The Application of Local E-Dimension/E-Vector Analysis to Near-Operational Ensembles of the Global Data Assimilation System

Adam Kubaryk

Department of Atmospheric and Oceanic Science, University of Maryland College Park, Maryland

Advisor Dr. Kayo Ide

A scholarly paper in partial fulfillment of the requirements for the degree of Masters of Science

26 April 2016

Acknowledgements

I would like to express truly endless thanks and appreciation to the many people who have assisted me in my journey at the University of Maryland College Park. Most principally I would like to thank Dr. Kayo Ide, my academic and research advisor, whose patience, support, advice, guidance, and direction formed my foundation at the University. I received invaluable support and advice from a tireless Dr. Daryl T. Kleist, who provided too many replies to questions that nobody else could answer; his deep breadth of knowledge and network of contacts underlied much of the research here. I thank, broadly, all of the faculty and staff of the Department of Atmospheric and Oceanic Science at the University: I have learned much in my time here both as a student and as a professional that has made me who I am today, and it will continue define me as I transition to the working world. I would like to thank Dr. Tomoko Matsuo of the University of Colorado Boulder, the PI of the project under which I was funded, for contributing much in the way of support and ideas for this project. I would also like to thank my family and friends, girlfriend and cat, without whom none of this would have been possible.

Table of Contents

Abstract

Ensemble forecasting has become an integral tool for numerical weather prediction (NWP) centers. Much emphasis is placed on the structure and behavior of the ensemble covariance matrix through the forecast and assimilation cycle, as the ensemble estimate of covariance is integral to the quality and effectiveness of a probabilistic forecast. This study seeks to investigate the local dimensionality of the Global Forecast System (GFS) model by performing principal component analysis on near-operational ensemble perturbation matrices to identify the both the local dimensionality of the model and the structure of its unstable manifolds by inspecting its E-Vectors. The unstable manifolds of the GFS represent the dominant modes of error growth, and are critical to near-term forecasts. We seek to expand on prior seminal works to adapt Local E-Dimension/E-Vector (LEDEV) Analysis, placing particular emphasis on the structure and consistency of E-Vectors in four dimensions, as well as through the data assimilation process, to identify spatiotemporal properties inherent to GFS ensemble perturbations.

List of Figures

List of Tables

List of Abbreviations

EnSRF	Ensemble Square Root Filter
EnKF	Ensemble Kalman Filter
GDAS	Global Data Assimilation System
GFS	Global Forecast System
LEDEV	Local E-Dimension/E-Vector
LETKF	Local Ensemble Transform Kalman Filter
LND	Level of Non-Divergence
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
SAIR	Space Atmosphere Interaction Region

1. Introduction

Central to NWP is the use of dynamical models of the nonlinear Earth systems to diagnose, understand, and predict the state and evolution of atmospheric flow. Over the last two decades, ensemble forecasting has been increasingly utilized in NWP centers both as a tool to provide probabilistic forecasts of meteorological events, and as a method to reduce computational cost of data assimilation relative to more expensive approaches like 4DVar while retaining the information provided by so-called errors of the day.

The mathematical concept underlying ensemble forecasting is based in Monte Carlo analysis: repetitive trials of non-deterministic forecasts to estimate the likelihood of given forecast conditions occurring. Consider *k* slightly perturbed initial conditions $\mathbf{x}_i$ for $1 < i < k$. Each ensemble member is initially a function of the true state $\mathbf{x}_t$ and some error $\boldsymbol{\varepsilon}_i$. The true state is intrinsically unknown, and so too is the error. As each member is integrated forward in time, the initial error grows exponentially, and comes to represent both the integrated initial representativeness error as well as introduced model error.

Assuming there is no introduced bias, the best estimate of the true state is the ensemble mean $\bar{\mathbf{x}}$, and the error can be estimated by the ensemble perturbation vector $\mathbf{x}_i - \bar{\mathbf{x}}$. Using each perturbation vector as the *k*-columns of a matrix leads to the ensemble perturbation matrix, **X**. If the initial perturbations are structured in such a way that they accurately represent the space spanned by the uncertainty intrinsic to the initial conditions, this perturbation matrix **X** can be used to construct a covariance matrix that is intended to span the uncertainty present from the model forecast, and can be used in a data assimilation system to optimally update the background forecasts with observational data. These methods will be covered briefly in a later chapter.

The methodology as described relies on several assumptions that are often not true in a real-world system. Even ignoring those issues, there are still several meaningful challenges to successfully implementing an ensemble forecasting system. Most pressingly, modern NWP models running at operational deterministic resolutions have $O(10^9)$ degrees of freedom. While ensemble forecasting is often performed at a lower resolution to control the computational cost, this would seemingly pose an insurmountable difficulty in achieving an ensemble space that accurately represents the uncertainty in initial conditions.

Thus it is necessary to estimate the inherent complexity of the system in the interest of determining what constitutes an appropriate number of ensemble members for spanning the space of uncertainty in an ensemble forecast. It is common practice with all nonlinear dynamical systems to determine their complexity via an estimate of its dimensionality. A traditional measure is the Lyapunov or Kaplan-Yourke (1979) dimension. Numerical calculation of the Lyapunov dimension relies upon estimating the global exponential divergence of chaotic trajectories over the entire model field. As a global measure it represents the entire system rather than spatiotemporally localized regions, and it becomes exponentially more difficult to calculate with additional model variables. It is thus computationally prohibitive to compute and essentially meaningless for the purposes of diagnosing a system with a high order of degrees of freedom and spatiotemporally localized chaotic dynamics.

Thus there existed substantial motivation to define a measure of dimensionality specific to the needs of a NWP model. Patil et al. (2001) defined the Bred Vector (BV) Dimension for this purpose. It is a continuous statistical measure of the singular values resulting from a principal component analysis of the ensemble perturbation matrix. Critically, by localizing the region of analysis and timestep at which the perturbations are defined, they succeeded in defining a truly spatiotemporally localized measure of dimensionality.

This measure will be defined both generally -- in an intuitive manner -- and rigorously -- in a mathematical manner -- in Chapter 2. Additionally, the chapter will review of several papers that have used this measure in application to NWP. Then, the direction of research for this paper and the experimental design will be described in Chapter 3, and the results will be presented in Chapter 4. Finally, a summary of the results in the context of prior works and the intended direction of future research be in the concluding Chapter 5. An Appendix is provided with additional information about the analysis package that was developed for this research.

Here we follow and build upon the naming conventions of the Oczkowski et al. (2005), referring to the singular value statistic as the ensemble dimension, or E-Dimension. This naming convention is continued for the left singular vectors -- E-Vectors.

2. Local E-Dimension and E-Vector Analysis

2.1. Principal Component Analysis

Principal component analysis (PCA) is commonly used in the Earth Science community to determine empirical relationships between variables across space or time at all different scales. Examples include the study of the El Niño Southern Oscillation (Westra et al. 2010), urban analyses of atmospheric chemistry (Costabile et al. 2009) and climatological studies (Benzi et al. 1996). For the application presented here -- PCA of ensemble perturbations -- similar concepts are being utilized, and we aim to draw a parallel between these uses of PCA. First, we seek to understand the ensemble perturbation matrix and its properties. Then, we will explore the matrices of the principal component analysis and what they represent. Finally, we will use the definition of Patil et al. (2001) as the basis for E-Dimension analysis.

The ensemble perturbation matrix $\mathbf{X} \in R^{N \times k}$ -- with N being the entire model variable field and *k* ensemble members -- constructed as described Section I, has impractical dimension and geographical footprint to extract relevant information about synoptic-scale meteorological features. Instead, an appropriate subset of the perturbation matrix, $\mathbf{B} \in R^{L \times k}$, is used -- with L being the number of chosen variables -- containing all of the chosen variables local to and surrounding the analysis gridpoint; such a choice is arbitrary, but if the general purpose of the analysis is to diagnose the dimensionality of local chaotic dynamics in the system, it is suggested that the analysis grid size is appropriate to capture relevant synoptic scale features. This will be discussed in more detail in Chapters 3 and 4.

PCA is a decomposition procedure that transforms an arbitrarily sized set of *n*-dimensional data into a set of at-most *n* linearly uncorrelated components, from which each observation can be reconstructed as a linear combination of the orthogonal components. By definition, the PCA decomposition is empirical in nature, with the leading mode necessarily corresponding to the vector which represents the plurality of variance present in the data set. This procedure is repeated recursively such that each additional mode represents the largest remaining amount of variance in the transformed data set until all of the variance is accounted for. A computationally-efficient method of obtaining the principal components is by singular value decomposition.

Thus consider a factorization of the matrix $\mathbf{B} = \mathbf{U\Sigma V}^\mathbf{T}$ where $\mathbf{U}$ and $\mathbf{V}^\mathbf{T}$ are orthonormal matrices, and $\mathbf{\Sigma}$ is a rectangular diagonal matrix whose entries are referred to as the singular values of $\mathbf{B}$. $\mathbf{U}$ is referred to as the left singular vectors of $\mathbf{B}$, and $\mathbf{V}^\mathbf{T}$ the right singular vectors of $\mathbf{B}$. The covariance matrix $\mathbf{C} = \mathbf{BB}^\mathbf{T}/(L-1)$ has non-negative values on its diagonal representing each variable's variance, and its off-diagonal values are symmetric about the diagonal, containing the ensemble covariance values. As a symmetric positive semi-definite matrix, its eigenvalues are non-negative. Inspecting a scaled covariance matrix:

$$\mathbf{BB}^\mathbf{T} = \mathbf{U\Sigma V}^\mathbf{T}(\mathbf{U\Sigma V}^\mathbf{T})^\mathbf{T}\ , \quad \text{(Eqn 1)}$$

$$\mathbf{BB}^\mathbf{T} = \mathbf{U\Sigma V}^\mathbf{T}\mathbf{V\Sigma U}^\mathbf{T}\ , \quad \text{(Eqn 2)}$$

$$\mathbf{BB}^\mathbf{T} = \mathbf{U\Sigma}^2\mathbf{U}^\mathbf{T}\ . \quad \text{(Eqn 3)}$$

Similarly, $\mathbf{B}^\mathbf{T}\mathbf{B} = \mathbf{V\Sigma}^2\mathbf{V}^\mathbf{T}$. We find that $\mathbf{U}$ is a set of orthonormal eigenvectors for $\mathbf{BB}^\mathbf{T}$ which spans the columns of $\mathbf{B}$, and $\mathbf{V}^\mathbf{T}$ is a set of orthonormal eigenvectors for $\mathbf{B}^\mathbf{T}\mathbf{B}$ which spans the rows of **B**. Additionally, the singular values of **B** correspond to the square root of the eigenvalues of $\mathbf{BB}^{\mathbf{T}}$, which have been established as non-negative; thus the singular values of **B** are real.

If the structure of **Σ** is such that $\mathbf{\Sigma}_{ii}$ >= $\mathbf{\Sigma}_{jj}$ for $i < j$, the left singular vectors which orthonormally span the column space of **B** -- or, specifically, the physical space spanned by the ensemble perturbations -- are ordered in decreasing representation in the ensemble variance structure, in line with the definition of PCA given above. The left singular vectors are weighted by the singular values, and together they represent the amount of variance contained in each direction indicated by the columns of **U**. This variance structure can be considered a representation of the physical uncertainty in a forecast: the left singular vectors alone represent the major orthogonal modes of ensemble variance, and are thus referred here to E-Vectors; additionally, the relative magnitude of the singular values is a proxy for how important the individual E-Vectors are in describing the variance structure. For example, a random variance structure in which each model variable is independently and equally varying in the ensemble would have equal singular values; likewise, a variance structure that has particularly dominant modes of growth would have very large singular values corresponding to those modes, while

2.2. E-Dimension

This structure of the singular values is what lead Patil et al. (2001) to define their ensemble dimension as follows:

$$E(\sigma_1, \sigma_2, ..., \sigma_k) = \frac{(\sum_{i=1}^{k} \sigma_i)^2}{\sum_{i=1}^{k} \sigma_i^2} \quad . \quad \text{(Eqn 4)}$$

Note that the E-Dimension is a continuous function strictly of the singular values. In the extreme case that all of the singular values only vary slightly, signifying small, random error growth in the system, the E-Dimension will approach *k*, and for cases of dominant error growth in the system along significant spatial vectors, the E-Dimension will be closer to its minimal value of 1. Thus the E-Dimension is an intuitive estimate of the true local dimensionality of chaotic dynamics in the system, and the E-Vectors represent the physical manifestation of that uncertainty. Together, they represent an analysis technique under one umbrella: Local E-Dimension and E-Vector (LEDEV) analysis, which empirically investigates the structure of variance -- and thus uncertainty and error growth -- for an ensemble of forecasts, and implicitly estimates the localized dimensionality of the dynamical system. Much like statistical analyses of surface air temperatures during ENSO cycles, the PCA is performed here with the intent of investigating how the model variables covary, and the physical and numerical structure of the variance defines the LEDEV analysis.

For example, we might consider performing LEDEV analysis for a NWP model in a mid-latitudinal region ahead of trough axis as a frontal passage occurs. In near-term forecasts, uncertainty grows in accordance with local instabilities in the atmospheric flows. In this case, we intuitively expect that dominant modes of error growth will present along the trough axis as it passes through, causing an increase in modally-focused uncertainty in that region. This would lead to transient local low-dimensionality in that region, and correspondingly the E-Dimension would drop. Visually inspecting a few of the leading E-Vectors might provide physical intuition for the structure of error growth in the system, and explicitly identify the main orthogonal modes of growth in the system. Further causes of local low-dimensionality and the properties of their persistence will be discussed in subchapter 2.4.

2.3. Projection Matrix

Finally, for the sake of completeness, we consider a partial representation of **B** in the matrix **T** = **UΣ**. **T** represents the weighted E-Vectors of **B**, and it entirely spans the ensemble space. The multiplication by $\mathbf{V}^\mathbf{T}$ to recover **B** can thus be considered a projection of the weighted E-Vectors onto the individual ensemble members. We thus call $\mathbf{V}^\mathbf{T}$ the projection matrix, although its relevance to LEDEV analysis is not explored in this study.

2.4. Extended Literature Review

It is important to understand the historical context of the study of E-Dimension, as it has entirely shaped the goals of this study. The roots of E-Dimension are defined by the works by Toth and Kalnay (1993, 1997) on ensemble forecasting, and -- quite specifically -- bred vectors. Bred vectors are somewhat analogous to E-Vectors, as they are constructed by running the model simultaneously with infinitesimal errors introduced. At regular intervals, the error vectors are rescaled and allowed to continue to grow, with the assumption that they are aligned with the fastest growing modes of error growth, thus representing the largest modes of uncertainty in the model. Keller et al. (2010) noted independently that “bred vectors … mainly point in the phase space direction of the leading Lyapunov vector and therefore favor one direction of growing errors,” and thus devised a method that closely mirrors the one derived above, using orthogonalization to obtain further vectors different from the leading one by applying a singular value decomposition.

Beginning with Patil et al. (2001), the focus of E-Dimension studies has appropriately been on data assimilation, as the E-Dimension analysis expresses such useful information about the structure of a perturbation matrix and thus covariance matrix. As a defining point, the authors cite Bishop and Toth (1999), who expressed that one should seek to move the initial state “closer to the observations along the direction of the unstable sub-space since that is where the true state most likely lies.“ In a practical application of E-Dimension, wherein one assumes that a low dimensionality implies a highly unstable subspace of the model, adaptively increasing ensemble spread in these regions as to better adhere to observational data, Corazza et al. (2002) found that in simple model experiments, improvements of up to 40% could be registered with this approach.

Patil et al. (2001) suffered primarily from a severely limited ensemble size due to computational restraints, only working with five ensemble members initially. The authors conjectured about the impact that increased ensemble size would have on the E-Dimension analysis, suggesting that there might be a fairly low upper bound to the E-Dimension of six. This was explored by Oczkowski et al. (2005), as they varied the ensemble size in increments between five members and 150 members. Indeed, the upper-bound expressed itself, as some 30 ensemble members proved sufficient to reach the upper-bound of the E-Dimension field minimum, but only at around 130 ensemble members did the maximum E-Dimension settle into the mid-30s. This suggests that the current operational ensemble of 80 members should be sufficient to explore the highly unstable space in the model if the ensemble members are ideally constructed, but the ensemble is insufficient to explore high-dimensionality regions. Indeed, this hints at a reason for lagging verification scores in the tropics, where E-Dimension is typically higher than mid-latitudinal and polar regions.

Oczkowski et al. (2005) also explored the significance of changes in domain size, finding that increased domain size tended to increase E-Dimension, and reduced domain size tended to decrease E-Dimension, but the structure of the E-Dimension field remainged largely consistent. There is both a mathematical and physical explanation for such a phenomenon. Mathematically, the introduction or removal of additional degrees of freedom into the analysis will tend to increase or decrease its dimensionality respectively. Physically, including variables on the fringe of a meteorological feature -- that may no longer represent the dynamic instability of that feature -- will induce perceived randomness in the analysis: the variables on the edge of the domain will appear uncorrelated, thus increasing the apparent E-Dimension. In seeking to analyze transient local low-dimensionality features, it is critical that one consider their footprint; thus, the authors chose an analysis grid size corresponding to the synoptic scale -- roughly 1000km x 1000km in the mid-latitudes.

With a case study analysis, the authors identified several different atmospheric processes that lead to transient low local dimensionality, including baroclinic instabilities, interacting baroclinic and barotropic instabilities, ageostrophic geopotential flux divergence, upper-tropospheric wave packet propagation, and multi-scale wave interactions in the atmospheric flow. Ultimately, they concluded that a feature of eddy diffusivity was responsible in many cases for the reduction in E-Dimension, implying that small-scale chaotic processes are responsible for changing local dimensionality at synoptic scales. However, the relative persistence of low dimensionality on the order of 1-3 days implied that in spite of highly non-linear dynamics, there is much to be gained by improving the initial conditions in scenarios of low dimensionality.

Szunyogh et al. (2005) explores the relationship between explained variance and E-Dimension as a basis to study the dynamical properties of an ensemble as they respond to, among other experiments, the differing dynamics present in the tropics compared to mid-latitudes. The authors suggest the overwhelming complexity of local sub-grid scale dynamics causes the ensemble to insufficiently capture the relevant true ensemble spread, weakening the estimate of the true state in those regions. However, they suggest this is an inherent limitation of the Ensemble Kalman Filter, and that care should be used when differentiating between model error and error introduced by the data assimilation process.

Other work has been done on ensemble sensitivity analysis using derivatives of the principal component analysis process. For example, Enomoto et al. (2015) define an ensemble singular vector (ESV) sensitivity metric, which is rooted in an eigenvector calculation. With this, they are able to successfully identify regions of high sensitivity to initial conditions, suggesting these might be regions of low dimensionality: high uncertainty in divergent directions. The same lead author contributed to Yang et al. (2015), which attempted to utilize these ESVs as additive noise in ensemble inflation, a topic which will be discussed in a later chapter. The concept is that the leading ESVs represent the fastest growing modes of error in the system, and thus provide useful information to the ensemble in order to better align its members along the unstable manifolds of the model. The authors show that this can be used to substantially reduce the spin-up time of an ensemble forecasting system, although it has other potential interesting applications. We suggest -- and will highlight during the results -- that a similar approach using the local E-Vectors could be applied, although this is primarily a topic for future research.

3. Experimentation Design

3.1 Goals

As has been shown, the basis of LEDEV analysis is concurrently a statistical and physical interpretation of a *k*-member set of atmospheric ensembles, spatiotemporally localized to a given timestep and the geographic region around the point of analysis. While the E-Dimension and E-Vectors are mathematically linked by the orthogonal decomposition process, they represent different styles of analysis. E-Dimension is a statistic dependent on the physical structure of the ensemble perturbations which span the ensemble space in the given region of analysis. The leading orthogonal E-Vectors represent a largely physical manifestation of the ensemble perturbations. On the basis of previous work which has primarily focused on E-Dimension analysis, this paper ultimately seeks to utilize both methods of analysis to explore two independent ideas on an operational reduced-resolution version of the National Centers for Environmental Prediction's (NCEP) Global Data Assimilation System (GDAS). Table 1 lists the experiments which will be described below in detail.

3.1.1 Data Assimilation Comparison

We will apply the LEDEV analysis to two common flavors of the Ensemble Kalman Filter (EnKF): the Ensemble Square Root Filter (EnSRF) and the Local Ensemble Transform Kalman Filter (LETKF). While a theoretical comparison of the two might suggest analytical equivalency (Anderson and Collins 2006), real-world systems often do not behave in such an agreeable manner. We seek to investigate any notable difference between the two schemes by comparing their handling of the ensemble perturbation matrix through the data assimilation process.

3.1.2 Data Assimilation Assessment

After identifying a data assimilation system of choice for these experiments, the process will be broken into two components, reflecting the operational configuration of the Gridpoint Statistical Interpolation (GSI) system. The first component is the ensemble analysis update strictly a product of the Ensemble Kalman Filter algorithm, incorporating the ensemble background forecasts with observational data. The second component is the inflation process. While inflation is used to maintain sufficient spread to prevent filter divergence and the collapse of the ensemble forecasting system (Anderson 2006), we sought to separate its effects from the actual ensemble update step as to assess the ensemble data assimilation algorithm independently. The specifics of these two components will be discussed in greater detail in the following subchapter.

3.1.3 Consistency: Vertical, Horizontal, Temporal

As a second point of emphasis, we wish to study the behavior of E-Vectors in a spatiotemporal manner. While one of the key features of LEDEV analysis is the spatially localized and temporally static nature of the analysis, it is important that the results presented be fairly robust; that is, in four dimensions, the E-Vectors should present the level of persistence suggested by previous studies: vertically -- within the same column of the model; laterally -- within the same model level; and temporally -- as the model fields evolve from time step to time step.

Vertical consistency is specifically a product of the single-level analysis that will be universally applied in these experiments. While the Oczkowski paper included the entire column of model variables in his gridpoint analysis, we note that the structure of meteorological features is very commonly tilted with height, and the behavior of the model necessarily changes from model level to model level. This issue has been exacerbated in recent years as the top of the model has been pushed into the upper stratosphere. Thus we do not seek uniform vertical consistency, but vertical consistency appropriate to synoptic dynamics: for example, in the case of a passing front we would expect to see broad consistency in the boundary layer, and vertically consistent regions above that indicate a common tilt of the system with height.

In the same way that slight changes to the domain size should not substantially affect the analysis, the overlapping segments of adjacent horizontal gridpoints that largely share the same analysis domain should remain largely consistent in the overlapped regions. Often in our results, the single level that is presented for non-vertical analysis will be correlative to the 850mb level. This was chosen for dynamical reasons, largely seeking a dynamically active, highly energetic region of the atmosphere that is commonly above the boundary layer in the mid-latitudes. Notably, regions of strong thermal gradient at the 850mb level signify regions of convergence in the atmosphere. Other candidates, like the level of non-divergence or the tropopause, will be shown, but not emphasized.

Finally, we hope that the meteorological features that define transient local low dimensionality have temporally consistency within a reasonable level of tolerance: it should not be the case that from one timestep to the next that the system has evolved so drastically that the current LEDEV analysis is entirely incoherent from the prior LEDEV analysis. This follows specifically from the results of Oczkowski et al. (2005), which identified persistence of local low dimensionality for 24-72 hours; we hope to identify similar consistency in the E-Vectors themselves, suggesting the modes of growing error have a physical basis and are not coincidental mathematical artifacts.

Table 1: Experiment List and Description

Experiment	EnKF Algorithm	Dates	Analysis Tool
Data Assimilation Comparison	LETKF&EnSRF	2013/12/09 18Z - 2013/12/10 00Z	E-Vector Correlation
Data Assimilation Assessment	EnSRF	2013/12/09 18Z - 2013/12/10 00Z	E-Vector Correlation
Temporal Consistency	EnSRF	2013/12/09 18Z - 2013/12/12 00Z	E-Vector Inspection
Vertical Consistency	EnSRF	2013/12/11 06Z	Vertical Cross-Section Analysis
Horizontal Consistency	EnSRF	2013/12/11 06Z	Overlapping E-Vector Comparison

3.2. Analysis Design

In contrast to the Oczkowski et al. (2005) use of an energy norm: $\| \cdot \|_E^2 = U^2 + V^2 + \frac{c_p}{T_r} T^2$, here generalized to a single-level analysis, we opt to treat the model variables individually. This is borne out of a desire to not introduce undue bias into the LEDEV analysis system by transforming the variables. Additionally, in line with the Oczkowski et al. (2005), which used a 5x5 grid at T62, we utilize an analogous 19x19 grid at T254, seeking to capture synoptic without introducing potentially spurious correlations at the edge of the domain.

These experiments will largely be done with both qualitative and quantitative approaches. Particularly in the study of temporal consistency, it is most useful to visually inspect the E-Vectors as they evolve after the 6-hour analysis timestep. Largely, though, an absolute correlation analysis will be applied to sets of E-Vectors:

$$|corr(v_1, v_2)| = |\tfrac{v_1 \cdot v_2}{||v_1|| \cdot ||v_2||}| . \quad \text{(Eqn 5)}$$

Such analysis allows us to quantitatively compare an entire set of E-Vectors with a qualitative eye, and it will often form the crux of comparative LEDEV analysis with the aid of the individually plotted E-Vectors to illustrate certain features.

3.3. System Configuration

Critical to any experiment is the choice of model and its configuration. All of the experiments presented here were performed on a near-operational 80 member ensemble of a hybrid 3D EnVar implementation of the Global Data Assimilation System (GDAS), a coupled GFS/GSI system. Hybrid data assimilation is a now-common approach that combines facets of variational, deterministic data assimilation with a supporting ensemble (Wang et al. 2008). When this research was in its infancy, the operational GDAS utilized a Eulerian grid at T574 for the deterministic run, and T254 for the ensemble, and we use the same resolutions here. The GFS has 64 hybrid sigma levels stretching from the surface to roughly 0.5mb.

The hybrid system operates as follows: the ensemble members and deterministic forecasts are integrated from the time of analysis out to nine hours. The deterministic run at T+6h is updated incrementally from the deterministic 3DVar solver with some of its covariance matrix coming from the static 3DVar matrix, and some coming from the ensemble covariance estimate with localization applied. The ensemble at T+6h is updated according to the EnKF algorithm and then recentered around the deterministic analysis. In these experiments we use equal weighting of the static and dynamic covariance matrices, and make use of two different EnKF algorithms: the EnSRF and the LETKF.

The operational EnSRF was implemented using the algorithm defined in Hamill and Whitaker (2002). By its nature, the EnSRF must serially assimilate observations instead of ingesting them in en masse as traditional EnKF algorithms do. With little computational cost, it is possible to estimate the impact that any given observation will have on the resulting analysis covariance matrix: one operational solution to improve computational cost of the EnSRF is to ignore all observations with a projected impact of less than 1% on the analysis matrix. Generally speaking this winds up being 10% of all observations. Critically, however, the LETKF does not have this feature, and therefore we have opted to disable it for all EnSRF runs.

The non-operational LETKF algorithm was implemented according to Hunt et al. (2007), and is configured identically to the EnSRF algorithm where appropriate.

Spread inflation is a necessary part of ensemble forecasting: without any treatment, ensemble spread tends to collapse to values that all but reject observational data during the EnKF update, and the ensemble forecast quickly diverges from the true nature state. The inflation algorithm is chosen to be a traditional relaxation to prior spread (RTPS). This is a form of multiplicative inflation in which -- on a variable-by-variable basis at each gridpoint -- the spread of the analysis ensemble is artificially inflated a certain percentage to the background ensemble spread. No constraints were applied to this process, and the multiplicative constant was chosen to be 0.8, consistent with the previously operational configuration. Other choices for ensemble spread inflation include stochastic methods such as additive inflation -- in which stochastic noise from the 3DVar static covariance is added to the ensemble -- and stochastic physics parameterizations -- in which sub-grid-scale processes are varied within their ranges of uncertainty to produce additional artificial spread in ensemble forecasts. SPPT was included in its operational configuration, but additive inflation was not explored in these experiments.

3.4. Initial Conditions

The initial conditions for December 1 2013 00Z were taken from an archive of then-operational GDAS analyses, and integrated forward in time seven days in order to alleviate any model shock introduced by differences between the 2013 GFS and the version used in this study. Then, separate runs of the EnSRF and LETKF were done for an additional seven days. The times of analysis chosen for each experiment correspond to the beginning, middle, and end of the passage of a deep trough in its mature phase, in the process of weakening due to inclusion.

4. Results

4.1. Case introduction

We will first justify the choice of case study for this analysis and introduce the standard set of figures that will be used for E-Vector inspection. Figure 1 shows a global view of E-Dimension, as well as the ensemble mean and ensemble spread of one of the three variables used in this analysis. As has been noted, there is a definite latitudinal structure evident, with the tropics having substantially higher E-Dimension than mid-latitudinal and polar regions. This follows from the discussion of highly chaotic sub-grid scale processes present in the tropics that lead to increased dimensionality. We call attention to the correlation between regions of high ensemble spread and regions of low dimensionality: e.g. the region in the Central Indian Ocean, the area off the coast of New Zealand, and much of the area around Japan. We note that there is a depression in the E-Dimension field in the Gulf of Alaska despite a lack of increased ensemble spread. This is a region of highly active cyclogenesis in the Northern Hemispheric winter.

Ultimately, though, we opted to investigate the region in the North Atlantic where a pocket of high ensemble spread exists with a corresponding depression in E-Dimension. Together with the structure of the mean field -- particularly looking at the trough evident in the temperature field -- we identify an extratropical cyclone. Additional analysis of the synoptic setup in this timeframe indicates that this is a mature system, having delivered a mild snowstorm to the Northeast US before progressing into the North Atlantic.

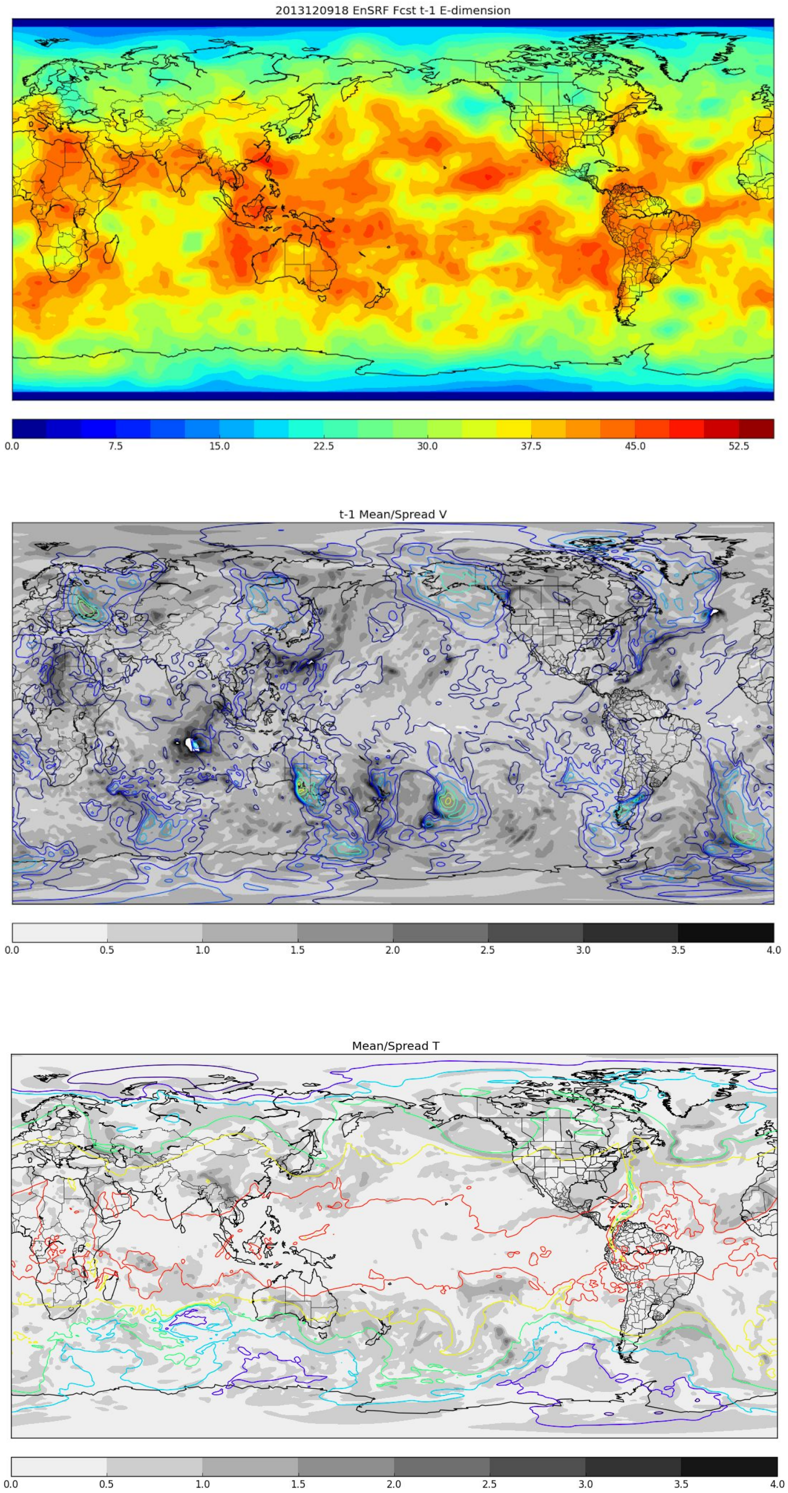


Figure 1: *Top*: A global map of single-level (ML15) E-Dimension for a background forecast valid at 2013120918 utilizing the EnSRF dataset using shaded contours with 2.5 interval. *Middle*: A global map of the ensemble spread in $m^2 s^{-2}$ (shaded contours) and ensemble mean zonal wind (colored contours, 5m s^{-1} intervals). *Bottom:* Same as middle, for temperature (spread in shaded contours, mean in colored contours, 10K intervals).

Performing LEDEV analysis in this region around 37.9N, 31.9W, we extract its corresponding E-Vectors, shown in Figures 2-4. We note that the E-Vectors for temperature and wind variables tend to be relatively independent; while there are notable structural similarities -- particularly with the negative tilt in the E-Vector fields -- ultimately there is no immediate striking correlation. The zonal and meridional E-Vector fields share mild structural similarities, demonstrating some geographically-offset consistency in regions with swirling wind fields. Ultimately, though, direct cross-variable correlation in the ensemble perturbation matrix exists, but is not a dominant feature. We will choose to target the temperature E-Vector field at this model level, as the temperature E-Vectors express dominant representation, while the leading wind E-Vectors can be subdued or even flat, but it should be emphasized that choice of variable does not substantially alter the analysis presented here.

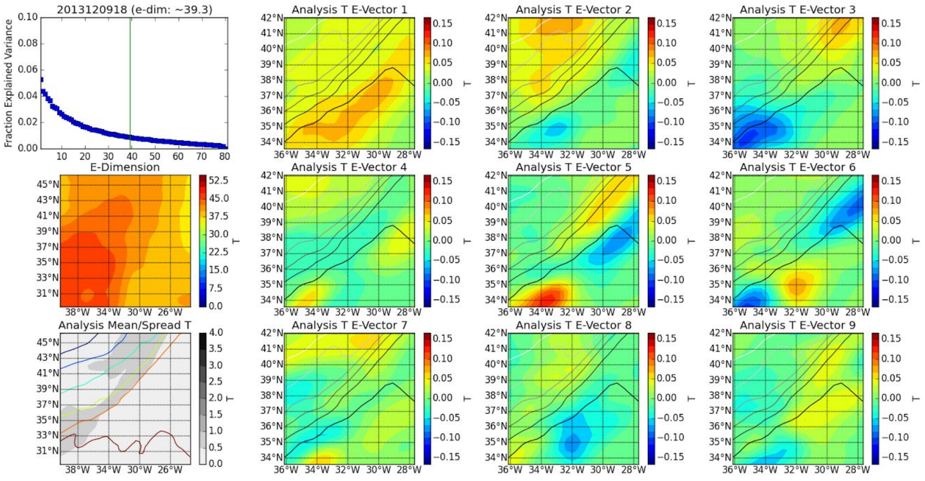
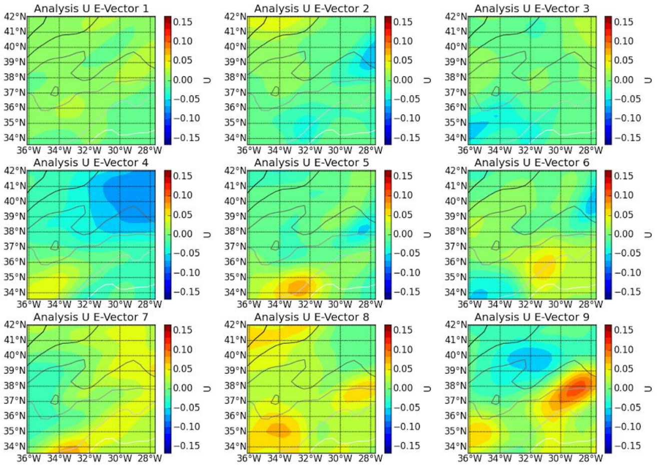


Figure 2: *Top-Left*: A plot of the singular values of the localized ensemble perturbation matrix, with a vertical line denoting the E-Dimension, at the defined region of interest in the North Atlantic, valid at 2013120918 utilizing the EnSRF dataset. *Middle-Left*: A plot of E-Dimension in the region immediately surrounding the defined region of interest, zoomed out so that the analysis grid is represented by the boxed outline. *Bottom-Left*: A local map of ensemble spread in K^2 (shaded contours) and ensemble mean temperature (colored contours, 10K intervals). *Others:* First nine leading E-Vectors in temperature (shaded contours) with ensemble mean temperature overlaid (colored contours).



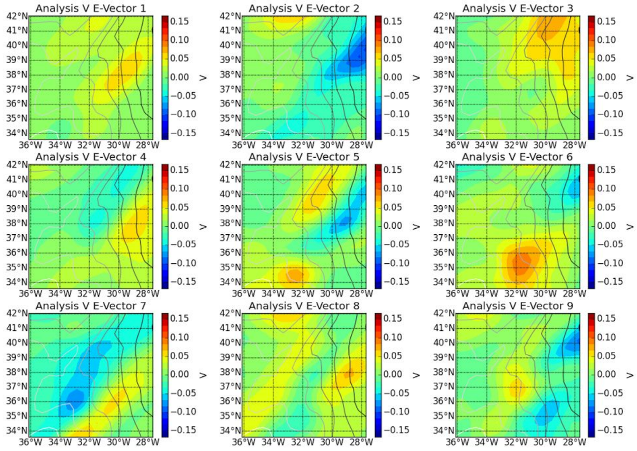
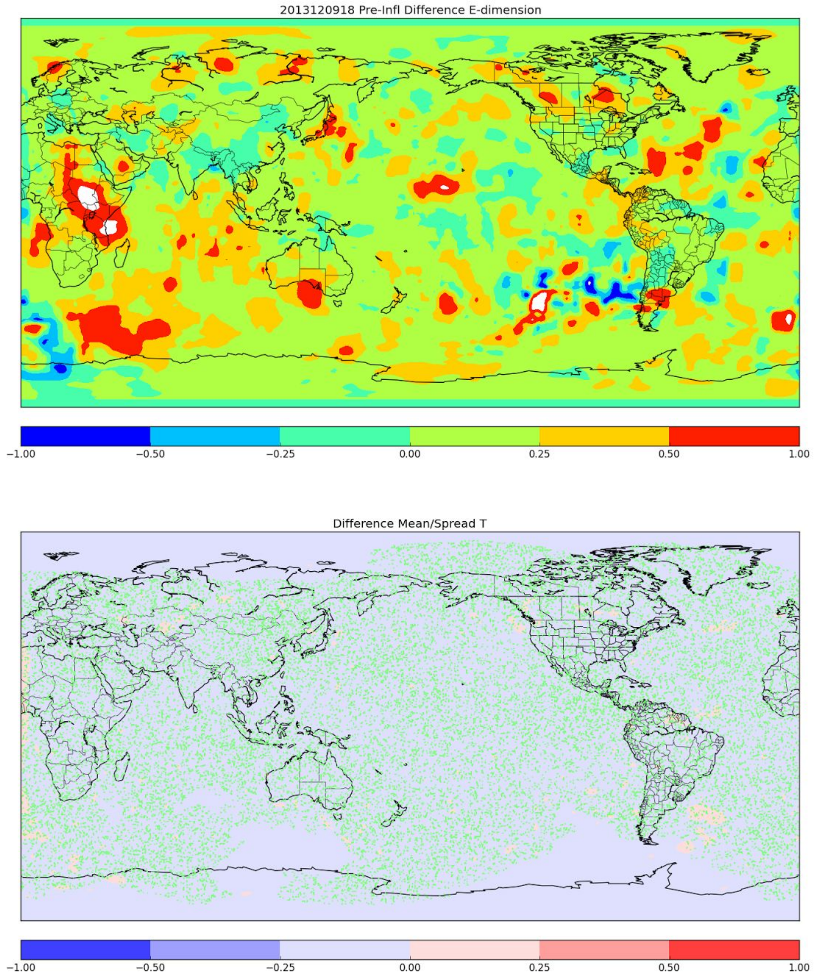


Figure 3: As the Figure 2 E-Vector plot, but for meridional wind; ensemble mean meridional wind (colored contours, 5m s^{-1} intervals).

Figure 4: As Figure 3 but for zonal wind.

4.2. EnKF comparison

In continuing the brief discussion presented in subchapters 3.1.1 and 3.3, we wished to perform a brief comparison of the operational EnSRF and LETKF systems available in the current version of the EnKF-enabled GSI. We compare here the pre-inflation analyses as to isolate the EnKF algorithm against external effects. As they are configured identically and assimilating the exact same set of observations, as well as consistently applying RTPS, we expected to find near-identical results with the ensemble mean, and indeed that is borne out in the bottom panel of Figure 5: the green and yellow outlines are near-machine-precision differences in the model mean field. The spread difference field is notably biased negative near-globally, showing how the LETKF algorithm better maintains ensemble spread through the analysis step.


2013120918 Pre-Infl Difference E-dimension

Difference Mean/Spread T

Figure 5: In a similar style to Figure 1. *Top*: A difference map of E-Dimension valid for the analyses represented by the EnSRF minus LETKF datasets at 2013120918. The white regions represent departures with absolute value greater than 1. *Bottom*: A difference map of ensemble spread (shaded contours, irregular intervals) and ensemble mean (colored contours). The ensemble mean difference between EnSRF and LETKF is zero to machine precision.

This is an important distinction between the two algorithms that is actually largely washed out by the choice of multiplicative inflation. One of the purported advantages of the LETKF algorithm is how it better maintains ensemble spread through its analysis, lessening the need for artificial inflation to prevent filter divergence and failure of the ensemble system. In the operational setup with RTPS, however, this advantage is largely diminished.

Nevertheless, we inspect the E-Vectors of the LETKF dataset in Figure 6, valid at the same time as Figure 2. We identify very similar structure all throughout the first nine leading E-Vectors shown: not only is the same characteristic tilt seen in the E-Vector field, but specific features appear identical. Looking at the second E-Vector for each data set, we identify a fairly common issue: E-Vector flipping. Specifically, there is a large feature in the north of the analysis field with a feature of the opposite sign in the southwest, but these features are opposite between the EnSRF and LETKF fields. We consider these features to be largely consistent, and an artifact of the mathematical process rather than a meaningful physical difference. Thus, when taking the correlation between the two vectors, we do so with an absolute value.

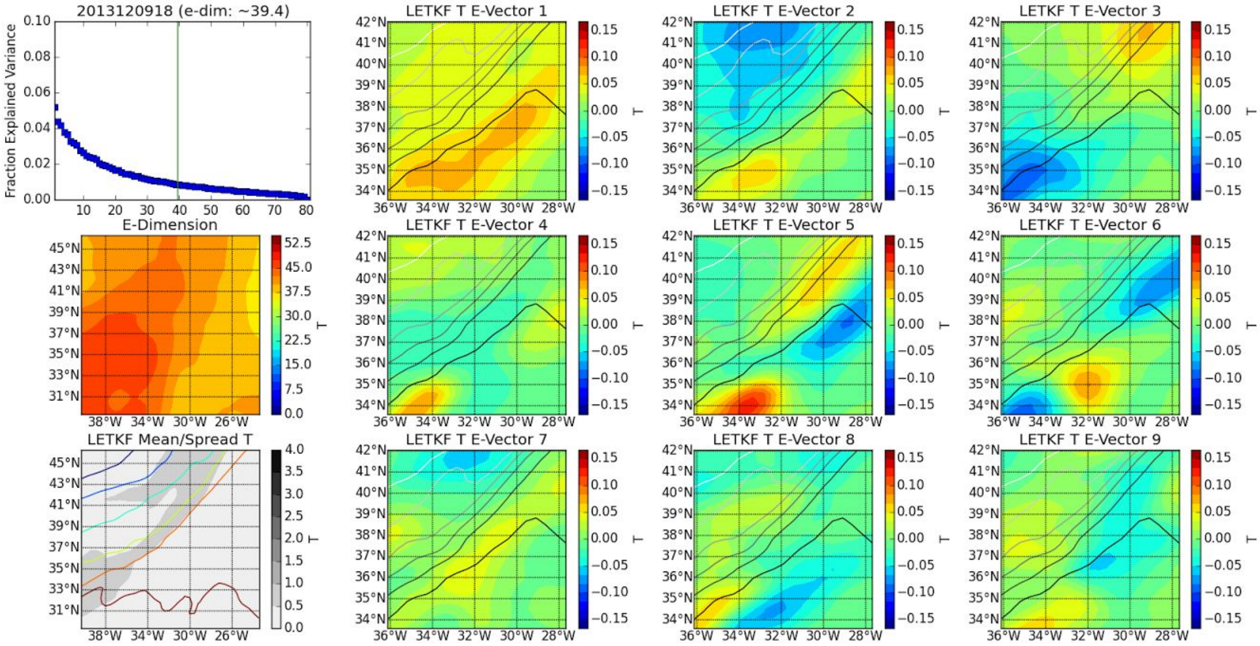


Figure 6: As Figure 2, but for the LETKF dataset valid at the same time.

While visual inspection allows for intuition toward the physical structure of the E-Vectors, it is insufficient for a rigorous comparison of the two fields, especially when trying to compare across all variables, and when comparing modes further from the dominant ones which have increasingly complex, non-physical structures. Figure 7 introduces the correlation box plot, which compares each of the 80 E-Vectors between the two datasets. We see a particularly strong diagonal structure, implying very strong correlation between the two data sets. The off-diagonal elements are occasionally relevant in the leading E-Vectors, such as the 7th and 8th E-Vectors which represent some combined information between themselves, but largely the broad inconsistencies are confined to E-Vectors with mode number higher than the local E-Dimension, and thus contribute very little to the overall ensemble variance structure.

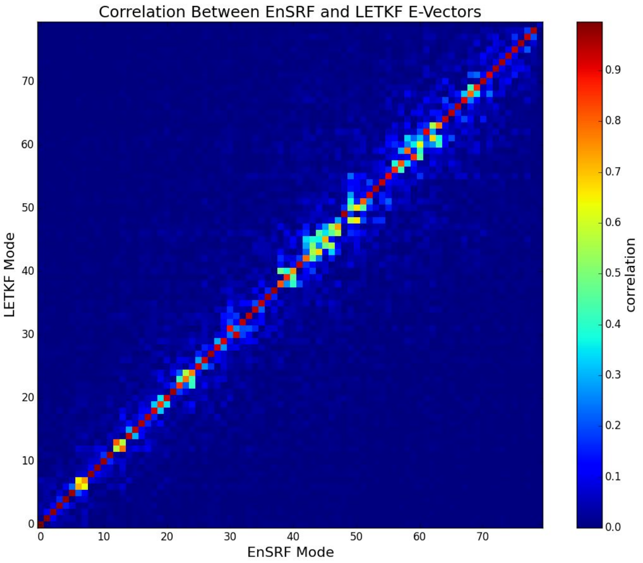


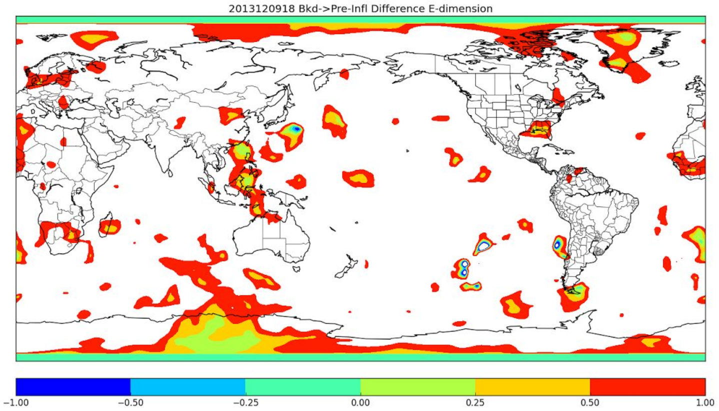
Figure 7: A correlation boxplot valid at 2013120918. With the axes increasing from the origin in the bottom-left, the x-axis represents the EnSRF E-Vectors and the y-axis represents the LETKF E-Vectors, with the leading modes beginning at the origin and going outward. Eqn. 5 has been applied to compute an absolute correlation between each E-Vector for each LEDEV analysis.

This is only one timestep, but these results are remarkably consistent across multiple timesteps and in various regions, both geographically through polar and tropical regions, and in various dynamical setups. Particularly under the constraints of RTPS in which the ensemble mean and spread are essentially identical for each timestep, there is little to differentiate the two. For the purposes of this study, we chose to continue with strictly the EnSRF for minimizing computational costs and to conform to the current operational system under the assumption that it has negligible effects on the analysis.

4.3. Data Assimilation Assessment

Having selected an EnKF algorithm to perform further analysis with, we wish to inspect each part of the ensemble update process to identify its effects on the E-Vectors. Continuing prior discussions, we reiterate that it is critical for the data assimilation process to adhere closely to observations in regions of greatest uncertainty, improving the ability of the ensemble to grow in the regions of greatest error growth to provide the most meaningful information possible to the EnKF algorithm. Specifically, we hope to find that the ensemble update process maintains the perturbation structure in the leading modes of error growth: ideally, there would be strong correlation along the diagonals of the correlation boxplot.

First looking at the EnKF step, Figure 8 shows the striking if unsurprising full-field differences: a near-global increase in E-Dimension upon the ingestion of the observational data. This follows from the reduction in ensemble spread inherent to the data assimilation process. The pattern shows little bias in any given region, except for the polar regions -- where observational data is relatively sparse -- showing less impact.



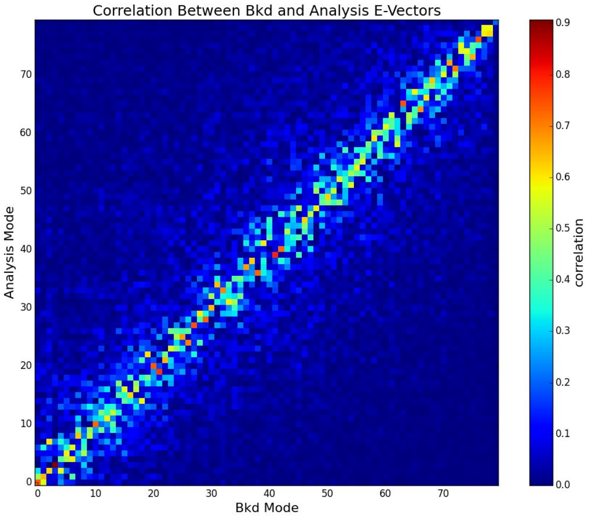


Figure 8: Global difference in two E-Dimension fields: with the background E-Dimension field subtracted from the pre-inflation analysis field.

Figure 9: As Figure 7, but comparing the EnSRF background forecast (x-axis) to the pre-inflation analysis (y-axis) valid at the same time.

Inspecting Figure 9, we find that with the decreased ensemble spread and substantial impact to the E-Dimension field, we find that there is notably more spread among the two sets of E-Vectors compared to the relatively tight diagonal structure of Figure 7. Some of the leading modes are altered entirely, and contain information from E-Vectors that were previously representing some half of the variance that they do in the analysis ensemble. Still, there is a reasonably strong diagonal structure, with a grouping along the off-diagonal limited to only five or six modes; this will not be the case in later experiments, so it should be noted here that this structure is actually quite consistent, even if it represents a loss of error-growth information from the background ensemble.

Revisiting Yang et al. (2015), whose results showed that using perturbation matrix singular vectors as additive inflation noise -- that is, adding stochastic covariance using the leading singular vectors as a basis -- could better align the ensemble members with the unstable manifold of the model. The same line of thinking applies to this analysis, where the loss of information inherent to the spread reduction of the EnKF could be mitigated by explicitly adding that information back into the ensemble perturbation matrix. This would reduce, or perhaps eliminate, the need for the multiplicative inflation, which we explore next.

Figures 10 and 11 are presented in an identical manner. The RTPS globally increases the spread of the ensemble, although it actually has a fairly inconsistent effect on the E-Dimension, substantially increasing the spread in some regions while substantially decreasing it in others. There is no latitudinal structure evident, although the regions in which E-Dimension is substantially altered by the ensemble spread are fairly consistent when comparing several timesteps next to one another, most notably the large region in the southeastern Pacific. It is tempting to ascribe this to its notorious marine stratocumulus, as another marine stratoculumus region off the western coast of southern Africa also exhibits this behavior, which indicates this may be a property of the model, although it is not particularly evident in the Northern Hemisphere.

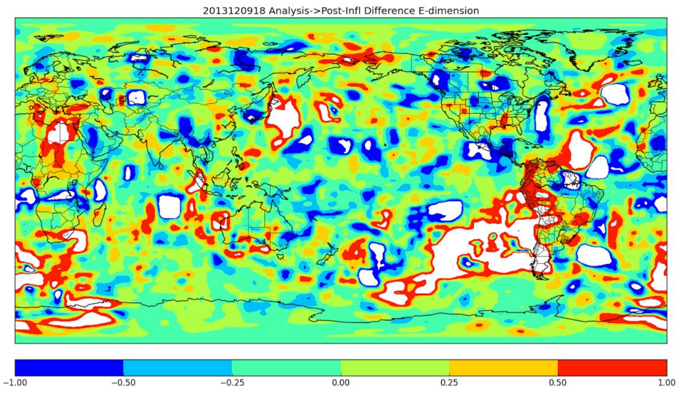


Figure 10: As Figure 8, with the pre-inflation analysis field subtracted from the post-inflation analysis field.

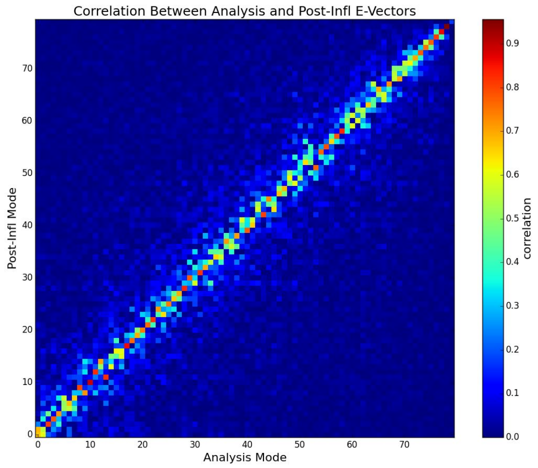
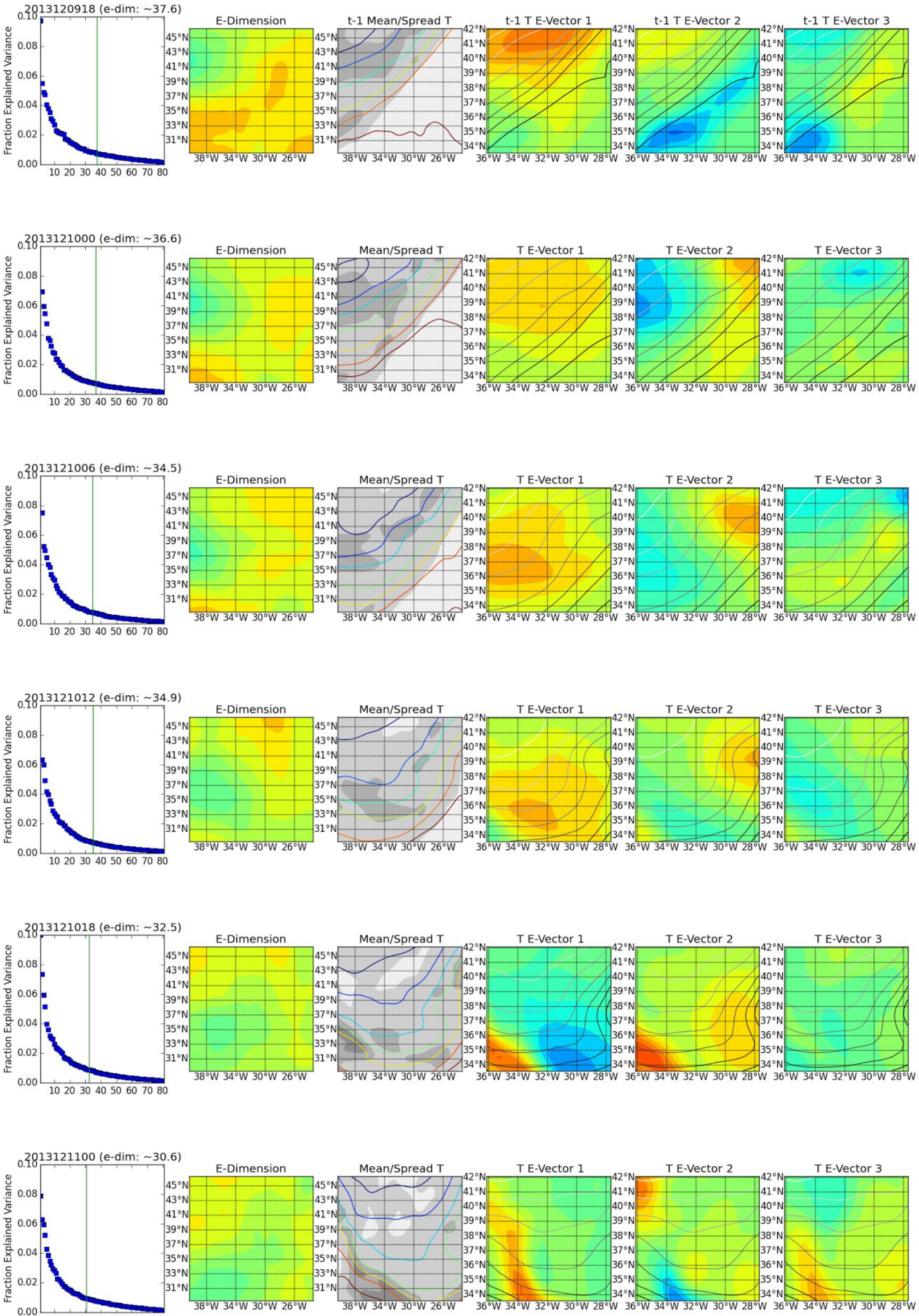


Figure 11: As for Figure 9, but for comparing The EnSRF pre-inflation analysis (x-axis) to the post-inflation analysis (y-axis) valid at the same time.

Inspecting the correlation boxplot, we find a diagonal structure tighter than that for the EnKF but still expressing some off-diagonal elements at the leading modes. Thus, we consider the RTPS inflation process to represent a minor loss of error-growth information, but not nearly as much as the EnKF process itself. However, in the context of prior literature suggesting that aligning the ensemble members with the unstable manifolds of the background forecast offering a meaningful improvement to the ensemble forecast, in conjunction with the relatively random alteration of the E-Dimension field, it would be ideal if the inflation process could be minimized, either through an additive inflation process as described above, another method like stochastic physics, or perhaps with a switch to the LETKF which in general requires less inflation than the EnSRF analysis.

4.4 Temporal Consistency

Now that we have explored the impacts of the EnKF on the structure of the ensemble perturbation matrix, we will exclusively look at 6-hour forecasts from this point forward, as they most clearly represent error growth in the model itself: the forecasts are bound to the model attractor, while mathematical increments to the ensemble perturbation matrix -- even when applying dynamical constraints -- might represent artificial, non-physical results. It is necessary to assess the temporal consistency of the E-Vectors to ensure that the process of cycling the model between forecasts and a data assimilation step does not successively destroy the E-Vectors, and that the evolution of the E-Vectors occurs in a manner physically consistent with atmospheric flow.


2013120918 (e-dim: ~37.6)
E-Dimension
t-1 Mean/Spread T
t-1 T E-Vector 1
t-1 T E-Vector 2
t-1 T E-Vector 3
2013121000 (e-dim: ~36.6)
E-Dimension
Mean/Spread T
T E-Vector 1
T E-Vector 2
T E-Vector 3
2013121006 (e-dim: ~34.5)
2013121012 (e-dim: ~34.9)
2013121018 (e-dim: ~32.5)
2013121100 (e-dim: ~30.6)
Fraction Explained Variance
0.00
0.02
0.04
0.06
0.08
0.10
10 20 30 40 50 60 70 80
45°N 43°N 41°N 39°N 37°N 35°N 33°N 31°N
38°W 34°W 30°W 26°W
42°N 41°N 40°N 39°N 38°N 37°N 36°N 35°N 34°N
36°W 34°W 32°W 30°W 28°W

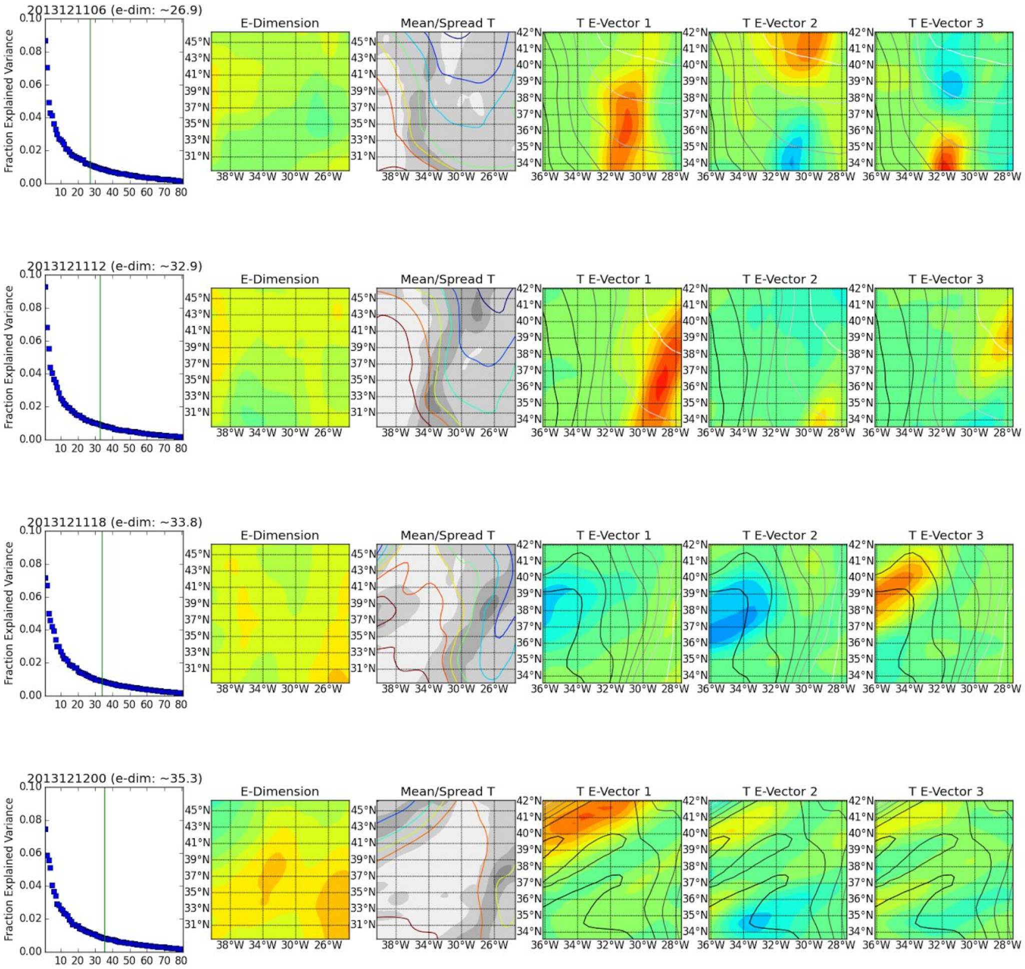


Figure 12: A multi-panel plot representing six-hour forecasts of the EnSRF dataset. Each row represents an increment of six hours from the prior row. The first column is a plot of the singular values, with a vertical line denoting the E-Dimension. The second column is a zoomed-out view of the local E-Dimension field. The third column is a plot of the ensemble spread in temperature (shaded contours, units of K^2) and ensemble mean temperature (contours, 10K intervals). The fourth through sixth columns represent the first three leading E-Vectors in temperature with a mean overlay.

Figure 12 presents ten successive six-hour forecasts, representing a total span of 36 hours, in our region of interest in the North Atlantic. Looking first at the evolution of the mean field, we can identify a region of high ensemble spread moving in from west and passing out to the east, eventually out of the analysis frame entirely. The shape of the temperature field indicates the passage of a trough, which begins negatively tilted and evolves into a positive tilt before continuing east out of the analysis region. The E-Dimension field itself evolves consistently, with a patch of low E-Dimension following the passage of the trough. The minima passes slightly south of the center analysis gridpoint. The E-Dimension at the gridpoint itself drops as the trough passes, and then increases as it leaves the frame.

Inspecting the evolution of the leading E-Vector, the first four frames represent a relatively uninteresting, fairly static setup. The leading E-Vector changes little in this time, and is likely the product of the mathematical SVD process rather than a physical phenomena. However, starting with the fifth frame, we see a marked change in the E-Vector structure as a strong feature enters the southwest portion of the analysis region. This feature appears to have a negative tilt, and the next frame affirms the assessment: the leading E-Vector is aligned with the trough axis. Through the next three frames, the feature progresses eastward, and maintains its alignment with the trough axis as it begins its motion toward a positive tilt.

The leading temperature E-Vectors for this case study appear to be a product of both the ensemble mean and the ensemble spread: they are often aligned with the mean field, but are expressed most strongly in regions of strong ensemble spread. The model coherently evolves the ensemble members -- and indirectly the mean field -- and the EnKF is tuned with the purpose of maintaining a proper ensemble spread: it follows that the evolution of the E-Vectors is consistent through time with the evolution of the atmospheric flow. Despite a reduction in error-growth information introduced at each time step by the data assimilation process, the hybrid cycling system proves to be self-consistent with the evolution of the ensemble perturbation matrix.

4.5. Vertical Consistency

With the result from the prior section that the structure of the E-Vectors combining information from the mean field and the ensemble spread, it should be logically consistent that the E-Vectors at different model levels express a coherence as well. We will investigate the spatial consistency of the E-Vectors by visually inspecting the vertical structure of the system, examining the structure of the E-Vectors at several model levels throughout the troposphere, specifically looking at two adjacent model levels, and then including two more that are substantially removed from 850mb level. Figures 13-16 represent model levels 15, 16, 25, and 33, corresponding to ~850mb, ~825mb, ~500mb, and ~200mb respectively.

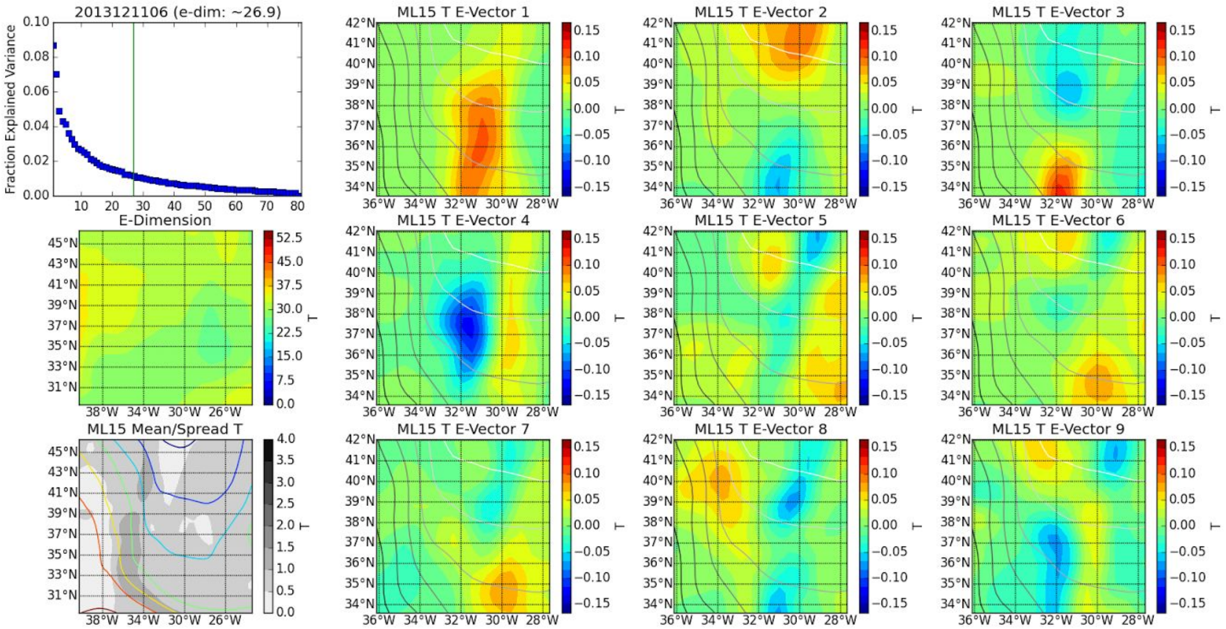


Figure 13: As Figure 2, but valid at 2013121106.

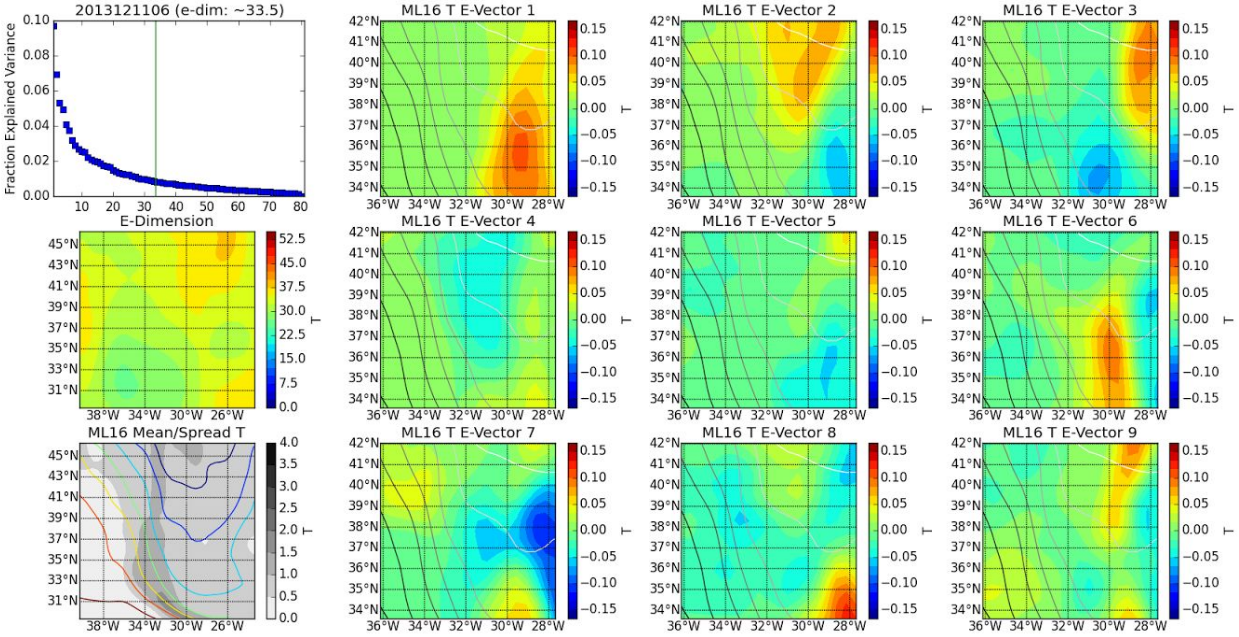


Figure 14: As Figure 13, but valid for model level 16.

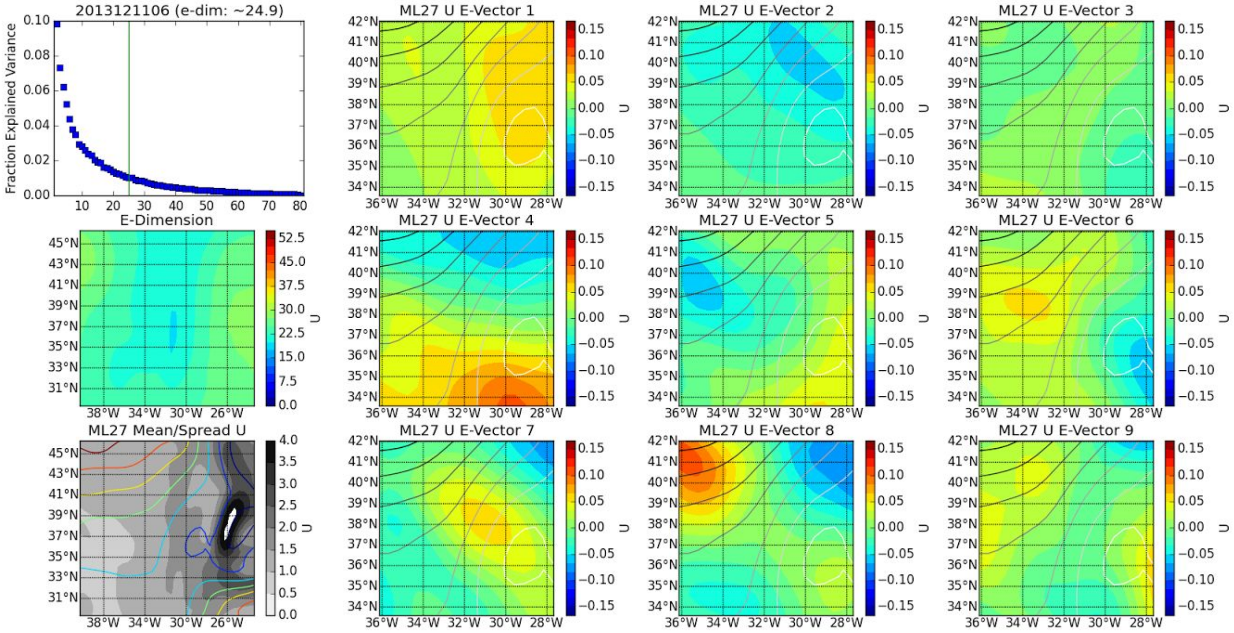


Figure 15: As Figure 13, but valid for model level 27 and for zonal wind U.

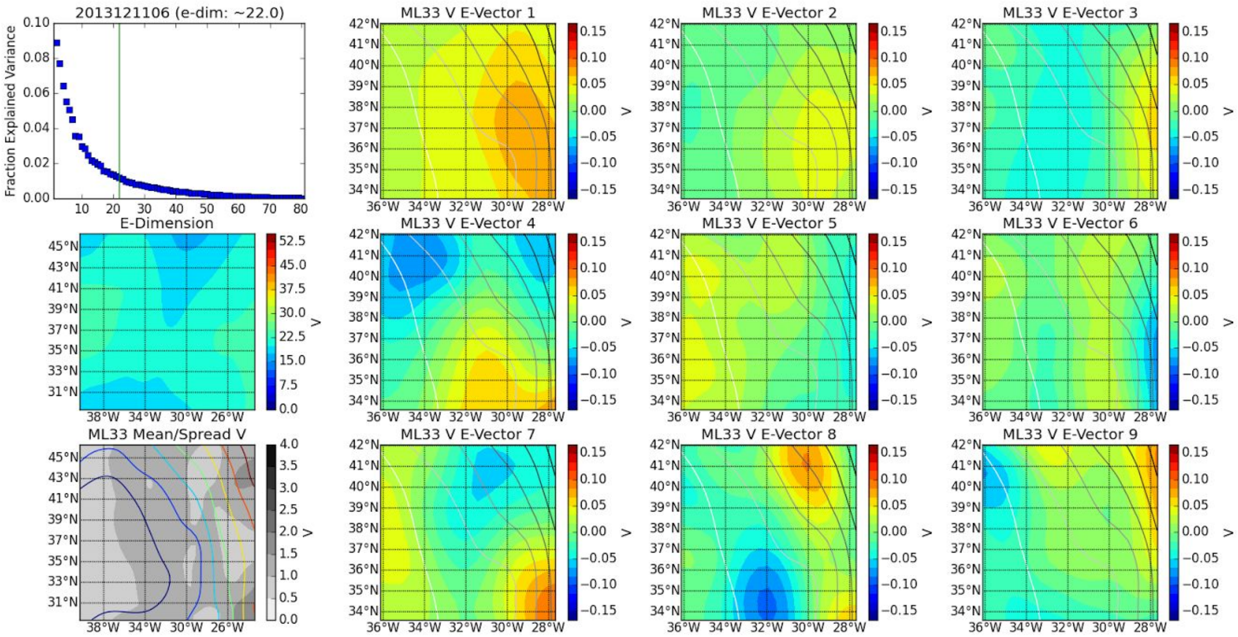


Figure 16: As Figure 13, but valid for model level 33 and meridional wind V.

We have chosen 2013121106 so that the leading E-Vector at model level 15 would be expressed as strongly as possible and in the center of the analysis field. Comparing Figure 13 to Figure 14, we find that the feature is strikingly present in largely the same quality but shifted slightly to the east, suggesting that there is a tilt toward the east with height. This would indicate that the low is in the process of including and will soon weaken; indeed, this is exactly the fate of the extratropical cyclone as it ceases to present at the surface in just 36 hours from the time of this analysis. This apparent consistency between E-Vectors continues for several modes, with a slight shift to the east, before starting to diverge around the fifth E-Vector.

Looking further up into the mid-troposphere, we turn to Figure 15, located roughly around the level of non-divergence (LND). We have switched to the zonal wind E-Vector here, as the temperature E-Vector exhibits very little dynamic activity, thus very little ensemble spread; it is in fact nearly flat across its leading E-Vectors. The zonal wind, however, has rather significant ensemble spread, and its E-Vectors are strongly expressed. Given that this level is substantially removed our ~850mb LEDEV analyses, it is not surprising that the level of consistency is substantially reduced; there is little evidence apart from the first E-Vector that the two are strongly correlated. However, that first E-Vector is again slightly shifted east, existing less sharply in the eastern half of the analysis domain.

The LEDEV analysis for Figure 16 is done roughly at the tropopause, some 5 kilometers in altitude above the LND. Again, the temperature E-Vectors have essentially no expression at this level, but the wind E-Vectors present incredibly strongly. For the sake of consistency we could have shown the zonal wind E-Vector, but the meridional wind E-Vector better highlights continued eastward tilt of the system with height: the zonal geostrophic wind in the mid-troposphere turns into meridional geostrophic wind on the western edge of the upper-level cyclone, and the first four meridional wind E-Vectors at the tropopause exhibit some correspondence to the zonal wind E-Vectors at the LND.

Straightforward boxplot correlations, as we have used previously, are entirely insufficient for this type of analysis. Because the expression of the unstable manifold differs so wildly with height, particularly with the added wrinkle of a substantial tilt introduced by the aging extratropical cyclone, naive comparison of two model levels may be entirely insufficient to assess the E-Vector consistency. We have identified a logical consistency to the E-Vectors with altitude, bolstering the LEDEV analysis and suggesting further physical consistency to the ensemble perturbation matrix.

4.6. Spatial Consistency

Finally, with both the dynamical and informational consistency of the LEDEV analysis having been explored, we seek to identify the robustness of our results. Oczkowski et al. (2005) varied the size of their analysis domain and found that although the E-Dimension changes substantially with the addition or removal of variables, the structure of the E-Dimension field largely does not. We will briefly extend this analysis to include the E-Vectors with two separate approaches, first involving a variation in domain size, and then inspecting adjacent gridpoints.

Returning to the ~850mb level, Figure 17 shows the E-Vectors with the analysis grid shrunk by 1, i.e. instead of being a 19x19 analysis grid, it is a 17x17 analysis grid, thus losing some 20% of the variables present in the regular analysis. Comparing to Figure 13, we find quite striking similarities throughout the entire first nine E-Vectors, a rather remarkable result for losing so much of the analysis grid. The intuition is confirmed with Figure 18, as a very strong diagonal structure is evident through the first fifteen E-Vectors before any substantial off-diagonal expression is present.

It is notable that the box plot is largely upper-rectangular in nature. This implies that the shrunken grid's E-Vectors contain information from less-relevant E-Vectors, but the opposite is not true: in other words, the smaller domain is incorporating some of the randomness present in the larger domain into its analysis, but the larger domain sees no such randomness from the smaller E-Vectors, instead spreading its variance downward. This may suggest that the standard grid size selected for most of these LEDEV analyses was inappropriately large, although such a conclusion is certainly not definitive.

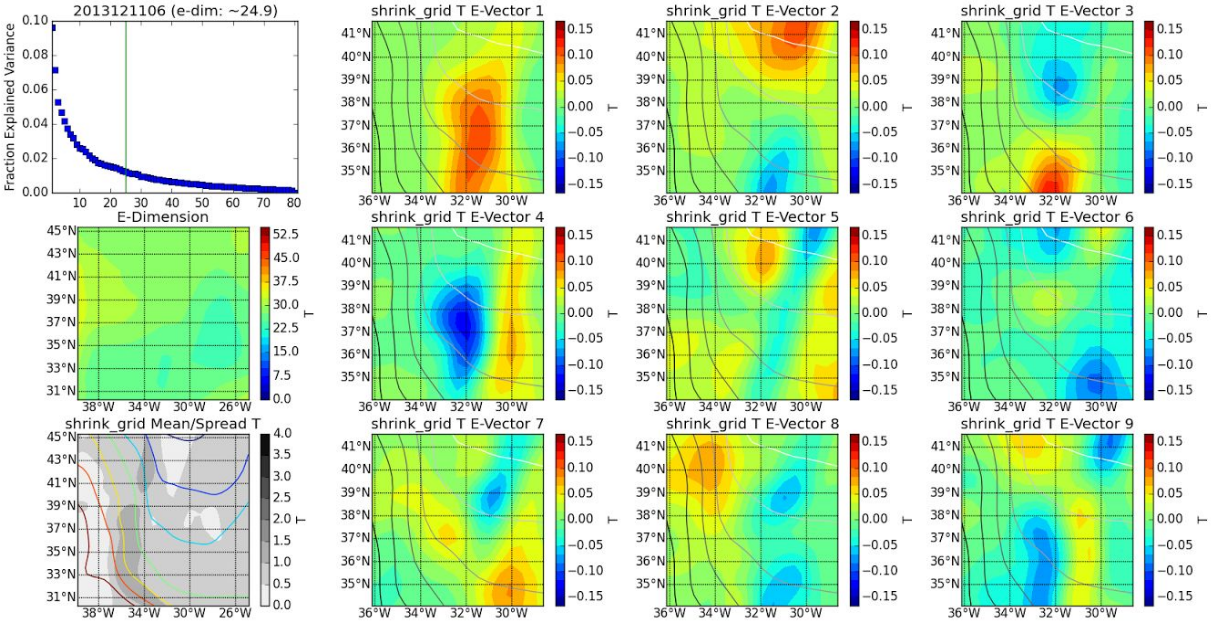


Figure 17: As Figure 13, for a reduced domain size.

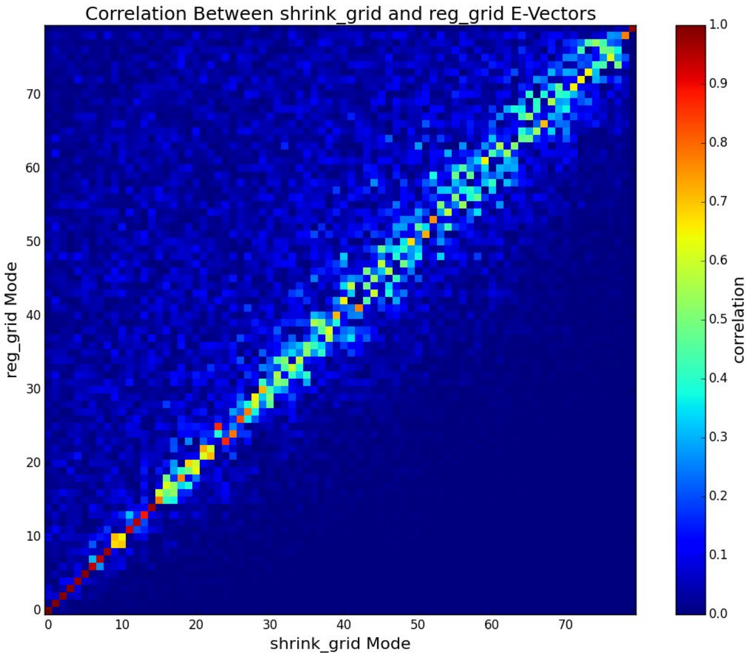


Figure 18: As Figure 7, but corresponding to Figures 13 and 17, applying Eqn 5 on the overlapping components of the E-Vector domains.

Finally, we inspect an adjacent gridpoint's E-Vectors with the exact same analysis conditions otherwise. Figure 19 shows the E-Vectors of the gridpoint immediately east of our original gridpoint. Such an analysis loses the left edge of the analysis domain while gaining a right edge not present in the original analysis, representing a ~10% change from the original domain. Similar to the previous analysis, the E-Vectors look strikingly similar, but visually shifted to the left, suggesting a nearly-exact representation of the E-Vectors. Similar to the domain-change experiment, the box correlation plot in Figure 20 -- comparing only the portions of the domains that overlap -- shows an incredibly strong diagonal structure over the first twenty E-Vectors with fairly tight off-diagonal grouping beyond that, indicating very strong agreement, emphasizing the substantially robust nature of these analyses.

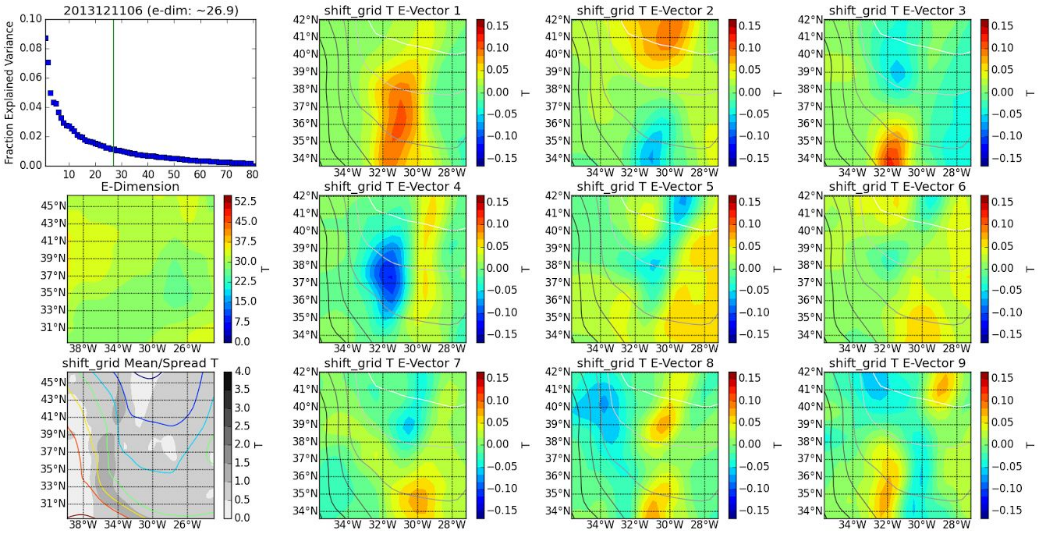


Figure 19: As Figure 13, but with the analysis gridpoint shifted zonally by one.

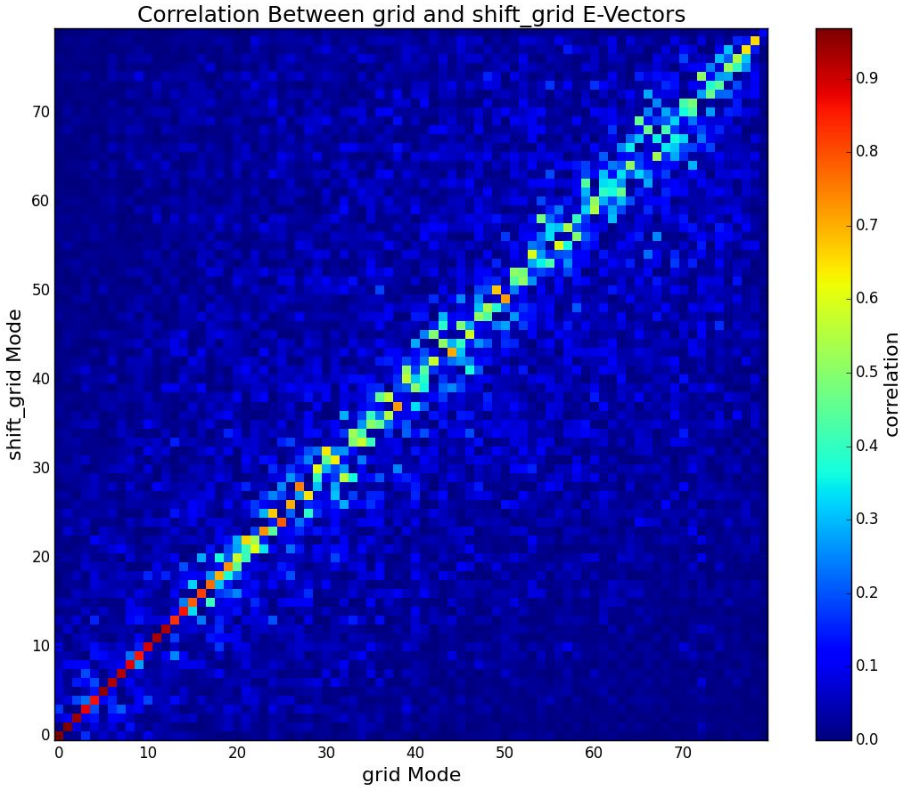


Figure 20: As Figure 7, but corresponding to Figures 13 and 19, specifically applying Eqn 5 on the overlapping components of the E-Vector domains.

6. Conclusions

6.1 Summary of Research

It has been demonstrated by repeated analysis of a passing mature extratropical cyclone in the Atlantic Ocean that LEDEV analysis can be used to identify a physical interpretation of the unstable manifolds of the GFS, and such analyses are both spatiotemporally robust and resistant to the inherent loss of error-growth information during the data assimilation process.

Specifically, we sought to confirm that spatiotemporal consistency in the model would lead to spatiotemporal consistency in LEDEV analysis. Indeed, we discover that the features present in LEDEV analysis advect coherently with the prevailing atmospheric flow and evolve in structure according to dynamical changes of the non-linear system. Additionally, we find that the vertical structure of single-level LEDEV analysis is consistent with dynamic expectations, as the mature cyclone was tilting eastward in height, and the E-Vectors -- although changing in nature with height -- showed substantial logical consistency. We also performed some basic horizontal consistency checks, determining the robustness of the results; the results in the experiments would not substantially change due to minor alterations to the experimental design.

Additionally, after a brief comparison of the LETKF and EnSRF, determining that their treatment of E-Vectors is largely consistent in the context of a hybrid cycling system with RTPS inflation, we found that the EnKF inherently causes a loss error-growth information when compressing the ensemble spread. Additionally, the RTPS inflation heterogeneously alters the E-Dimension field despite expanding the ensemble spread everywhere, and we suggest that any method of limiting the need for such an inflative process -- whether with additive inflation, stochastic physics, or a different EnKF algorithm -- can only improve the behavior of the EnKF overall.

5.2. Future Research Direction

While this research shows immediate robustness, it would substantially benefit from additional case studies. Only looking at one Northern Hemisphere mid-latitude case in one season is insufficient to draw any definitive conclusions, and as such we emphasize that any speculative conclusions drawn here are perhaps more speculative than conclusive.

This work, in conjunction with Yang et al. (2015) suggests an excellent basis for performing a local breeding method. Traditionally, Kalnay and Toth (1997) performed their breeding experiments with near-hemispheric bred vectors, but with the at-times excellent coherency of the E-Vectors and simple translation between the E-Vectors and the ensemble perturbation matrix, a more localized approach now seems appropriate.

One potential application of this feature could be to apply an ensemble forecasting technique to a whole-atmosphere model. The upper atmosphere is notoriously dissipative, meaning that initial conditions are often largely irrelevant on the atmospheric timescales of the lower atmosphere. In order to operate a successful ensemble in this region in a single model system, it is likely that more advanced techniques will be required to generate enough ensemble spread to successfully operate a hybrid data assimilation system.

5.3. Acknowledgements

Computing resources were provided by the National Atmospheric and Oceanic Administration (NOAA) through funding provided through a National Aeronautics and

Space Administration (NASA) Award #NNX14AI17G for "The Heliophysics Grand Challenges Research - Predictability and Ensemble Modeling of the Space-Atmosphere Interaction Region (SAIR).” The LEDEV analysis tool that was developed for this project was built upon existing code from Dr. Daryl T. Kleist and Dr. Jeff Whitaker.

Appendix

The LEDEV analysis tool was developed in the interest of streamlining the process of analyzing large ensemble datasets using the LEDEV technique, attempting to automate every step of the process while providing standard visualization tools for progressive analysis and maintaining a broad flexibility in analysis options. It consists of two components: a Fortran tool that performs a global E-Dimension calculation, and a Python tool that is used for location-specific analysis and visualization. As demonstrated, there exist a variety of options to cater to a specific analysis needs, but some have not been presented in this study. The package has been designed to cater to GSI output, although the code is easily adaptable to other models.

The Fortran tool is intended to be run in a highly parallel environment, as it utilizes both MPI and OpenMP directives to vastly improve the computational efficiency of the program. A straightforward serial processing of a T254 global perturbation field requires O(10^5) SVDs to be performed on matrices with O(10^4) entries. Several rounds of parallelization followed that improved its performance by roughly two orders of magnitude.

The Python tool performs the bulk of the E-Vector work, as storage of the E-Vectors from the Fortran decompositions unnecessarily takes up several GB of disk space, exacerbating IO issues when processing many files at once. It has been tested on both NOAA and NASA machines, but should be easily portable to any machine provided libraries are available for interfacing with the model hybrid spectral files.

Comprehensive LEDEV Analysis Package documentation is available, including a tutorial. Further information regarding the tool is available upon request.

References

Anderson, J. L., and N. Collins, 2006: Scalable Implementations of Ensemble Filter Algorithms for Data Assimilation. *J. Atmos. Oceanic Technol.,* **24**, 1452–1463.

Anderson, J. L., 2007: An adaptive covariance inflation error correction for ensemble filters. *Tellus A,* **59**, 210–224.

Benzi, R., R. Deidda, and M. Marrocu, 1997: Characterization of temperature and precipitation fields over Sardinia with principal component analysis and singular spectrum analysis. *Int. J. Climatol.*, **17**, 1231–1262.

Bishop, C. H., and Z. Toth, 1999: Ensemble Transformation and Adaptive Observations. *J. Atmos. Sci.*, **56**, 1748–1765.

Corazza, M., E. Kalnay, and D. J. Patil, 2002: Application of Bred Vectors To Data Assimilation. *EGS General Assembly Conference Abstracts*, **27**, 775.

Costabile, F., W. Birmili, S. Klose, T. Tuch, B. Wehner, and A. Wiedensohler, *et al.,* 2009: Spatio-temporal variability and principal components of the particle number size distribution in an urban atmosphere. *Atmos Chem Phys*., **9**, 3163–3195.

Enomoto, T., S. Yamane, and W. Ohfuchi, 2015: Simple Sensitivity Analysis Using Ensemble Forecasts. *J. Met. Soc. Jap.*, **93**, 199–213.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I., 2007: Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.

Kaplan, J. L., and J. A. Yorke, 1979: Chaotic behavior of multi-dimensional difference equations. *Functional Differential Equations and Approximations of Fixed Points: Proceedings, Bonn, July 1978*, Vol. 730, H.-O. Peitgen and H.-O. Walther, Eds., Springer-Verlag, 204–227.

Keller, J. D., A. Hense, L. Kornblueh, and A. Rhodin, 2010: *On the Orthogonalization of Bred Vectors*. *Wea. and For.*, **25:4**, 1219–1234.

Oczkowski, M., I. Szunyogh, and D. J. Patil, 2005: Mechanisms for the Development of Locally Low-Dimensional Atmospheric Dynamics. *J. Atmos. Sci.,* **62**, 1135–1156.

Patil, D. J., B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2001: Local low-dimensionality of atmospheric dynamics. *Phys. Rev. Lett*., **86**, 5878–5881.

Szunyogh, I., E. J. Kostelich, G. Gyarmati, D.J. Patil, B. R. Hunt, E. Kalnay, E. Ott, and J.A. Yorke, 2005: Assessing a local ensemble Kalman filter: perfect model experiments with the National Centers for Environmental Prediction global model. *Tellus A*, **57**, 528–545.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*., **125**, 3297–3319.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc*., **74**, 2317–2330.

Wang, X., D. Barker, C. Snyder, and T. M. Hamill, 2008: A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. *Mon. Wea. Rev.*, **136**, 5116-5131.

Westra, S., Brown, C., Lall, U., Koch, I. and Sharma, A., 2010: Interpreting variability in global SST data using independent component analysis and principal component analysis. *Int. J. Climatol.*, **30**, 333–346.

Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev*., **130**, 1913–1924.

Yang, S.-C., E. Kalnay, and T. Enomoto, 2015: Ensemble singular vectors and their use as additive inflation in ENKF. *Tellus A*, **67**, 26536, http://dx.doi.org/10.3402/tellusa.v67.26536