo. Optimal reduced space for variational data assimilation. *Journal of Computational Physics*, 379:51 – 69, 2019.

[3] J. J. Tribbia and D. P. Baumhefner. Scale interactions and atmospheric predictability: An updated perspective. *Monthly Weather Review*, 132(3):703–713, 2004.

[4] C E Leith. Numerical models of weather and climate. *Plasma Physics and Controlled Fusion*, 35(8):919–927, aug 1993.

[5] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194.

[6] R. N. Bannister. A review of operational methods of variational and ensemble-variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):607–633.

[7] Olivier Talagrand and Philippe Courtier. Variational assimilation of meteorological observations with the adjoint vorticity equation. i: Theory. *Quarterly Journal of the Royal Meteorological Society*, 113(478):1311–1328.

[8] Dusanka Zupanski. A general weak constraint applicable to operational 4dvar data assimilation systems. *Monthly Weather Review*, 125(9):2274–2292, 1997.

[9] Olivier Talagrand. *4D-VAR: four-dimensional variational assimilation*, pages 3–30. 10 2014.

[10] F. Rawlins, S. P. Ballard, K. J. Bovis, A. M. Clayton, D. Li, G. W. Inverarity, A. C. Lorenc, and T. J. Payne. The met office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 133(623):347–362.

[11] Andrew C. Lorenc. Development of an operational variational assimilation scheme (gtspecial issueltdata assimilation in meteology and oceanography: Theory and practice). *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B):339–346, 1997.

[12] Yannick Trmolet. Model-error estimation in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, 133(626):1267–1280.

[13] R. N. Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. ii: Modelling the forecast error covariance statistics. *Quarterly Journal of the Royal Meteorological Society*, 134(637):1971–1996.

[14] Benjamin Mntrier and Thomas Aulign. Optimized localization and hybridization to filter ensemble-based covariances. *Monthly Weather Review*, 143(10):3931–3947, 2015.

[15] S. A. Haben, A.S. Lawless, and N.K. Nichols. Conditioning of incremental variational data assimilation, with application to the met office system. *Tellus A: Dynamic Meteorology and Oceanography*, 63(4):782–792, 2011.

[16] Mark Asch, Marc Bocquet, and Maelle Nodet. *Data assimilation: methods, algorithms, and applications*. 12 2016.

[17] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

[18] P. L. Houtekamer and Herschel L. Mitchell. A sequential ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 129(1):123–137, 2001.

[19] Gregory Gaspari and Stephen E. Cohn. Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757.

[20] Ross N. Bannister. How is the balance of a forecast ensemble affected by adaptive and nonadaptive localization schemes? *Monthly Weather Review*, 143(9):3680–3699, 2015.

A Ethics Considerations

Table 2: Ethics Checklist

	Yes	**No**
SECTION 1: HUMAN EMBRYOS/FOETUSES		
Does your project involve Human Embryonic Stem Cells?		x
Does your project involve the use of human embryos?		x
Does your project involve the use of human foetal tissues / cells?		x
SECTION 2: HUMANS		
Does your project involve human participants?		x
SECTION 3: HUMAN CELLS / TISSUES		
Does your project involve human cells or tissues? (Other than from Human Embryos/Foetuses i.e. Section 1)?		x
SECTION 4: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?		x

Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		x
Does it involve processing of genetic information?		x
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.		x
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		x
SECTION 5: ANIMALS		
Does your project involve animals?		x
SECTION 6: DEVELOPING COUNTRIES		
Does your project involve developing countries?		x
If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		x
Could the situation in the country put the individuals taking part in the project at risk?		x
SECTION 7: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		x
Does your project deal with endangered fauna and/or flora /protected areas?		
Does your project involve the use of elements that may cause harm to humans, including project staff?		x
Does your project involve other harmful materials or equipment, e.g. high-powered laser systems?		x
SECTION 8: DUAL USE		
Does your project have the potential for military applications?		x
Does your project have an exclusive civilian application focus?		x
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		x
Does your project affect current standards in military ethics e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		x
SECTION 9: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		x
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		x
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied??		x

Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		x
SECTION 10: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?		x
Will your project use or produce goods or information for which there are data protection, or other legal implications?		x
SECTION 11: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?		x