Assessment and quantification of HF radar uncertainty

Fearghal O'Donncha*, Sean McKenna*, Teresa Updyke†, Hugh Roarty‡ and Emanuele Ragnoli*
*IBM Research - Ireland
Email: feardonn@ie.ibm.com
†Old Dominion University
‡Rutgers University

***Abstract*—A large body of work exists concerning uncertainty in ocean current measuring high-frequency radar (HFR) systems. This study investigates the magnitude of uncertainty present in a HFR system in the lower Chesapeake Bay region of Virginia. A method of assessing the fundamental performance of the HFR is comparing the radial velocities measured by two facing HF radars at the centre point of their baseline. In an error-free network, radial vectors from the two sites would be equal and opposite at a point on the baseline, so the magnitude of their sum represents a measure of imperfection in the data. Often essential information lies not in any individual process variable but in how the variables change with respect to one another, i.e. how they co-vary. PCA is a data-driven modelling technique that transforms a set of correlated variables into a smaller set of uncorrelated variables while retaining most of the original information. This paper adopts PCA to detect anomalies in data coming from the individual HF stations. A PCA model is developed based on a calibration set of historical data. The model is used with new process data to detect changes in the system by application of PCA in combination with multivariate statistical techniques. Based on a comprehensive analysis the study presents an objective preconditioning methodology for preprocessing of HFR data prior to assimilation into coastal ocean models or other uses sensitive to the divergence of the flow.**

I. Introduction

The technology of measuring surface current by high frequency radar (HFR) has been rapidly expanding over the last decade [1], having been used to study nearshore circulation in a large variety of environmental conditions [2]–[6]. HFR allows measurement along the conductive sea surface for distances of up to 200km offshore at time intervals of 0.2-1h [7]. HFR systems have a number of unique advantages in terms of the observation of coastal ocean dynamics. These include: providing real-time data over large ocean areas at relatively low cost; enabling two-dimensional mapping of surface currents at resolutions that capture the complex structure related to coastal bathymetry and the intrinsic instability scales of the coastal circulation; as systematic input to operational ocean models via data assimilation [8]; while HFR systems can also play a role in environmental monitoring and event response systems.

A large body of work exists concerning uncertainty in ocean current measuring HFR systems. A study by Emery et al. (2004) [9] comparing HFR and moored current meters in the Santa Barbara basin indicated rms differences of $7-19cm/s$. In a similar study by Essen at al. (2000) [10], the accuracy of HFR was assessed by comparison with in situ current meters. RMSD were in the range of $10-20cm/s$; however, the theoretical error of the HFR based on the sea state was estimated to only be in the range $3-10cm/s$. The rest was assumed to be due to differences in the quantities measured, e.g. the spatial averaging, point in water column at which measurement recorded, etc.

Much of this work, however, focuses on direct comparisons of radar observation versus an alternate sensor measurement, be it ADCP, drifters or other current measuring instruments. However, these comparisons introduce inherent complexities due to additional errors being introduced from the second sensor and also what is termed target difference: discrepancies between both sensors due to the HFR typically measuring different spatial and temporal scales. This study aims to isolate individual errors in a HFR system; quantify the magnitude of the error in a historical dataset; and finally, develop a transportable algorithm that can be used to establish the uncertainty in a real-time measuring system.

This paper describes research conducted by the authors in assessing HFR uncertainty and the definition of a preconditioning technique to lessen the impact of potential errors on operational applications. A detailed dataset of HFR observed currents was collected at 60 minute intervals for a 12 month period (2012) encompassing a wide range of environmental conditions. This dataset is used to provide insight into error magnitudes associated with HFR systems. A multivariate analysis procedure, Principal Component Analysis (PCA) is used to detect anomalous measurements and reconstruct the data with a reduced number of modes.

The approach adopted by the authors is presented in the section on methodology; this section includes a description of both the HFR system and the PCA methodology. The process of reconstructing the data is described and the validation of the technique against new data discussed. The section on results presents a quantitative investigation of HFR error ranges; the viability of using PCA to identify and reduce anomalous data measurements is discussed. Finally, conclusions from this research are drawn and the recommendations for future research made.

II. Methodology

High frequency (HF) radar surface current data were provided by three radar systems located in the lower Chesapeake Bay region of Virginia. Figure 1 presents the geometric configuration of the three sites. These radar stations operate at 25 MHz and are a part of the Mid-Atlantic Regional Association Coastal Ocean Observing System (MARACOOS). At each site, radial current velocities were determined following the method described in Lipa et. al. (2006) [11]. Radial maps were generated with velocity vectors placed in 1.5 kilometre range bins and 5 degree directional bins. Radial processing algorithms utilized antenna response patterns measured at VIEW and CPHN stations. An ideal antenna response pattern was assumed at SUNS. Hourly surface current maps were produced by a standard un-weighted least squares method of combining radial data from individual radar sites onto a defined grid [12]. The grid in this case was a nominally 2 kilometer spaced grid developed by the U.S. National HF radar network [13]. Vector measurements returned hourly data and the data covered a one year period, January - December 2012 (8784 hours).

The study investigates a number of techniques to elucidate the inherent uncertainty of the system. As a means of assessing the fundamental performance of the HFR, analysis compares the radial velocities measured by two facing HF radars along their baseline. This serves to localise data uncertainty as the target difference is negligible if, both, the comparison is made at the middle of the baseline and the electromagnetic wave frequencies of the two sensors are the same. In an error-free network, radial vectors from the two sites would be equal and opposite at a point on the baseline, so the magnitude of their sum represents a measure of imperfection in the data.

Often essential information lies not in any individual process variable but in how the variables change with respect to one another, i.e. how they co-vary. PCA is a data-driven modelling technique that transforms a set of correlated variables into a smaller set of uncorrelated variables while retaining most of the original information. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

In computational terms the principal components are found by calculating the eigenvectors and eigenvalues of the data covariance matrix. In the case of vector observation (HFR velocities in the horizontal plane), it is convenient to represent the flow as complex number $\vec{u} = u + iv$, where u and v are the zonal and meridional components of flow respectively. The data matrix (X) is constructed where each row is one map of HFR measurements and each column is a time series of observations for a given location. The data are detrended so that each column has zero mean, the covariance matrix computed by calculating $R = X^T X$, and then we solve the eigenvalue problem

$$RP = P\lambda \tag{1}$$

λ is a real diagonal matrix containing the eigenvalues λ_i of R. The p_i column vectors of P are the eigenvectors of R corresponding to the eigenvalues of λ_i.

For each eigenvalue λ_i chosen we find the corresponding complex eigenvector p_i. Each of these eigenvectors can be regarded as a map. These eigenvectors are the principal components (PC) of the data. Each eigenvalue λ_i gives a measure of the fraction of the total variance explained by the mode. This fraction is found by dividing the λ_i by the sum of all the other eigenvalues.

The pattern obtained when an eigenvector is plotted as a map represents a standing oscillation. The time evolution of an eigenvector shows how this pattern oscillates in time. To see how PC_1 'evolves' in time we calculate

$$\overrightarrow{t_1} = X\overrightarrow{p_1} \tag{2}$$

The n components of the vector $\overrightarrow{t_1}$ are the projections of the maps in X on PC_1, and the vector is a time series for the evolution of PC_1. In general for each calculated PC_j, we can find a corresponding $\overrightarrow{a_j}$. These are the *principal component time series* or the *expansion coefficients* of the PCs. Just as the PCs were uncorrelated in space, the expansion coefficients are uncorrelated in time. We can reconstruct the data from the PCs and the expansion coefficients:

$$X = \sum_{j=1}^{p} \overrightarrow{a_j}(p_j) \tag{3}$$

A common use of PCA is to reconstruct a cleaner version of the data by truncating this sum at some $j = N << p$, that is, we only use the PCs of the few largest eigenvalues. The rationale is that the first N eigenvectors are capturing the dynamical behaviour of the system.

III. Results

A. Baseline Comparisons

Prior to more detailed comparisons, a direct comparison of the radial velocity measured by the individual radar station along the baseline between sites is investigated. Previous studies have demonstrated significant differentials when baseline radial values are compared away from the central region due to disparate horizontal averaging scales within the radial cells [14], [15]. In this study, a midpoint between the two radars is selected and all radial measurement within a 1km radius of that point gathered from both sites.

Figure 1 presents the geometry of the radar sites and baselines. Figure 2 shows scatterplots of hourly radial velocities at the midpoint of SUNS–CPHN (top), SUNS–VIEW (middle) and CPHN–VIEW (bottom). All statistics were computed for a one month period in December 2012. The solid line is the regression line obtained from the principal component analysis (PCA) which minimizes the sum of the square distance from the point (x, y) to the regression line $(y = Ax + B)$. PCA is particularly suitable for this analysis because it provides the symmetric regression line with respect to the two variables in scatterplots, as opposed to other measures of regression such

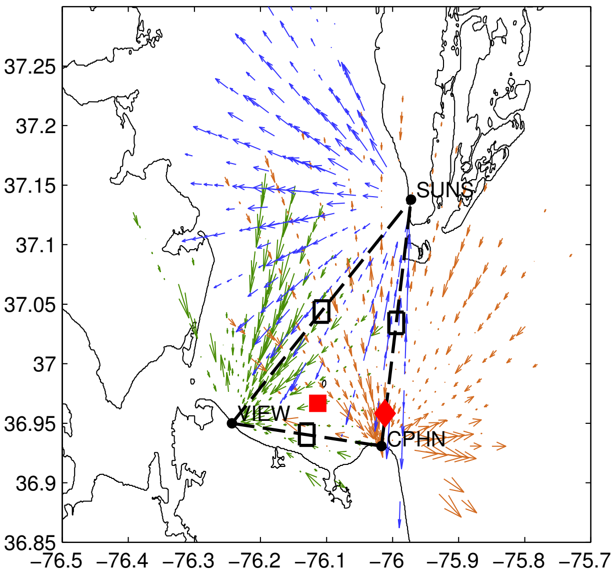

Fig. 1: Radial current synoptic vector map along with the baseline between HF radar sites along and the mid-point sampling region where radial values were compared (black rectangle). The red diamond and rectangle denotes the location of ADCP and weather station, respectively, used for the study.

as ordinary least squares which are more suitable for predictor-observed comparisons. In addition, rms distances from the regression line can be readily computed as an estimate of the uncertainty in the radar. The regression coefficients (A and B), correlation (COR), root-mean-square differentials (RMS), and number of samples (NUM) are also presented.

Good agreement is observed between two of the radar pairs (namely, SUNS–CPHN and SUNS–VIEW) reflected in correlation scores of 0.81 and 0.84 respectively. The baseline between SUNS–VIEW demonstrates very high agreement with regression coefficients of $(A = 0.91, B = -3.23 cm/s)$. Regression line coefficients from the SUNS-CPHN site $(A = 0.61, B = 2.46 cm/s)$ suggests that the variance from the SUNS site is almost 40% greater than the CPHN site. The relatively high rms figures between these sites further illustrates this. These agreement metrics are similar to comparable studies in other HFR systems. In comparisons of four baseline geometries in the Monterey Bay region, Paduan et al. [16] observed a linear regression relationship ranging in slope from 0.63 to 0.98, while correlation coefficients ranged from $0.6 - 0.8$. Similar analysis of HFR accuracy in the Tsushima Strait [15], observed correlation in the range of 0.63 - 0.88 was returned while the RMS varied between $5.75 - 13.71 cm/s$.

Baseline comparisons between CPHN–VIEW provides an interesting contrast. There is no evident agreement between values measured by the facing radar stations. Further investigation of this identified the cause to be a thin strip of land approximately 600m long beside the CPHN station over which the baseline HFR signal travels before reaching open water. This serves to distort the signal in this direction and result in contaminated data measurement.

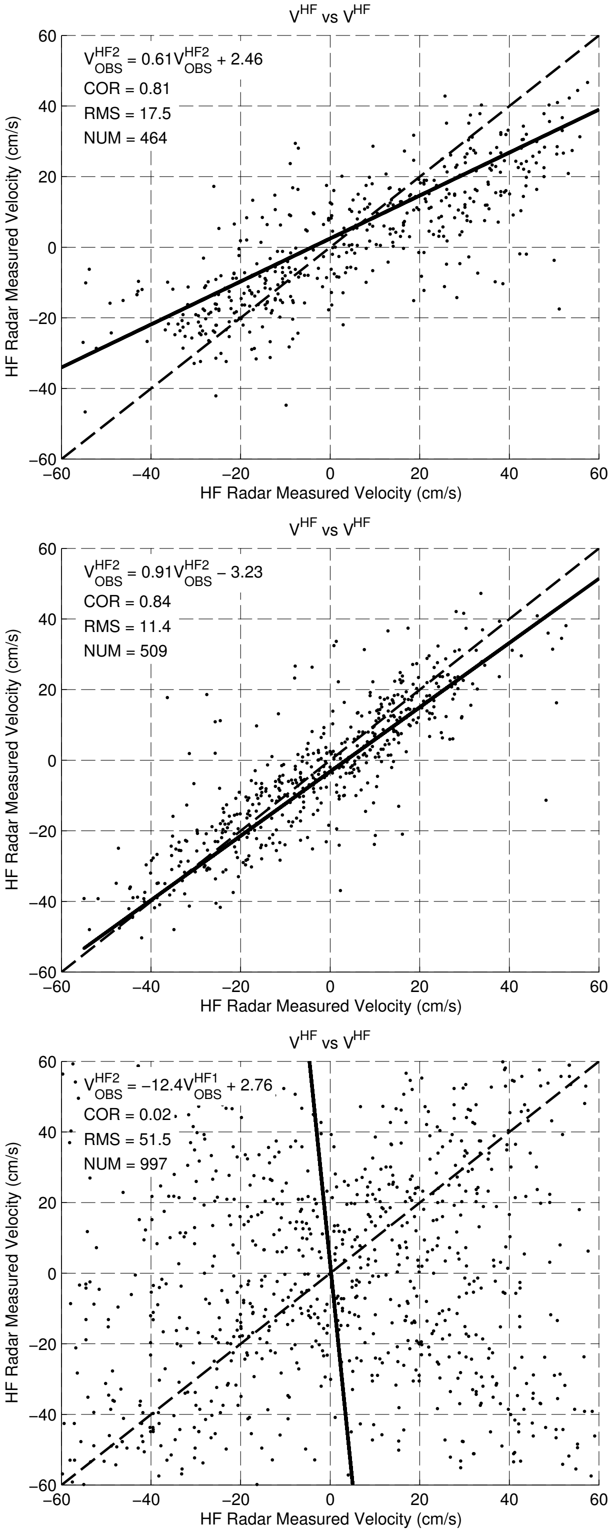

Fig. 2: Scatter plot of radial measurements from the three radar sites, SUNS–CPHN (top), SUNS–VIEW (middle) and CPHN–VIEW (bottom) (see Figure 1) along their baseline are presented. The solid line denotes the linear regression computed from Principal Component Analysis. Radial measurements returned at 30 minute intervals from the CPHN and VIEW stations while SUNS operated at 60 minute intervals.

This analysis highlights the inherent uncertainty present in HFR systems. In addition the CPHN–VIEW comparison demonstrates the additional complexities involved and one of the many factors that may impact on measurement accuracy of a remote sensing installation. The next section investigates this uncertainty further and discusses techniques to identify and eliminate these measurement errors.

B. PCA

First analysis of HFR data focused on a two month period June-July 2012. This time window was chosen since it was hypothesised that flows would be at their most stationary during this period avoiding both energetic winter storm events and high river outflows during spring ice melts. As common with sensor data percent coverage varies considerably over the course of the study period. Gaps in the data need to be accounted for prior to the application of PCA. Two approaches were adopted:

- Only data from grid cells that returned data $> 60\%$ of the time was used.
- Missing data in remainder of cells are interpolated from neighbouring grids using standard linear interpolation technique.

The PCA method was then applied to the data as described in section II.

Figure 3 presents the spatial patterns of the first three PCs for the time period June-July 2012 while Figure 4 displays the associated time expansion coefficients. Cumulatively, these 3 PCs account for 74% of the total variance. Mode I is the most dominant mode accounting for 54% with mode II and III accounting for 13% and 7% respectively.

The consistent direction of flows in PC_1 along with the high proportion of variance explained suggests it to be connected with tidal flows in the region. To investigate this hypothesis further, PC_1 was compared with an independent estimate of the tidal signal. To estimate the tidal signal, data from an Acoustic Doppler Current Profiler (ADCP) located in the Southern Region of the inner-Bay was used (red diamond in Figure 1). The ADCP data were processed via the t_tide software [17]; this decomposed the data into its harmonic (tidal) and residual component. In conjunction with this the HFR flow was reconstructed using PC_1 only from the grid cell nearest the ADCP location. Figure 6 presents time series plot comparing the two datasets. The tidal signal is clearly evident within the reconstructed data displaying close agreement with the extracted tidal signal.

It is reasonable to expect subsequent PCs to be closely related to wind forcing in the bay. Correlation coefficients between PC_2 and measured wind speeds from a weather station located at the Chesapeake Bay Bridge Tunnel (Figure 1) however, did not provide significant correlation. Computing a complex correlation coefficient [18] between the two vector time series (wind speed and flows reconstructed with PC_2 only) returned a correlation of 0.28 (where 0 indicates no correlation and 1 represents perfect agreement) with higher

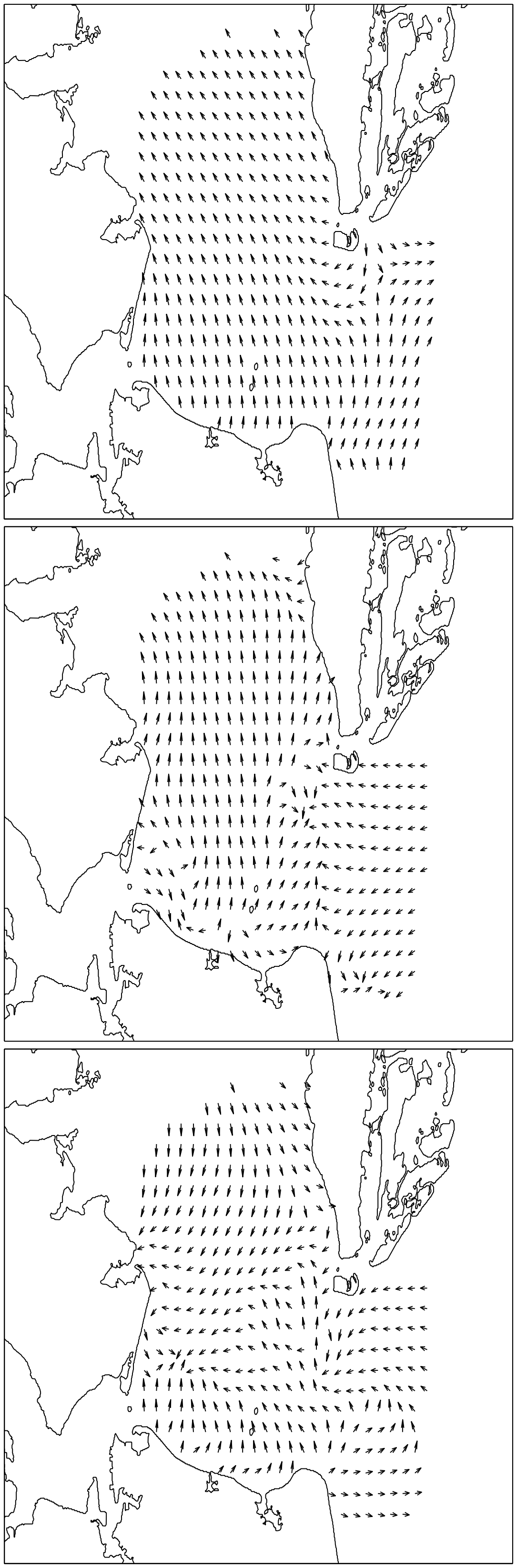

Fig. 3: PC spatial map patterns for modes I(top), II(middle) and III(bottom)

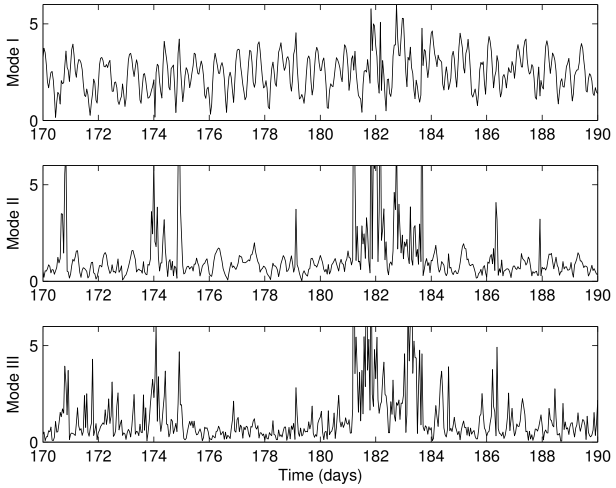

Fig. 4: First three principal component expansion coefficients computed for June-July period (20 day window presented for display purposes). The modal amplitudes are normalized by their respective standard deviations.

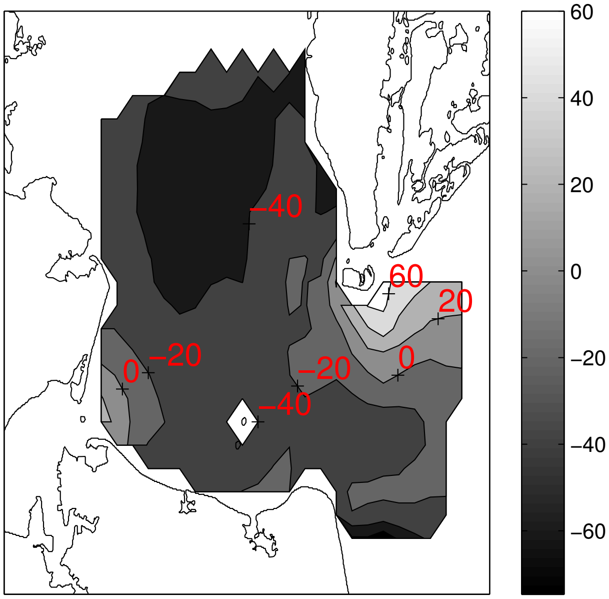

Fig. 5: Average counterclockwise angle derived from correlation computations between wind speed and PC_1 of the low pass filtered HFR dataset.

agreement observed in the North-South direction when investigating correlation independently in the zonal and meridional direction.

Analysis of the temporal evolution of the principal components (Figure 4) indicates this to be a result of the residual presence of tidal signal in this PC. To permit analysis of the signal distinct from the tidal component we returned to the original HFR data and low-pass filtered using a cosine-Lanczos filter with a 40-hr halfpower point [19] to remove the tidal signal from the data. Applying PCA to the filtered data

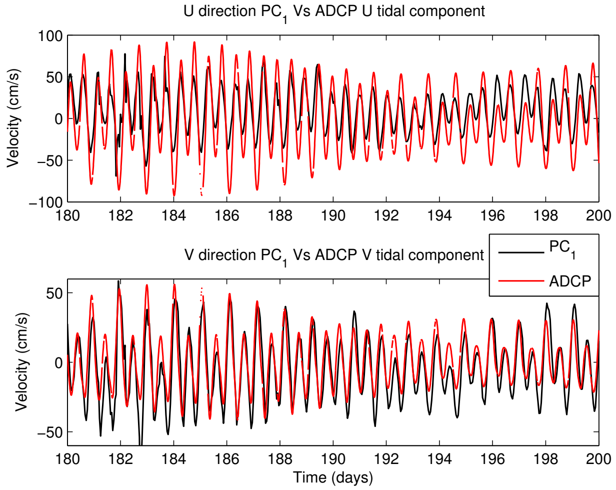

Fig. 6: Plotting flows reconstructed from the first principal component only against the estimated tidal signal in the bay. Flows reconstructed for the HFR grid cell closest to the ADCP probe. The tidal signal is computed by applying a harmonic analysis to the near-surface ADCP data from the Chesapeake Bay Bridge Tunnel.

gives insight into variability in the bay excluding the dominant tidal signal. The expectation in this case was that PC_1 would be primarily a result of wind effects. Recomputing complex correlation between PC_1 and measured wind speeds returned a value of 0.73 with this mode accounting for 50% of the total variance of the filtered data. The PC pattern associated with this mode is presented in Figure 7. The phase angle of the complex correlation coefficient, by definition, gives a measure of the average counterclockwise angle of the second vector (wind speed) with respect the first. Figure 5 presents the phase angle of correlation. Analysing the figure suggests reasonable agreement between angle of flows and wind forcing. In the outer bay, the angle is quite close to zero while in the inner bay the discrepancy is plausibly a result of topographical steering of the flow as it enters the bay and is directed Northwards into the bay.

The development of a PCA model that is representative of the raw data while excluding high frequency "noise" has two important considerations

- the number of PCs to include in the reconstruction
- the choice of temporal window width to which to apply the linear technique

The choice of number of PCs to retain is often times empirical and case specific. The simplest criterion is to retain enough PCs to represent a sufficient fraction of the total variance. Jolliffe [20] suggests the range of fractional variance between 0.7 and 0.9 may be a reasonable range. Applying total variance explained cut off points of 70, 90 and 95% results in retaining 2, 12, and 29 PCs respectively

Another subjective approach is based on the shape of the graph of the eigenvalues. The method looks for a "knee point" in the residual percent variance (RPV) plotted against the

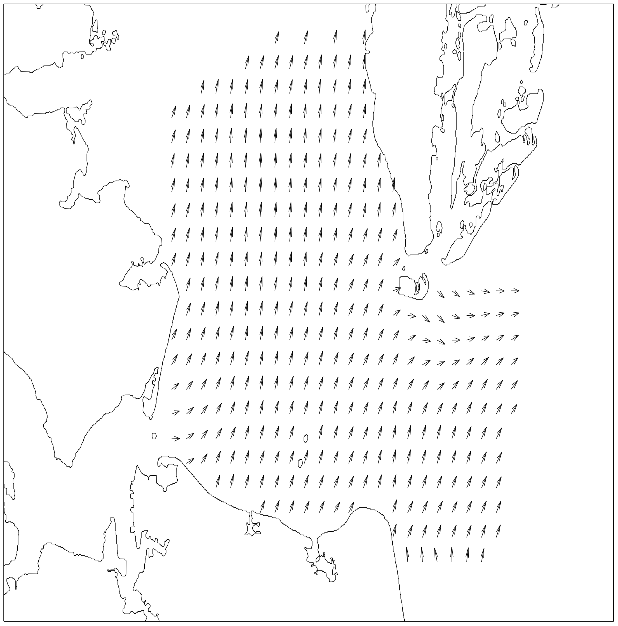

Fig. 7: PC spatial map patterns for modes PC_1 when the raw data is low-pass filtered prior to the application of PCA.

number of principal components. The method is based on the idea that the residual variance should reach a steady state when the factors begin to account for random errors. When a break point is found or when the plot stabilizes that corresponds to the number of principal components to represent the process. The RPV is computed based on residual eigenvalue:

$$RPV(k) = 100 \left[\frac{\sum\limits_{j=k+1}^{m} \lambda_j}{\sum\limits_{j=1}^{m} \lambda_j} \right] \% \tag{4}$$

Analysing graph of the RPV (not presented) suggests that steady state develops after 7 PCs.

An alternative criterion dictating which principal components to retain is the Guttman-Kaiser criterion [21]: Principal components associated with eigenvalues that are larger in magnitude than the average, $\overline{\lambda}$, of the eigenvalues or, better, a somewhat lower cut-off such as $\lambda^* = 0.7\overline{\lambda}$, are retained. Applying these criterion to this dataset would retain 20 and 25 of the principal components respectively. North et al. [22] argue that a set of principal components with similar eigenvalues should either be all retained or all excluded. The size of gaps between successive eigenvalues is thus an important consideration for any decision rule, and North et al. (1982) [22] provide a rule-of-thumb for deciding whether gaps are too small to split the principal components on either side. The rule states that if the sampling error of a particular eigenvalue $\lambda \left[\partial\lambda - \lambda \left(\frac{2}{N}\right)^{1/2}\right]$ is larger than the spacing between λ and a neighbouring eigenvalue, then the associated PCs will have comparable sampling errors. This implies that these eigenvectors are a random mixture of the true eigenvectors and could be excluded from the set. Applied to this data results in retention of 9 PCs.

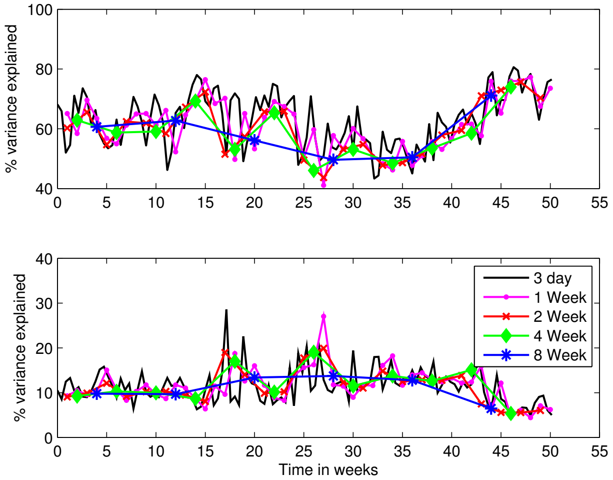

Fig. 8: Time evolution of fraction of variance explained by PC_1 (top) and PC_2 (bottom) for a range of window width. The window width used are of three days (72 time points), one, two, four and eight weeks.

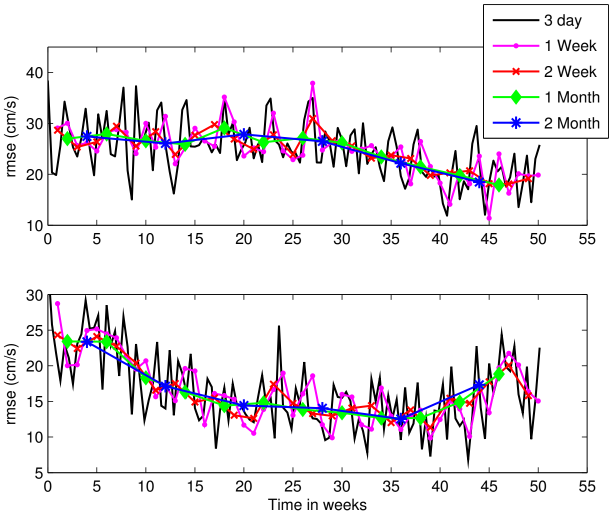

Fig. 9: RMSE computed between flows reconstructed from PC_1 (only) and the harmonic component of ADCP data for zonal (top) and meridional (bottom) components. The flows are reconstructed for the yearly dataset using five different PCA window widths of three days (72 time points), one, two, four and eight weeks.

The second point demanding attention is the window width of the PCA model. Up to now, we adopted a two month window and assumed the data had near-stationary mean and covariance structure for this time period. However, in such a dynamic system as ocean surface currents, this assumption is an area that requires further investigation.

To investigate how the process drifts with time we returned to the original one year dataset and applied PCA to the entire

TABLE I: Mean and standard deviation (σ) of RMSE computed between PC_1 and harmonic component of Cape Henry ADCP for a range of PCA window widths. Results are presented decomposed into their zonal and meridional components

Window Width	Zonal rmse	Merid rmse	Zonal σ	Merid. σ
1 day	25.04	16.96	7.84	6.91
3 day	24.71	16.99	5.57	7.20
1 week	24.74	16.46	4.89	4.47
2 week	24.72	16.48	3.72	3.82
1 month	24.89	16.41	3.48	3.81
2 month	24.75	16.46	3.68	3.85
3 month	24.67	16.53	3.75	3.56
6 month	24.77	16.07	2.02	3.55
1 year	26.61	16.49	-	-

year with a range of window widths, namely: one day, three day, and 1, 2, 4, 8, 12, 24, and 48 weeks.

Of interest was both the evolution in time of the PCs with different time windows and also the degree of compression provided by PCA as a function of time. As a preliminary step the degree of compression was investigated by evaluating how much of the total variance was explained by the first modes. Figure 8 presents the variance explained by different PCA models for the duration of the 48 week period. Analysing the figure indicates that while the 4 and 8 week sampling windows captures the general trend of the data, the linearity of the technique results in a considerable amount of information relevant to shorter time scales being neglected.

As a further measure of the amount of relevant information extracted by the different applications we returned to the information on the tidal signal gleaned from Figure 6. Considering that the information contained in the PC_1 is strongly correlated with tide, it is reasonable to associate the optimum compression of the data to that which best represents the tidal signal extracted from the ADCP. Again, the flows were reconstructed using PC_1 only, at the grid cell nearest the ADCP location at a range of window widths. To quantify performance, root-mean-square-error (RMSE), was computed between the reconstructed data and the tidal component and the progression in time analysed. Figure 9 plots the resultant differential.

As expected the general trend of the tidal signal is captured with large sampling times (two months). Table I presents the mean and standard deviation computed for the RMSE for the year. While the means are in very close agreement, there is considerable differences in standard deviation as would be expected from a visual inspection of Figure 9. Apparent is that with a high frequency sampling time, there are short periods when the RMSE is considerably higher. This may be a result of dynamicity present in the flow that cannot be captured by PC_1 or alternatively "noise" in the signal that a larger window width effectively averages out.

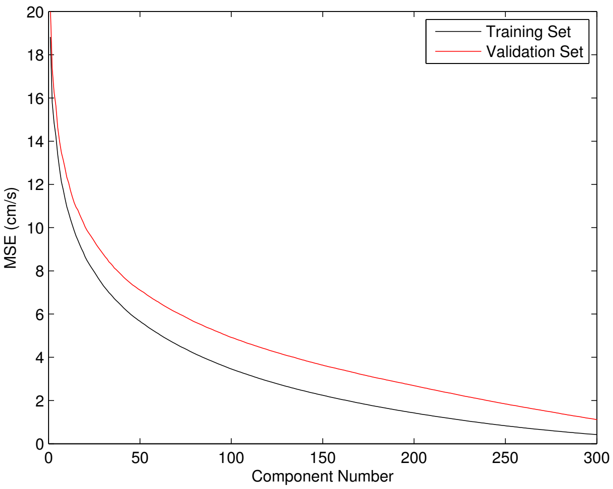

Fig. 10: MSE computed between reconstructed data for training and validation datasets

C. Application of model to validation set

Cross-validating the PCA model using new data is a means of providing further objective insight into PCA model performance. The basic idea of cross-validation is the use of different datasets for estimation and validation of each PC model [23]. For all applications the data was split into two equal time partitions: the training set used to construct the PCA model and the validation set to assess performance of the model with new data. PC models were determined using the training data and then evaluated on the validation data. The application of the method to new data involves making use of the scores of the PCA model. The scores of the model are the projections of the samples in the new coordinate system defined by the PCs. Projecting the validation data X_{val} onto the same PCs gives a reconstruction of the validation dataset $\tilde{X}_{val} = X_{val}P^T P^T$ which can be used to monitor changes in the system. The skill of the model (as function of window width and mode truncation) was evaluated with regards to optimum model selection. The skill of the model in returning the raw data can be represented by the mean squared reconstruction error (MSE) defined as:

$$MSE = \frac{1}{nm}||X - \tilde{X}||_F^2 \tag{5}$$

where X is the raw data, $\tilde{X}$ is the data reconstructed from PCA, n, m the dimensions of the matrix and $||X||_F$ is the Frobenius (or matrix) norm.

Figure 10 presents a comparison of the MSE computed for both the training set and the validation set. Apparent is the equivalent trend evident in both training and validation data MSE. This suggests that the signal of the HFR contains such similarities that prevent a simple decomposition of the noise from the distinct signal. It also does not provide any useful insight into the number of components required to describe the process. To further the usefulness of the PCA model in noise reduction a choice on number of PCs to retain must be

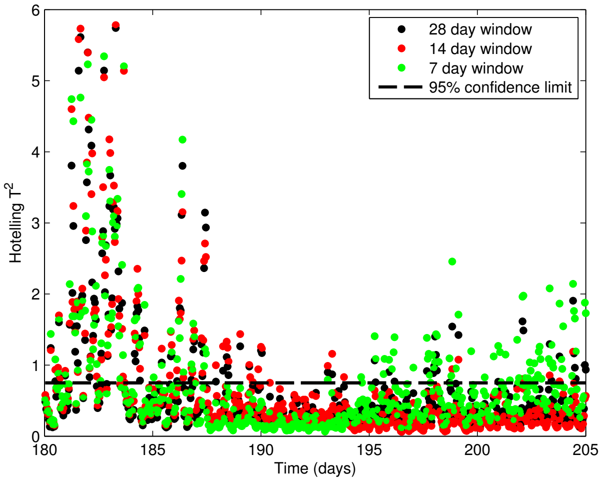

Fig. 11: Hotelling's T^2 statistic computed for a validation set of 28 days. The PCA model was computed using a window width of 7, 14 and 28 days and the validation set reconstructed. For the 7 and 14 day window widths, the PCA model was applied repeatedly using the previous dataset to best capture the evolution of the mean of the dataset. The dashed line represents the computed 95% confidence limit above which the dataset is considered an "outlier"

made. Considering the similarities with other cutoff choices and to permit for automated applications, the Guttman-Kaiser criterion [21] discussed earlier, that retains all eigenvalues, q, where $\lambda > 0.7\overline{\lambda}$ was adopted. The validation set was then reconstructed as $\tilde{X}_{val} = X_{val}P_q^TP_q^T$

To provide further insight into outliers in the dataset and their origins, Hotelling's T^2-test which is a multivariate representation of Student's t-test is adopted. It gives a measure of the variation *not* captured by the model and can be expressed as:

$$T_i^2 = t_i^T\Lambda^{-1}t_i \tag{6}$$

where $\lambda = diag(\lambda_1, \lambda_2...\lambda_k)$ are PC eigenvalues. A range of PCA models was constructed using different window widths as described earlier and deviations from the model computed using the T^2 measure. A multivariate process is considered to be anomalous at the i_{th} sampling time if T_i^2 exceeds an upper control limit. A limit for the 95% confidence level can be expressed as:

$$T_{lim}^2 = \frac{K(N-1)}{N-K}F(K, N-K, \alpha) \tag{7}$$

where$F(K, N-K, \alpha)$ corresponds to the probability point on the F-distribution with (K,N-K) degrees of freedom and confidence level α, K is the number of principal components, and N is the number of observations.

Figure 11 presents the Hotelling's T^2 statistic computed for validation set of 28 days. The PCA model was computed using a window width of 7, 14 and 28 days and the validation set reconstructed. For the 7 and 14 day window widths, the PCA model was applied repeatedly using the previous dataset to best capture the evolution of the mean of the dataset. The 95% confidence limit is also denoted. Apparent is the considerable number of returns that exceed the computed confidence intervals. In itself, the metric is of limited value as it only provides a measure for the entire dataset at each time return. To provide meaningful insight, the contribution of each HFR grid cell to the total is more practical. A spatial representation of the contribution can be computed as

$$t_{con,i} = t_i\lambda_i^{-1/2}P_k^T \tag{8}$$

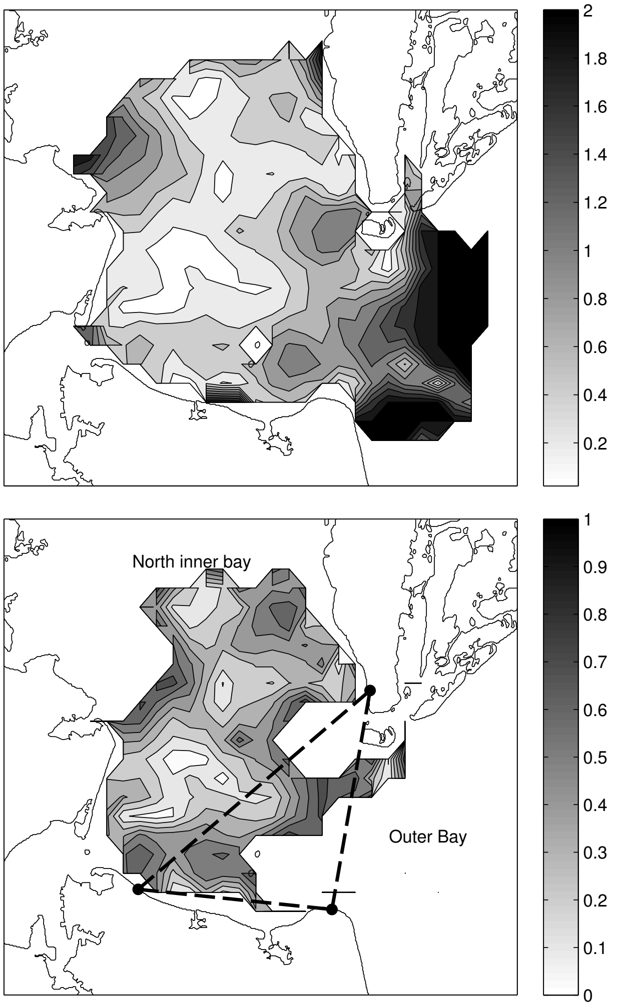

Fig. 12: Spatial maps of T2 contributions to (a) total and (b) eliminating cells that exceed limit until confidence interval met corresponding to HFR measurements for julian day 181 at 05:00am (selected due high volume of anomalous returns for representative purposes)

From Figure 11, it is apparent that the earlier portion of the time window contains a number of points that exceed the confidence limit by multiple orders of magnitude. For illustration purposes we adopted the time return that corresponds to the largest T^2 value (day 181 at 05:00am); i.e. the time when the model performs poorest in capturing the variation of the data.

Figure 12 presents the spatial contribution of each grid cell to the T2 score for this time.

It is evident that a region in the outer bay contributes a large proportion of the total variation computed. The methodology adopted for this study is the iterative elimination of cells with maximum T^2 contribution until the confidence interval of the dataset is met. Figure 12b presents the spatial map of T^2 after the data is processed as described. For this particular case, the elimination of outlier data reduces the number of return by 40%. Analysis of the map of processed data suggests that the data identified by the PCA model as being anomalous is physically meaningful. Known issues exist regarding the performance of the HFR in the outer Bay. The SUNS station does not return radial measurements in this region due to no direct over water line of sight (see Figure 1). Hence, the meridional component of velocity is not well resolved in the region resulting in an uncertain reconstruction of the flow. Monitoring the incoming data with process control metrics identify and eliminate measurements from this region. Other areas with potential issues that are successfully identified include the North inner Bay where distance from radar station plausibly impacts performance and beside the headland in the Northern Bay where the signal is distorted by the nearby land.

It is important to note that this data return represents the most extreme outlier of all the data analysed. Hence, the exclusion ratio of 40% can be considered a worst case scenario. It also requires stressing that no pre-processing of the data was conducted prior to analysis. Typically in HFR applications, the data is pre-processed to eliminate particular cells based on known performance issues such as low signal-to-noise ratio, areas of high geometric dilution of precision [9], extreme distance from radar measuring site, etc. No pre-processing was performed in this study as the goal was an objective analysis that would identify and eliminate outlier data in an automated manner.

IV. Conclusions

This paper presents the application of multivariate process control techniques to the analysis of surface current flows collected by HFR system in the Chesapeake Bay area. To better understand flows in the region. PCA de-constructs the data based on the amount of variance present. Analysis shows that this data-driven approach inherently links measured flows to physical processes in the bay. The decomposition into distinct spatial and temporal patterns serves as a means to better understand and describe flow patterns and further relate synoptic patterns to local environmental variables. It also supports the viability of adopting PCA to partition the physically driven signal present in the HFR measurement from underlying noise.

Application of the technique to the validation dataset correctly identifies area that have known performance issues. In this study we chose to remove these cells thereby reducing the measurement area; an alternative option is to filter these anomalous cells by truncating the reconstruction at fewer PCs or applying a weighting coefficient to reflect the increased uncertainty of these cells.

The research also highlights challenges in the application of PCA to HFR data that requires further investigation. The high spatial and temporal variability of the data makes a distinct decomposition of flows into uncorrelated variables in space and time difficult. The relative close proximity in time of the measurements (hourly) imply that there is likely to be correlation between measurements at adjacent time points, resulting in non-independence between observations. Several techniques exist that take account of correlation between observations such as Singular Spectrum Analysis (SSA) or frequency domain PCA [20]. Future work will focus on a more detailed investigation of these relationships and combination with multivariate process control metrics.

References

[1] M. Yaremchuk and A. Sentchev, "A combined EOF\variational approach for mapping radar-derived sea surface currents," Stennis Space Center, MS, 39529, Naval Research Laboratory, Tech. Rep., 2011.

[2] D. Prandle, S. G. Loch, and R. Player, "Tidal flow through the Straits of Dover," *J. Phys. Oceanogr.*, vol. 23, no. 1, 1993.

[3] J. T. Kohut, H. J. Roarty, and S. M. Glenn, "Characterizing observed environmental variability with HF radar surface current mappers and acoustic Doppler current profilers: Environmental variability in the coastal ocean," *IEEE J. Oceanic. Eng.*, vol. 31, no. 4, pp. 876–884, Oct. 2006.

[4] D. S. Ullman, J. O'Donnell, J. Kohut, T. Fake, and A. Allen, "Trajectory prediction using HF radar surface currents: Monte Carlo simulations of prediction uncertainties," *J. Geophys. Res.*, vol. 111, p. 14 PP., Dec. 2006.

[5] J. Kohut, H. Roarty, S. Licthenwalner, S. Glenn, D. Barrick, B. Lipa, and A. Allen, "Surface current and wave validation of a nested regional HF radar network in the Mid-Atlantic bight," in *IEEE/OES 9th Working Conference on Current Measurement Technology, 2008. CMTC 2008.* IEEE, Mar. 2008, pp. 203–207.

[6] B. Lipa, C. Whelan, B. Rector, and B. Nyden, "HF radar bistatic measurement of surface current velocities: Drifter comparisons and radar consistency checks," *Remote Sensing*, vol. 1, no. 4, pp. 1190–1211, Dec. 2009.

[7] M. Yaremchuk and A. Sentchev, "Mapping radar-derived sea surface currents with a variational method," *Cont. Shelf Res.*, vol. 29, no. 14, pp. 1711–1722, Jul. 2009.

[8] J. D. Paduan and L. Washburn, "High-frequency radar observations of ocean surface currents," *Annual Review of Marine Science*, vol. 5, pp. 115–136, 2013.

[9] B. M. Emery, L. Washburn, and J. A. Harlan, "Evaluating radial current measurements from CODAR high-frequency radars with moored current meters," *J. Atmos. Ocean. Tech.*, vol. 21, pp. 1259–1271, 2004.

[10] H. H. Essen, K. Gurgel, and T. Schlick, "On the accuracy of current measurements by means of HF radar," *IEEE J. Oceanic. Eng.*, vol. 25, no. 4, pp. 472–480, 2000.

[11] B. Lipa, B. Nyden, D. S. Ullman, and E. Terrill, "Seasonde radial velocities: derivation and internal consistency," *Oceanic Engineering, IEEE Journal of*, vol. 31, no. 4, pp. 850–861, 2006.

[12] B. Lipa and D. Barrick, "Least-squares methods for the extraction of surface currents from codar crossed-loop data: Application at arsloe," *Oceanic Engineering, IEEE Journal of*, vol. 8, no. 4, pp. 226–253, 1983.

[13] J. Harlan, A. Allen, E. Howlett, E. Terrill, S. Kim, M. Otero, S. Glenn, H. Roarty, J. Kohut, J. O'Donnell *et al.*, "National ioos high frequency radar search and rescue project," in *OCEANS 2011.* IEEE, 2011, pp. 1–9.

[14] B. Lipa, "Uncertainties in seasonde current velocities," in *Current Measurement Technology, 2003. Proceedings of the IEEE/OES Seventh Working Conference on.* IEEE, 2003, pp. 95–100.

[15] Y. Yoshikawa, A. Masuda, K. Marubayashi, M. Ishibashi, and A. Okuno, "On the accuracy of hf radar measurement in the tsushima strait," *Journal of Geophysical Research: Oceans (1978–2012)*, vol. 111, no. C4, 2006.

[16] J. D. Paduan, K. C. Kim, M. S. Cook, and F. P. Chavez, “Calibration and validation of direction-finding high-frequency radar ocean surface current observations,” *Oceanic Engineering, IEEE Journal of*, vol. 31, no. 4, pp. 862–875, 2006.

[17] R. Pawlowicz, B. Beardsley, and S. Lentz, “Classical tidal harmonic analysis including error estimates in matlab using t_tide,” *Comput. Geosci.*, vol. 28, no. 8, pp. 929–937, 2002.

[18] P. Kundu, “Ekman veering observed near the ocean bottom,” *J. Phys. Oceanogr.*, vol. 6, pp. 238–242, 1976.

[19] W. J. Emery and R. E. Thomson, *Data analysis methods in physical oceanography*. Elsevier Science, 2001.

[20] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2005.

[21] J. E. Jackson, *A user’s guide to principal components*. John Wiley & Sons, 2005, vol. 587.

[22] G. R. North, T. L. Bell, R. F. Cahalan, and F. J. Moeng, “Sampling errors in the estimation of empirical orthogonal functions,” *Monthly Weather Review*, vol. 110, no. 7, pp. 699–706, 1982.

[23] G. Diana and C. Tommasi, “Cross-validation methods in principal component analysis: a comparison,” *Statistical Methods and Applications*, vol. 11, no. 1, pp. 71–82, 2002.