Using Machine Learning to Model Uncertainty for Water-Vapor Atmospheric Motion Vectors

Joaquim V. Teixeira[1], Hai Nguyen[1], Derek J. Posselt[1], Hui Su[1], Longtao Wu[1]

[1]Jet Propulsion Laboratory, California Institute of Technology

Abstract. Wind-tracking algorithms produce Atmospheric Motion Vectors (AMVs) by tracking clouds or water vapor across spatial-temporal fields. Thorough error characterization of wind-tracking algorithms is critical in properly assimilating AMVs into weather forecast models and climate reanalysis datasets. Uncertainty modelling should yield estimates of two key quantities of interest: bias, the systematic difference between a measurement and the true value, and standard error, a measure of variability of the measurement. The current process of specification of the errors in inverse modelling is often cursory and commonly consists of a mixture of model fidelity, expert knowledge, and need for expediency. The method presented in this paper supplements existing approaches to error specification by providing an error-characterization module that is purely data-driven and requires few tuning parameters. Our proposed error-characterization method combines the flexibility of machine learning (random forest) with the robust error estimates of unsupervised parametric clustering (using a Gaussian Mixture Model). Traditional techniques for uncertainty modeling through machine learning have focused on characterizing bias, but often struggle when estimating standard error. In contrast, model-based approaches such as k-means or Gaussian mixture modelling can provide reasonable estimates of both bias and standard error, but they are often limited in complexity due to reliance on linear or Gaussian assumptions. In this paper, a methodology is developed and applied to characterize error in tracked-wind using a high-resolution global model simulation, and it is shown to adequately capture the error features of the tracked wind.

1. Introduction

Reliable estimates of global winds are critical to science and application areas, including global chemical transport modeling and numerical weather prediction. One source of wind measurements consists of feature-tracking based Atmospheric Motion Vectors (AMVs), produced by tracking time sequences of satellite-based measurements of clouds or spatially distributed water vapor fields (Mueller et al., 2017; Posselt et al., 2019). The importance of global measurements of 3-dimensional winds was highlighted as an urgent need in the NASA Weather Research Community Workshop Report (Zeng et al., 2016) and was identified as a priority in the 2007 National Academy of Sciences Earth Science and Applications from Space (ESAS 2007) Decadal Survey and again in ESAS 2017. For instance, wind is used in the study of global CO_2 transport (Kawa et al., 2004), numerical weather prediction (NWP; Cassola and Burlando, 2012), as inputs into weather and climate reanalysis studies (Swail and Cox, 2000), and for estimating current and future wind-power outputs (Staffell and Pfenninger, 2016).

Thorough error characterization of wind-track algorithms is critical in properly assimilating AMVs into forecast models. Prior literature has explored the impact of 'poor' error-characterization in Bayesian-based approaches to

34 remote sensing applications. Nguyen et al. (2019) proved analytically that when the input bias is incorrect in Bayesian

35 methods (specifically, optimal estimation retrievals), then the posterior estimates would also be biased. Moreover,

36 they proved that when the input standard error is 'correct' (that is, it is as close to the unknown truth as possible), then

37 the resulting Bayesian estimate is 'efficient'; that is, it has the smallest error among all possible choices of prior

38 standard error. Additionally, multiple active and passive technologies are being developed to measure 3D winds, such

39 as Doppler wind lidar (DWL), radar, and infrared/microwave sensors that derive AMVs using feature-tracking of

40 consecutive images. Therefore, an accurate and robust methodology for modeling uncertainty will allow for more

41 accurate assessments of mission impacts, and the eventual propagation of data uncertainties for these instruments.

42 Velden and Bedka (2009) and Salonen et al. (2015) have shown that height assignment contributes a large component

43 of uncertainty in AMVs tracked from cloud movement and from sequences of infrared satellite radiance images.

44 However, with AMVs obtained from water vapor profiling instruments (e.g., infrared and microwave sounders),

45 height assignment error cannot be directly assessed purely through analysis of the AMV extraction algorithm. Height

46 assignment is instead an uncertainty in the water vapor profile itself. Unfortunately, without the quantified

47 uncertainties on the water vapor profile necessary to pursue such a study, that is well beyond the scope of this paper.

48 As such, this study will focus on errors in the AMV estimates at a given height. Previous work has demonstrated

49 several different approaches for characterizing AMV vector error. One common approach is to employ quality

50 indicator thresholds, as described by Holmund et al (2001), which compare changes in AMV estimates between

51 sequential timesteps and neighboring pixels, as well as differences from model predictions, to produce a quality

52 indicator to which a discrete uncertainty is assigned. The Expected Error approach, developed by Le Marshal et al.

53 (2004), builds a statistical model using linear regression against AMV-radiosonde values to estimate the statistical

54 characteristics of AMV observation error.

55 In this study, we outline a data-driven approach for building an AMV uncertainty model using observing system

56 simulation experiment (OSSE) data. We build on the work by Posselt et al. (2019) in which a water vapor feature-

57 tracking AMV algorithm was applied to a high-resolution numerical simulation, thus providing a global set of AMV

58 estimates which can be compared to the reference winds produced by the simulation. In this case, a synthetic "true"

59 state is available with which AMVs can be compared and errors are quantified, and it is shown that errors in AMV

60 estimates are state dependent. Our approach will use a conjunction of machine learning (random forest) and

61 unsupervised parametric clustering (Gaussian mixture models) to build a model for the uncertainty structures found

62 by Posselt et al. (2019). The realism and robustness of the resulting uncertainty estimates depend on the realism and

63 representativeness of the reference dataset. This work builds upon the work of Bormann et al. (2014) and Hernandez-

64 Carrascal and Bormann (2014), who showed that wind tracking could be divided into distinct geophysical regimes by

65 clustering based on cloud conditions. This study supplements that approach with the addition of machine learning,

66 which, compared with traditional linear modeling approaches, should allow the model to capture more complex non-

67 linear processes in the error function.

68 Traditional techniques for modeling uncertainty through machine learning have focused on characterizing bias but

69 often struggle when estimating standard error. By pairing a random forest algorithm with unsupervised parametric

70 clustering, we propose a data-driven, cluster-based approach for quantifying both bias and standard error from

71 experimental data. According to the theory developed by Nguyen et al. (2019), these improved error characterizations

72 should then lead to improved error characteristics (e.g., lower bias, more accurate uncertainties) in subsequent analyses

73 such as flux inversion or data assimilation.

74 This paper does not purport that the specific algorithm detailed here should supplant error characterization approaches

75 for all AMVs; indeed, most commonly assimilated AMVs are based on tracking cloud features, not water vapor

76 profiles. In addition, this algorithm is trained and developed for a specific set of AMVs extracted from a water vapor

77 field associated with a particular range of flow features. As such, application of our algorithm to modeled or observed

78 AMVs will be most appropriate in situations with similar dynamics to our training set. However, we intend in this

79 paper to demonstrate that the methodology is successful in characterizing errors for this set of water vapor AMVs and

80 suggest that this approach— that is, capturing state-dependent uncertainties in feature-tracking algorithms through a

81 combination of clustering and random forest— could be implemented in other feature-tracking AMV extraction

82 methods and situations.

83 The rest of the paper is organized as follows: In Section 2, we give an overview of the simulation which provides the

84 training data for our machine learning approach. We then motivate and define the specific uncertainties this study

85 aims to characterize. In Section 3, we describe the error characterization approach with the specifics of our error

86 characterization model, including both the implementation of and motivations for employing the random forest and

87 Gaussian mixture model. In Section 4, we provide a validation of our methods, attempting to assess the bias of our

88 predictions. In Section 5, we discuss the implications of our error characterization approach, both on AMV estimation

89 and data assimilation more broadly.

90 **2. Experimental Set-up**

91 **2.1 Simulation and Feature-Tracking Algorithm**

92 We trained our model on the simulated data used by Posselt et al. (2019), which applied an AMV algorithm to outputs

93 from the NASA Goddard Space Flight Center (GSFC) Global Modeling and Assimilation Office (GMAO) GEOS-5

94 Nature Run (G5NR; Putman et al. 2014). The Nature Run is a global dataset with ~7 km horizontal grid spacing that

95 includes, among other quantities, three-dimensional fields of wind, water vapor concentration, clouds, and

96 temperature. Note that throughout the text we will use the term 'Nature Run wind' to refer to reference winds in the

97 simulation dataset used to train the uncertainty model. The AMV algorithm is applied on four pressure levels (300hPa,

98 500hPa, 700hPa, and 850hPa) at 6-hourly intervals, using three consecutive global water vapor fields spaced one hour

99 apart, and for a 60-day period from 07/01/2006 to 08/30/2006. The water-vapor fields from GEOS5 were input to a

100 local-area pattern matching algorithm that approximates wind speed and direction from movement of the matched

101 patterns. The algorithm searches a pre-set number of nearby pixels to minimize the sum-of-absolute-differences
102 between aggregated water vapor values across the pixels. Posselt et al. (2019) describes the sensitivity of the tracking
103 algorithm and the dependency of the tracked winds on atmospheric states in detail. The coordinates of the data are on
104 a 5758 x 2879 x 240 spatio-temporal grid for the longitude, latitude, and time dimension, respectively.

105 It is important to note that the AMV algorithm tracks water vapor on fixed pressure levels. In practice, these would be
106 provided by satellite measurements, whereas in this paper we use simulated water vapor from the GEOS-5 Nature
107 Run. In this simulation height assignment of the AMVs is assumed to be perfectly known. This assumption is far from
108 guaranteed in real world applications but, as previously discussed, its implications are not pursued in this paper. As
109 such, we focus solely on observational AMV error and not on height assignment error. We note that in practice, one
110 approach to understanding the behavior and accuracy of the wind-tracking algorithm is to apply it to modeled data
111 (e.g., Posselt et al., 2019). Our approach seeks to complement this approach by providing a machine-
112 learning/clustering hybrid approach that can further divide comparison domains into 'regimes' which may provide
113 further insights into the behavior of the errors and/or feedback into the wind-tracking algorithm.

114 A snapshot of the dataset at 700hPa is given in Figure 1, where we display the water vapor from Nature Run (top left
115 panel), the wind speed from Nature Run (top right panel), the tracked wind from the AMV-tracking algorithm (bottom
116 right panel), and the difference between the Nature Run and tracked wind (bottom left panel). Note that the wind-
117 tracking algorithm tends to have trouble in region where the Nature Run water vapor content is close to zero. It is clear
118 that while the wind-tracking algorithm tends to perform well in most regions (we can classify these regions as areas
119 where the algorithm is skilled), in some regions the algorithm is unable to reliably make a reasonable estimate of the
120 wind speed (unskilled). We will examine these skilled and unskilled regimes (and their corresponding contributing
121 factors) in section 3.

122 **2.2 Importance of Uncertainty Representation in Data Assimilation**

123 Proper error characterization for any measurement, including AMVs, is important in data assimilation. Data
124 assimilation often uses a regularized matrix inverse method based on Bayes' theorem, which, when all probability
125 distributions in Bayes' relationship are assumed to be Gaussian, reduces to minimizing a least-squares (quadratic) cost
126 function Eq (1):

127
$$\mathbf{J} = (\mathbf{x} - \mathbf{x_b})\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x_b}) + \left((\hat{\mathbf{y}} - \mathbf{a}) - \mathbf{H}[\mathbf{x}]\right)^{\mathrm{T}}\mathbf{R}^{-1}\left((\hat{\mathbf{y}} - \mathbf{a}) - \mathbf{H}[\mathbf{x}]\right) \tag{1}$$

128 where $\mathbf{x}$ represents the analysis value, $\mathbf{x_b}$ represents the background field (first guess), $\mathbf{B}$ represents the background
129 error covariance, $\mathbf{y}$ represents the observation, and $\mathbf{H}$ represents the forward operator that translates model space into
130 observation space. This translation may consist of spatial and/or temporal interpolation if $\mathbf{x}$ and $\mathbf{y}$ are the same variable
131 (e.g., if the observation of temperature comes from a radiosonde), or may be far more complicated (e.g., a radiative
132 transfer model in the case of satellite observations). $\mathbf{R}$ represents the observation error covariance, and $\mathbf{a}$ represents

the accuracy, or bias, in the observations. The right hand side of Eq. (1) can be interpreted as a sum of the contribution of information from the data ($\mathbf{y} - \mathbf{H[x]} - \mathbf{a}$) and the contribution from the prior ($\mathbf{x} - \mathbf{x_b}$), which are weighted by their respective covariance matrices. In our analysis, the AMVs obtained from the wind-tracking algorithm is used as 'data' in subsequent analysis. That is, the tracked wind data $\hat{\mathbf{y}}$ is a biased and noisy estimator of the true wind $\mathbf{y}$, and might be assumed to follow the model Eq. (2):

$$\hat{y} = y + \epsilon \qquad (2)$$

where ϵ is an error term, commonly assumed to be Gaussian with mean $\mathbf{a}$ and covariance matrix $\mathbf{R}$ (i.e., $\epsilon \sim N(\mathbf{a}, \mathbf{R})$), which are the same two terms that appear in Equation (1). As such, for data assimilation to function, it is essential to correctly specify the bias vector $\mathbf{a}$ and the standard error matrix $\mathbf{R}$. Incorrect characterizations of either of these components could have adverse consequences on the resulting data assimilation analyses with respect to bias and/or the standard error (Nguyen et al., 2019).

3 Methodology

3.1 Generalized Error Characterization Model

An overview of our approach is outlined in Figure 2. Given a set of training predictors X, training responses $\hat{Y}$, and simulated true response Y, our approach begins with two independent steps. In one step, a Gaussian mixture model is trained on the set of X, $\hat{Y}$, and Y. This clustering algorithm identifies geophysical regimes where the nonlinear relationships between the three variables differ. In the other step, a random forest is used to model Y based on X and $\hat{Y}$. This step produces an estimate of the true response (we call this $\tilde{Y}$) using only the training predictors and response. We then employ the Gaussian mixture model to estimate the clusters which the set of X, $\hat{Y}$, and $\tilde{Y}$ pertain to. Subsequently, we compute the error characteristics of each cluster of X, $\hat{Y}$, and $\tilde{Y}$ in the training dataset. Thereafter, given a new point consisting solely of X and $\hat{Y}$, we can assign it to a specific cluster and ascribe to it a set of error characteristics.

In this paper, we are primarily interested in the distribution of a retrieved quantity versus the truth. That is, given a retrieved value $\hat{Y}_i$, we are interested in the first and second moments (i.e., $E(\hat{Y}_i - Y)$ and $var(\hat{Y}_i - Y)$), respectively. We note that there is a large body of existing work on uncertainty modeling in the machine learning literature (e.g., Coulston et al., 2016; Tripathy et al., 2018; Tran et al., 2019; Kwon et al., 2020), although these approaches primarily define the uncertainty of a prediction as $var(\hat{Y}_i)$, or quantify how sensitive that prediction is to tiny changes in the models/inputs. Our approach, on the other hand, characterizes the error as $var(\hat{Y}_i - Y)$, which describes how accurate a prediction is relative to the *true value*. For this reason, our methodology is more stringent in that it requires knowledge of the true field (which comes naturally within OSSE framework) or some proxies such as independent validation data or reanalysis data. In return, the error estimates from our methodology fit naturally within the data assimilation framework (that is, it constitutes the parameter R in Eq. (1)).

165 What follows in this paper is an implementation of the error characterization model obtained for a subsample of the
166 GEOS-5 Nature Run at a fixed height of 700hPa. In particular, we trained the error characterization on a random
167 subsample from the first 1.5 months of the Nature Run, and show the results obtained when applying it to a test
168 subsample drawn from the subsequent 0.5 months of the Nature Run.

3.2 Error Regime

170 When examining the relationship between AMVs and Nature Run winds in Figure 3, it is clear that there are two
171 distinct 'error-regimes' present in the dataset. The majority of AMV estimates can be categorized as 'skilled', wherein
172 their estimate lies clearly along a one-to-one line with the Nature Run wind. However, there is also clearly an
173 'unskilled' regime, for which the AMV estimate is very close to zero when there are actually moderate or large Nature
174 Run wind values present. Our goal is to provide unique error characterizations for each error regime, because the error
175 dynamics are different within each regime. Furthermore, when we analyze this error and its relationship to water
176 vapor, we see that 'unskilled' regime correlates highly with areas of low water vapor in Figure 4. This matches the
177 error patterns discussed in Posselt et al. (2019). We note that the division between skilled and unskilled regimes does
178 not need to be binary. For instance, in some regions the wind-tracking algorithm might be unbiased with high-
179 correlation with the true winds, and in other regions the algorithm might still be unbiased relative to the true winds,
180 but with higher errors. The second situation is clearly less skilled than the first, although it might still be considered
181 'skilled', and this separation of the wind-tracking estimates into various 'grades' of skill forms the basis of our error
182 model.

3.3 Gaussian Mixture Model

184 These distinct regimes present an opportunity to employ machine learning. Bormann et al. (2014) and Hernandez-
185 Carrascal and Bormann (2014) demonstrated that cluster (also called regime) analysis is a successful approach for
186 wind-tracking error characterization, and so we aim to train a clustering algorithm that will cluster a given individual
187 AMV estimate to various 'grades' of skill. In particular, we use a clustering algorithm that can take advantage of the
188 underlying geophysical dynamics. To this end, we employ a Gaussian mixture model, an unsupervised clustering
189 algorithm based on estimating a training set as a mixture of multiple Gaussian distributions. A mathematical overview
190 follows:

191   1.   Define each location containing Nature Run winds, water vapor, and AMV estimates as a random variable
192   x_i
193   2.   Define θ as the population that consists of all x_i in the training dataset
194   3.   Model the distribution of the population $P(\theta)$ as:

$$P(\theta) = \sum_{j}^{K} \pi_j N(\mu_j, \Sigma_j) \tag{3}$$

196   Where $N(\mu_j, \Sigma_j)$ is the normal distribution with mean μ_j and covariance Σ_j of the *j*-th cluster,

197   K is the number of clusters, and π_j is the mixture proportion.

198   4.   Determine π_j, μ_j, Σ_j for K clusters using the Expectation–Maximization Algorithm

199   5.   From 3. and 4., estimate the probability of a given x_i belonging to the j-th cluster as $P(x_i \in k_j) = p_{ij}$

200   6.   Assign point x_i to the cluster with the maximum probability p_{ij}

201   The mixture model clustering is based on the R package 'Mclust' developed by Fraley et al. (2012), which builds upon
202   the theoretical work of Fraley and Raftery (2002) for model-based clustering and density estimation. The process uses
203   an Expectation-Maximization algorithm to cluster the dataset, estimating a variable number of distinct multivariate
204   Gaussian distributions from a sample dataset. Training the Gaussian mixture model on this dataset provides a
205   clustering function which outputs a unique cluster for any data point with the same number of variables.

206   In one dimension, a Gaussian mixture model looks like the distributions depicted in Figure 5: instead of modeling a
207   population as a single distribution (Gaussian or otherwise), the GMM algorithm fits multiple Gaussian distributions
208   to a population. One key aspect of this algorithm is the capability of assigning a new point to the most likely
209   distribution. For example, in the 1-D figure, a normalized AMV estimate with a value of 10 would be more likely to
210   originate from the broad cluster '2' than the narrow cluster '4'. In this case, we model the population as a Gaussian
211   mixture model in five-dimensional space, which consists of two Nature Run wind vector components (u and v), two
212   AMV estimates of these wind components (û and v̂), and the simulated water vapor values, all of which have been
213   standardized to have mean 0 and standard deviation of 1. Each cluster has a 5-dimensional mean vector for the center
214   and a 5x5 covariance matrix defining their multivariate Gaussian shape. The estimation of a covariance matrix allows
215   for the characterization of the relationships between the different dimensions within each cluster, and as such the
216   gaussian mixture model approach provides greater potential for understanding the geophysical basis of error regimes
217   than other unsupervised clustering approaches.

218   We note that the choice of inputs to the clustering methodology is limited, and that a more successful clustering may
219   be achieved by including additional meteorological or geographic information. However, the intention of this paper
220   is to study the ability of a purely data-driven approach, where no additional information or assumptions are passed to
221   the machine learning model outside of the inputs and outputs to the AMV algorithm itself. Posselt et al. (2019) showed
222   that state dependent uncertainties are a major source of error in water vapor AMVs; introducing further information
223   may cloud our ability to discern these specific uncertainties. While scaling this methodology to other applications may
224   incentivize tailoring to specific conditions, this paper aims to demonstrate that modifications are encouraged for
225   improvement, but not necessary for success.

226   Having trained the Gaussian mixture model on the 1.5 month training dataset, we applied the clustering algorithm to
227   a testing dataset sampled from the subsequent 0.5 months of the nature run. By re-analyzing the AMV estimate in
228   relation to the Nature Run winds within each cluster (), we find that the clustering approach successfully separates the

229 AMV estimates according to their 'skillfulness'. Essentially, we repeat Figure 3 but divide the AMV estimates by
230 cluster. We see that, for example, clusters 4, 5, and 7 clearly represent cases in which the feature-tracking algorithm
231 provides an accurate estimate of the Nature Run winds, with very low variance around the one-to-one line (i.e., low
232 estimation errors). Clusters 1, 2, 3, and 9 are somewhat noisier than the low-variance clusters, with error characteristics
233 similar to those of the entirety of the dataset. That is, they are considered less skilled, but their estimates still lie on a
234 one-to-one line with respect to the true wind. Clusters 6 and 8, on the other hand, are clearly unskilled in different
235 ways. Cluster 6 is a noisy regime, which captures much of the more extreme differences between the AMV estimates
236 and the Nature Run winds. Cluster 8, on the other hand, represents the low AMV estimate, high Nature Run wind
237 regime. This cluster is returning AMVs with values of zero where the Nature Run wind is clearly non-zero because of
238 the very low water vapor present. We further see the stratification of the regimes when analyzing the absolute AMV
239 error in relation to the water vapor content (Figure 7). We see that clusters that have similar behaviors in the error
240 pattern (such as 1, 2, and 3) represent different regimes of water vapor content.

241 We specified 9 individual clusters due to a combination of quantitative and qualitative reasons. Quantitatively, the
242 'Mclust' package uses the Bayesian Information Criterion (BIC), a model selection criterion based on the likelihood
243 function which attempts to penalize overfitting, to select the optimal number of clusters given an input range. Using
244 an input range of one through nine, the BIC indicated the highest number of clusters would be optimal. More
245 importantly, however, the 9 clusters can be physically distinguished and interpreted. Plots of the geophysical variables
246 in the testing set associated with each of the clusters are shown in Figures 8-11. Specifically, Figure 8 plots the
247 distribution of water vapor for each cluster, while Figure 9 plots the mean wind magnitude in each direction by cluster.
248 Figure 10 plots the correlation matrix for each cluster and Figure 11 show the geographic distribution of each cluster.
249 In looking at these in combination, we see discernable and discrete clusters with unique characteristics. For example,
250 cluster 1 captures the very dry, high-wind regime in the southern hemisphere visible in Figure 2. Cluster 7
251 encompasses the tropics, while cluster 3 captures mid-latitude storm systems. Clusters 6, 8, and 9 are all characterized
252 by a much worse performance of the AMV tracking algorithm, exhibited both in Figure 7 and in Figure 8 but all
253 encompass different geographic and geophysical regimes. We see that the clustering algorithm succeeds in capturing
254 physically interpretable clusters without having any knowledge of the underlying physical dynamics. We note that in
255 other applications, the optimal number of clusters will change and the researcher will need to explore various choices
256 of this parameter in their modeling, although this tuning process should be greatly simplified by the inclusion of an
257 information criterion (e.g., BIC) in the GMM algorithm.

258 **3.5 Random Forest**

259 The clustering algorithm requires the Nature Run wind vector component values (u and v) in order to classify the
260 AMV error. When applying the algorithm in practice to tracked AMV wind from real observations, the true winds are
261 unknown. To represent the fact that we will not know the true winds in practice, we develop a proxy for the Nature
262 Run winds using only the AMV estimates and the simulated water vapor itself. This is an instance in which the
263 application of machine learning is desirable, since machine learning excels at learning high-dimensional non-linear

264 relationships from large training datasets. In this case, we specifically use random forest to create an algorithm which

265 predicts the Nature Run wind values as a function of the tracked wind values and water vapor.

266 Random forest is a machine learning regression algorithm which, as detailed by Breiman (2001), employs an ensemble

267 of decision trees to model a nonlinear relationship between a response and a set of predictors from a training dataset.

268 Here, we chose random forest specifically because it possesses certain robustness properties that are more appropriate

269 for our applications than other machine learning methods. For instance, random forest will not predict values that are

270 outside the minimum and maximum range of the input dataset, whereas other methods such as neural networks can

271 exceed the training range, sometimes considerably so. Random forest, due to the sampling procedure employed during

272 training, also tends to be robust to overtraining in addition to requiring fewer tuning parameters compared with

273 methods such as neural networks.

274 We trained a random forest with 50 trees on a separate set of tracked winds and water vapor values to predict Nature

275 Run winds using the 'randomForest' package in R. While the random forest estimate as a whole does not perform

276 much better than the AMV values in estimating the Nature Run wind (2.89 RMSE for random forest vs 2.91 RMSE

277 for AMVs), as shown in Figure 12, it does not display the same discrete regimentation as the AMV estimates in Figure

278 3. As such, the random forest estimates can act as a proxy for Nature Run wind values in our clustering algorithm —

279 they remove the regimentation which is a critical distinction between the AMV estimates and the Nature Run wind

280 values.

281 **3.6 Finalized Error Characterization Model**

282 The foundation of the error characterization approach is to combine the random forest and clustering algorithm. We

283 apply the Gaussian mixture model, as trained on the Nature Run winds (in addition to the AMVs and water vapor), to

284 each point of water vapor, AMV estimate, and associated random forest estimate. This produces a set of clusters

285 which, when implemented, require no direct knowledge of the actual Nature Run state (Figure 13).

286 Naturally, the clustering algorithm performs better when applied to the dataset with the Nature Run winds, as

287 opposed to winds generated from the random forest algorithm. The former is created with direct knowledge of the

288 Nature Run winds, and any approximation will lead to increased uncertainties. In practice, the performance of the

289 cluster analysis can be improved by enhancing the performance of the random forest itself. As with any machine

290 learning algorithm, the random forest contains hyperparameters that can be optimized for specific applications. In

291 addition, performance could be improved by including additional predictor variables. Our intent is not to use the

292 random forest as a wind tracking algorithm; rather, the random forest is presented in this paper as a proof of concept.

293 Nonetheless, we see in Figure 13 and Figure 14 that the error characterization still discretizes the testing data set into

294 meaningful error regimes. The algorithm manages to separate the AMV estimates into appropriate error clusters. Once

295 again, clusters 6 and 8 manage to capture unskilled regimes, and cluster 7, and to a lesser extent clusters 4 and 5,

296 remain skillful. By taking the mean and standard deviation of the difference between AMV estimates and Nature Run

297   winds in each cluster, we develop error characteristics for each cluster (Figure 15); these quantities are precisely the

298   bias and uncertainty that we require for the cost function J in Eq (1). We see that the unskilled clusters have very high

299   standard errors and they correspond roughly to the areas of unskilled regimes in Figure 3. Similarly, skilled clusters

300   5, 4 and 7 have standard errors below that of the entire dataset. Since each cluster now has associated error

301   characteristics (e.g., bias and standard deviation), it is then straightforward to assign the bias and uncertainty for any

302   new tracked wind observation by computing which regime it is likely to belong to.

3.7 Experimental Set up

304   In this section we will describe our experimental setup for training our model on the GEOS-5 Nature Run data and

305   testing its performance on a withheld dataset. We divide the dataset into two parts: a training set consisting of the first

306   1.5 months of the GEOS-5 Nature Run, and a testing set consisting of the last 0.5 month of the Nature Run. Our

307   training/testing procedure for the simulation data and tracked wind is as follows:

308   1.   Divide the simulation data and tracked wind into two sets: training set of 1,000,000 points from the first 1.5
309   months of the Nature Run and a testing set of 1,000,000 points from the final 0.5 months of the Nature Run.
310   2.   We train a Gaussian Mixture Model on a normalized random sample of observations from the training dataset
311   of Nature Run winds (u and v direction), tracked winds (u and v direction), and water vapor with n=9 clusters.
312   3.   We train two separate random forests on a different random sample of 750,000 observations from the training
313   dataset. We use tracked wind (u and v direction) and water vapor to model, separately, Nature Run winds in
314   both the u and v directions.
315   4.   We apply the random forests to the dataset used for the Gaussian Mixture Model. This provides a random
316   forest estimate for each point, which is used as a substitute for Nature Run wind values in the next step.
317   5.   We predict the Gaussian mixture component assignment for each point of water vapor, tracked winds, and
318   random forest estimate using the GMM parameters estimated in Step 2.
319   6.   We compute the mean and standard deviation of the difference between the tracked winds and the Nature
320   Run winds, per direction, for each Gaussian mixture model cluster assignment. This provides a set of error
321   characteristics that are specific to each cluster.
322   7.   We can apply the random forest, and then the cluster estimation, to any set of water vapor and tracked AMV
323   estimates. Thusly, any set of tracked AMV estimates and water vapor can be mapped to a specific cluster,
324   and therefore its associated error characteristics.

4 Results and Validation

326   In this section, we compare our clustering method against a simple alternative, and we quantitatively demonstrate

327   improvements that result from our error characterization. Recall that in Section 3, we divided the wind-tracking

328   outputs into 9 regimes, which range from very skilled to unskilled. For the *i*-th regime, we can quantify the predicted

329   uncertainty estimate as a gaussian distribution with mean m_i and standard deviation σ_i, which has a well-defined

330   cumulative distribution function which we denote as F_i. To test the performance of our uncertainty forecast, we divide

331   the dataset described in Section 2 into a training dataset (first 1.5 month) and a testing dataset (last 0.5 month). Having

332   trained our model using the training dataset, we apply the methodology to the testing dataset, and we compare the

333   performance of the predicted probability distributions against the actual wind error (tracked winds - Nature Run

334   winds). This is a type of probabilistic forecast assessment, and we assess the quality of the prediction using a scoring

335   rule called continuous ranked probability score (CRPS), which is defined as a function of a cumulative distribution

336   function F and an observation x as follows:

$$\mathrm{CRPS(F,x)} = \int_{-\infty}^{\infty} (\mathrm{F(x)} - \mathbb{1}(\mathbf{y}-\mathbf{x}))^2 \, \mathbf{dy} \tag{4}$$

338   Where $\mathbb{1}(\ )$ is the Heaviside step function and denotes a step function along the real line that is equal to 1 if the argument

339   is positive or zero, and it is equal zero if the argument is negative (Gneiting and Katzfuss, 2014) . The continuous rank

340   probability score here is strictly proper, which means that the function $\mathrm{CRPS(F,x)}$ attains the minimum if the data x

341   is drawn from the same probability distribution as the one implied by F. That is, if the data x is drawn from the

342   probability distribution given by F, then $\mathrm{CRPS(F, x)} < \mathrm{CRPS(G, x)}$ for all $G \neq F$.

343   The alternative error characterization method that we test against is a simple marginal mean and marginal standard

344   deviation of the entire tracked subtract Nature Run wind dataset. This is essentially equivalent to an error

345   characterization scheme that utilizes one regime, where m and σ are given as the marginal mean and the marginal

346   standard deviation of the residuals (i.e., tracked wind minus Nature Run winds). Here, we use a negatively oriented

347   version of the CRPS (i.e., Eq.(4) without the minus sign), which implies that lower is better. A histogram evaluating

348   the performance of our methodology against the naive error characterization method is given in Figure 16.

349   The relative behavior of the CRPS is consistent between u and v winds. The CRPS tends to have to wider distribution

350   when applied to the regime-based error characterization. Compared to the alternative error characterization scheme,

351   our methodology produces a cluster of highly accurate predictions (low CRPS scores), in addition to some cluster of

352   very uninformative predictions (high CRPS scores). These clusters correspond to the highly skilled cluster (e.g.,

353   Cluster 3) and the unskilled clusters (Cluster 6 and 8), respectively. Overall, the mean of the CRPS is lower for our

354   methodology than it is for the alternative method, indicating that as a whole our method produces a more accurate

355   probabilistic forecast.

356   Thus far we have shown that our method produces more accurate error-characterization than an alternative method

357   based on marginal means and variance. Now, we assess whether our methodology provides valid probabilistic

358   prediction; that is, we test whether the uncertainty estimates provided are consistent with the empirical distribution of

359   the validation data. To assess this, we construct a metric in which we normalize the difference between the Nature

360   Run wind and the tracked wind by the predicted variance. That is, for the i-th observation, we compute the normalized

361   values for u_i and v_i using the following equations:

362
$$z_{u,i} = \frac{u_i - \hat{u}_i}{\sigma_{u,i}}$$

363
$$z_{v,i} = \frac{v_i - \hat{v}_i}{\sigma_{v,i}} \tag{5}$$

364   Where u_i is the *i*-th Nature Run u wind from the Nature Run data, $\hat{u}_i$ is the tracked-wind, and $\sigma_{u,i}$ is the error as
365   assessed by our model (recall that it is a function of the regime index to which $\hat{u}_i$ has been assigned). The values for
366   the v-wind are defined similarly. The residuals in Eq (5) can be considered as a variant of the z-score, and it is
367   straightforward to see that if our error estimates are valid (i.e., accurate), then the normalized residuals in Eq. (5)
368   should have a standard deviation of 1. If our uncertainty estimates $\sigma_{u,i}$ and $\sigma_{v,i}$ are too large, then the standard deviation
369   of $z_{u,i}$ and $z_{v,i}$ should be less than 1; similarly, if our uncertainty estimates are too small, then the standard deviation
370   of $z_{u,i}$ and $z_{v,i}$ should be larger than 1. In *Figure 17,* we display the histogram of the normalized residuals z_u and z_v.
371   It is clear that for both types of wind, the standard deviation of $z_{u,i}$ and $z_{v,i}$ are 1.003 and 1.009, respectively, indicating
372   that our error characterization model is highly accurate when forecasting uncertainties.

373   **5 Conclusion**

374   Error characterization is an important component of data validation and scientific analysis. For wind-tracking
375   algorithms, whose outputs (tracked u and v) are often used as observations in data assimilation analyses, it is necessary
376   to accurately characterize the bias and standard error (e.g., see Section 2.2). Nguyen et al. (2019) illustrated that
377   incorrect specification of these uncertainties (**a** and **R** in Eq. (1)) can adversely affect the assimilation results –
378   mischaracterization of bias will systematically offset a tracked wind, while an erroneous standard error could
379   incorrectly weigh the cost function.

380   In this paper we demonstrate the application of a machine learning uncertainty modeling tool to AMVs derived from
381   hyperspectral sounder water vapor profiles. The methodology, based on a combination of gaussian mixture model
382   clustering and random forest, identified distinct geophysical regimes and provided uncertainties specific to each
383   regime. This was achieved in a purely data-driven framework; nothing was known to the model except the specific
384   inputs and outputs of the AMV algorithm, deducing the relationship between regime and uncertainty from the
385   underlying multivariate distribution of water vapor, Nature Run wind, and tracked wind. Our algorithm does require
386   one major tuning parameter in the number of clusters for the GMM algorithm, although the search for the 'optimal'
387   number of clusters can be aided by the inclusion of an information criterion (e.g., the BIC) in the GMM model.
388   Nonetheless this methodology was sufficient to produce improved error estimates of state-dependent uncertainties as
389   detailed in Posselt et al. (2019).

390   At its most general, our methodology consists of two parts: an emulator and a clustering algorithm. In this
391   implementation, random forest and Gaussian mixture modelling are the approaches; in theory, these two steps could
392   be accomplished using other algorithms belonging to the appropriate class. Indeed, improvements to the methodology

393 could surely be made with further research in both areas. Given the degradation in the uncertainty estimates between

394 those produce with and without Nature Run wind values, an improvement of the emulator could yield the most efficient

395 returns. This could either take the form of improving or replacing the random forest algorithm. As previously

396 discussed, improvements could also be made in both the inputs to and nature of the clustering algorithm.

397 We note that our methodology requires knowledge of the true field of interest (here u and v), or some proxy thereof.

398 This makes our methodology a natural supplement to OSSE-based studies where the true fields of interest are provided

399 by numerical weather models. Such studies are important components of algorithm validation (e.g., Posselt et al.,

400 2019), and our proposed methodology provides a framework for characterizing the error within different geophysical

401 regimes. In practice, we envision that the lack of true fields could be addressed by either using independent validation

402 data or reanalysis model data. Therefore, there would be an additional component of error due to the usage of proxy

403 data in lieu of the true field, but this error should be inversely proportional to the quality of the proxy data.

404 The error estimation algorithm presented in this paper is a proof of concept. While the methodology is expected to be

405 generally applicable to other AMV extraction methods (e.g., cloud-track winds or tracking of other trace gases), our

406 specific error functions are only valid for our specific training dataset. That said, the state-dependent errors identified

407 by Posselt et al. (2019) are also expected to apply to other water vapor AMVs. This is because, in general, AMV

408 algorithms have difficulty tracking fields with very small gradients, and will produce systematic errors in situations

409 for which isolines in the tracked field (e.g., contours of constant water vapor mixing ratio) lie parallel to the flow. To

410 the extent that our algorithm represents a general class of errors, the results may be applicable to other geophysical

411 scenarios and other AMV tracking methodologies. As mentioned in the introduction, robust estimates of uncertainty

412 are important for data assimilation, and we expect that our methodology could be used to provide more accurate

413 uncertainties for AMVs used in data assimilation for weather forecasting and reanalysis. To do so, we recommend

414 training on a dataset that is large enough to encompass the full range of features likely to be seen by the assimilation

415 and forecast system. To the extent that errors may be seasonally and regionally dependent, it may be more effective

416 to train the error estimation algorithm on data that is expected to represent the specific flow regimes and water vapor

417 features valid for a particular forecast or assimilation period. We have tested the error function described in this paper

418 in an AMV weather forecast and data assimilation OSSE study using GEOS5, and its impact will be reported in a

419 forthcoming paper.

420 **Author Contribution**

421 Teixeira conceived of the idea with inputs from Nguyen. Teixeira performed the computation. Wu provided the

422 experimental datasets along with data curation expertise. Posselt and Su provided subject matter expertise. All authors

423 discussed the results. Teixeira wrote the initial manuscript and updated the draft with inputs from co-authors.

424 **Competing Interest:** The Authors declare no conflict of interest.

425 **Funding Acknowledgment**: The research was carried out at the Jet Propulsion Laboratory, California Institute of

426 Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). © 2020.

427 California Institute of Technology. Government sponsorship acknowledged

References

429 Bormann, N., Hernandez-Carrascal, A., Borde, R., Lutz, H.J., Otkin, J.A. and Wanzong, S.: Atmospheric motion
430 vectors from model simulations. Part I: Methods and characterization as single-level estimates of wind, Journal of
431 Applied Meteorology and Climatology, 53(1), 47-64. https://doi.org/10.1175/JAMC-D-12-0336.1, 2014.

432 Breiman, L.: Random forests. Machine learning, 45(1), 5-32, 2001.

433 Cassola, F. and Burlando, M.: Wind speed and wind energy forecast through Kalman filtering of Numerical Weather
434 Prediction model output, Applied Energy, 99, 154-166, 2012.

435 Coulston, J.W., Blinn, C.E., Thomas, V.A. and Wynne, R.H., 2016. Approximating prediction uncertainty for
436 random forest regression models. *Photogrammetric Engineering & Remote Sensing*, *82*(3), pp.189-197.

437 Fraley, C. and Raftery, A.E.: MCLUST: Software for model-based clustering, density estimation and discriminant
438 analysis (No. TR-415). Washington University, Seattle Department of Statistics, 2002.

439 Fraley, C., Raftery, A.E., Murphy, T.B. and Scrucca, L: mclust version 4 for R: normal mixture modeling for model-
440 based clustering, classification, and density estimation, Washington University, Seattle Department of Statistics, 2012

441 Gneiting, T. and Katzfuss, M., 2014. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*,
442 pp.125-151

443 Hernandez-Carrascal, A. and Bormann, N.: Atmospheric motion vectors from model simulations. Part II:
444 Interpretation as spatial and vertical averages of wind and role of clouds, Journal of Applied Meteorology and
445 Climatology, 53(1), 65-82, 2014.

446 Holmlund, K., Velden, C. S., & Rohn, M.: Enhanced automated quality control applied to high-density satellite-
447 derived winds, Monthly Weather Review, 129(3), 517-529, 2001.

448 Kawa, S.R., Erickson, D.J., Pawson, S. and Zhu, Z.: Global CO2 transport simulations using meteorological data
449 from the NASA data assimilation system, Journal of Geophysical Research: Atmospheres, 109,
450 https://doi.org/10.1029/2004JD004554, 2004.
451
452 Kwon, Y., Won, J.H., Kim, B.J. and Paik, M.C., 2020. Uncertainty quantification using Bayesian neural networks in
453 classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, *142*,
454 p.106816.

455 Le Marshall, J., Rea, A., Leslie, L., Seecamp, R., & Dunn, M.: Error characterisation of atmospheric motion vectors,
456 Australian Meteorological Magazine, 53(2), 2004.

457 Mueller, K.J., Wu, D.L., Horváth, Á., Jovanovic, V.M., Muller, J.P., Di Girolamo, L., Garay, M.J., Diner, D.J.,
458 Moroney, C.M. and Wanzong, S.: Assessment of MISR cloud motion vectors (CMVs) relative to GOES and
459 MODIS atmospheric motion vectors (AMVs), Journal of Applied Meteorology and Climatology, 56(3), 555-572,
460 https://doi.org/10.1175/JAMC-D-16-0112.1, 2017.
461

462 Nguyen, Hai, Noel Cressie, and Jonathan Hobbs. "Sensitivity of Optimal Estimation Satellite Retrievals to
463 Misspecification of the Prior Mean and Covariance, with Application to OCO-2 Retrievals." *Remote Sensing* 11.23
464 (2019): 2770.

465 Posselt, D. J., L. Wu, K. Mueller, L. Huang, F. W. Irion, S. Brown, H. Su, D. , and C. S. Velden: Quantitative
466 Assessment of State-Dependent Atmospheric Motion Vector Uncertainties. J. Appl. Meteor. Clim., In Press.
467 https://doi.org/10.1175/JAMC-D-19-0166.1., 2019.

468 Putman, W., A.M. da Silva, L.E. Ott and A. Darmenov: Model Configuration for the 7-km GEOS-5 Nature Run,
469 Ganymed Release (Non-hydrostatic 7 km Global Mesoscale Simulation). GMAO Office Note No.5 (Version 1.0),
470 18, 2014.

471 Salonen, K., J. Cotton, N. Bormann, and M. Forsythe: Characterizing AMV Height-Assignment Error by Comparing
472 Best-Fit Pressure Statistics from the Met Office and ECMWF Data Assimilation Systems, J. Appl. Meteor.
473 Climatol., 54, 225–242, https://doi.org/10.1175/JAMC-D-14-0025.1, 2015.

474 Staffell, I. and Pfenninger, S.: Using bias-corrected reanalysis to simulate current and future wind power output,
475 Energy, 114,1224-1239, 2016.

476 Swail, V.R. and Cox, A.T.: On the use of NCEP–NCAR reanalysis surface marine wind fields for a long-term North
477 Atlantic wave hindcast, Journal of Atmospheric and oceanic technology, 17(4), 532-545, 2000.

478 Tran, D., Dusenberry, M., van der Wilk, M. and Hafner, D., 2019. Bayesian layers: A module for neural network
479 uncertainty. In *Advances in Neural Information Processing Systems* (pp. 14660-14672).
480
481 Tripathy, R.K. and Bilionis, I., 2018. Deep UQ: Learning deep neural network surrogate models for high
482 dimensional uncertainty quantification. *Journal of computational physics*, *375*, pp.565-588.

483 Velden, C.S. and K.M. Bedka,: Identifying the Uncertainty in Determining Satellite-Derived Atmospheric Motion
484 Vector Height Attribution. J. Appl. Meteor. Climatol., 48, 450–463, https://doi.org/10.1175/2008JAMC1957.1 2009.

485 Zeng, X., S. Ackerman, R.D. Ferraro, T.J. Lee, J.J. Murray, S. Pawson, C. Reynolds, and J. Teixeira:
486 Challenges and opportunities in NASA weather research. Bull. Amer. Meteor. Soc., 97, 137–140, 2016.

487

488

489

490

491

492

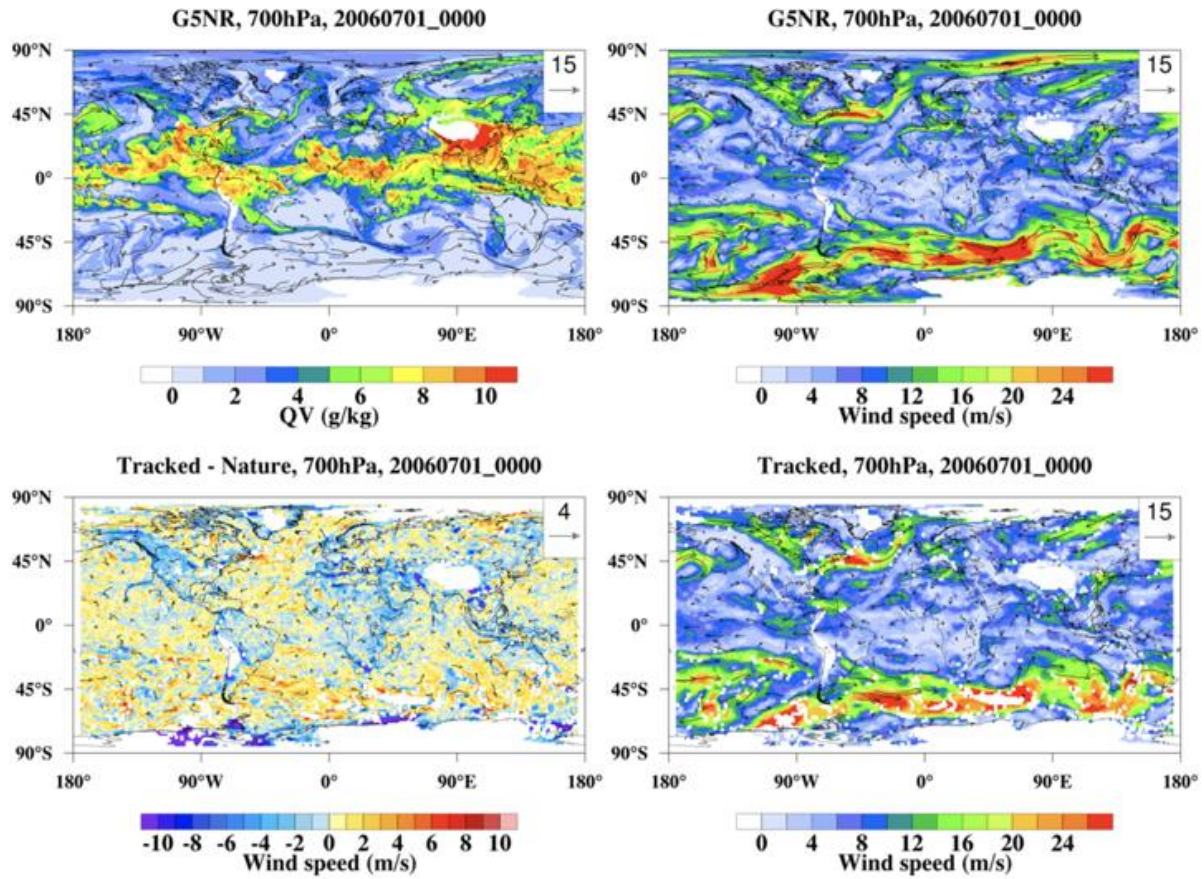
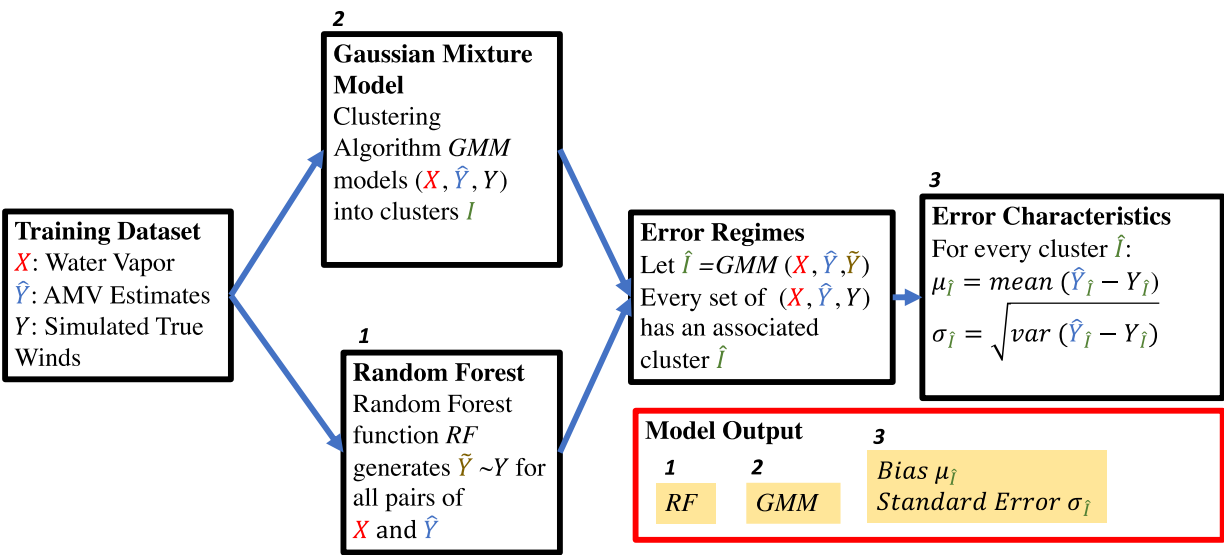
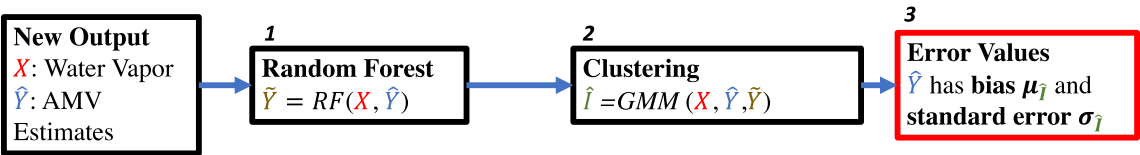



Figure 1: Map of Nature Run at one timestep at 700hPa (A): Water Vapor (B): Nature Run Wind Speed (C): Difference between Nature Run Wind Speed and AMV Estimate (D): AMV Estimate.

1. Training

2. Implementation

 Figure 2: Diagram of Training Approach and Diagram of Implementation steps.

497

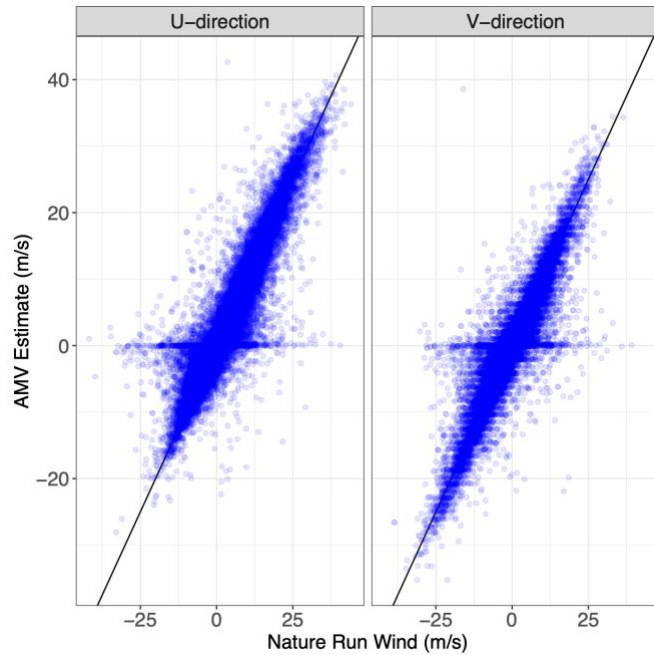



Figure 3: Scatter plot of the simulated Nature Run wind vs AMV estimates for u and v wind in the training

dataset.

500

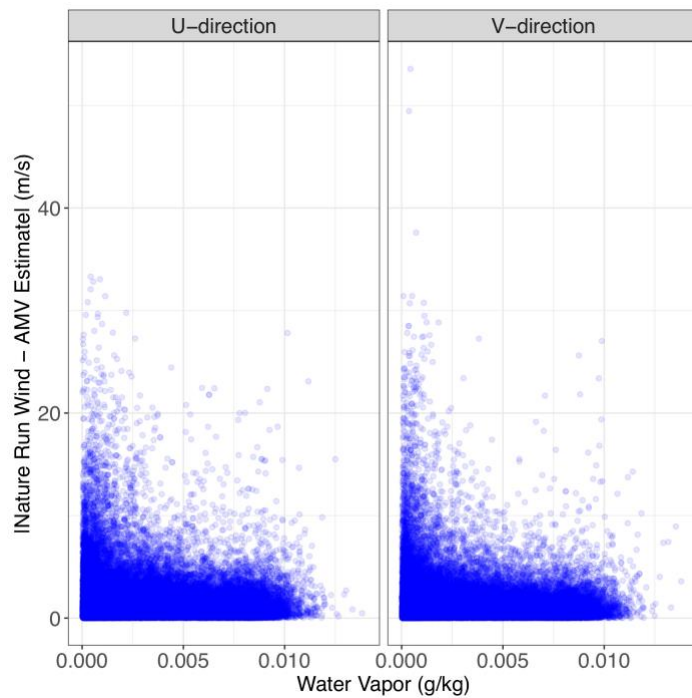



Figure 4: Simulated water vapor vs the absolute value of the difference between Nature Run and tracked

winds in the training dataset.

503

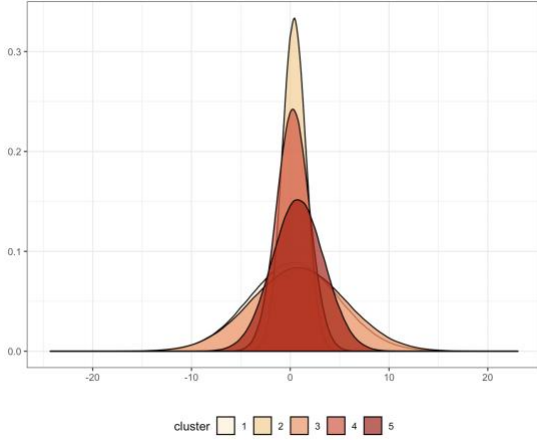



Figure 5: Example of Gaussian Mixture Model in one dimension. Density Figures for the U-Direction AMV Estimate dimension of fitted Gaussian mixture.

506

507

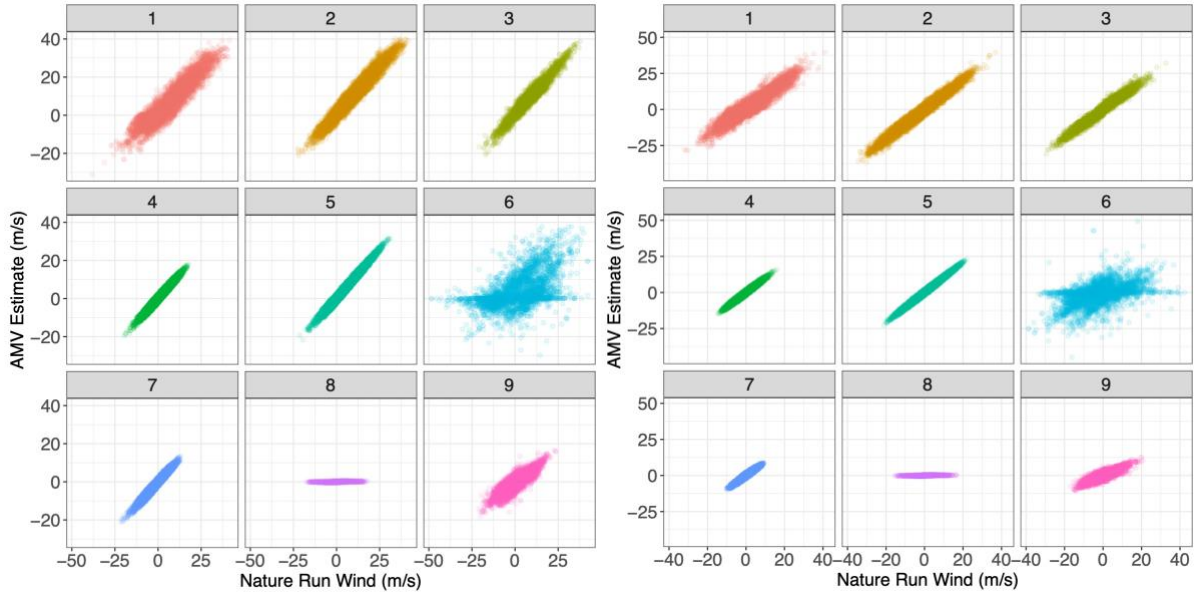



Figure 6: Scatterplot of simulated Nature Run wind vs AMV Estimates, each sub-panel corresponding to the specific Gaussian mixture component to which each point in the testing set has been assigned. (A): U-Direction Wind (B): V-Direction Wind.

511

512

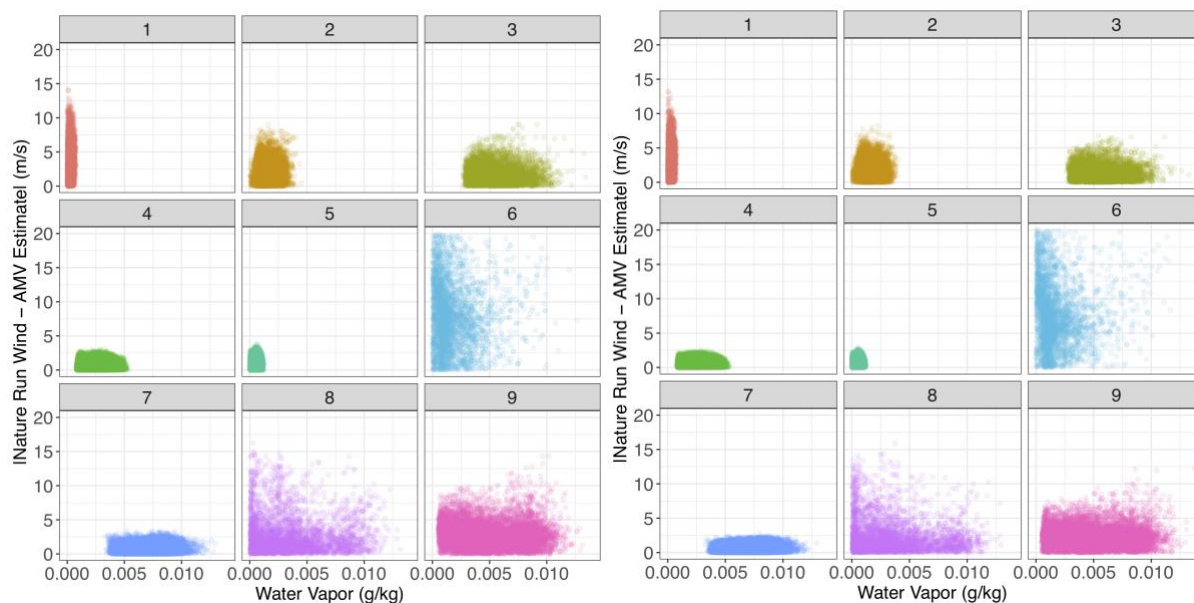
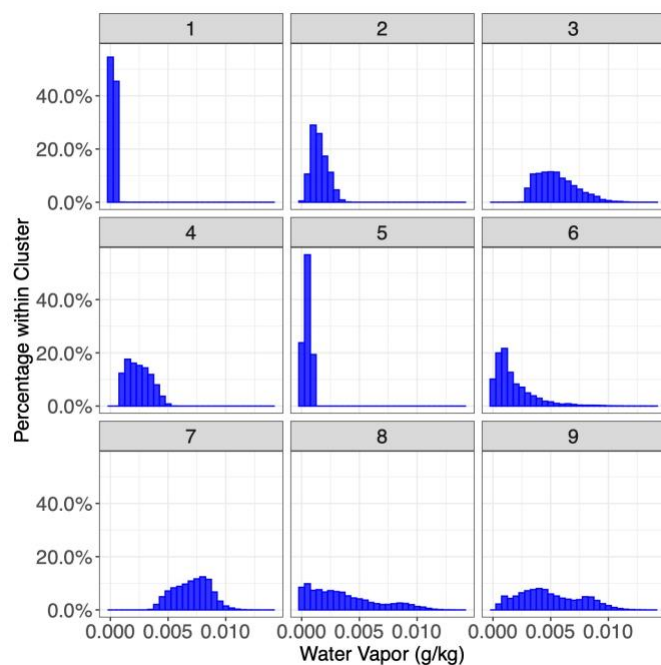
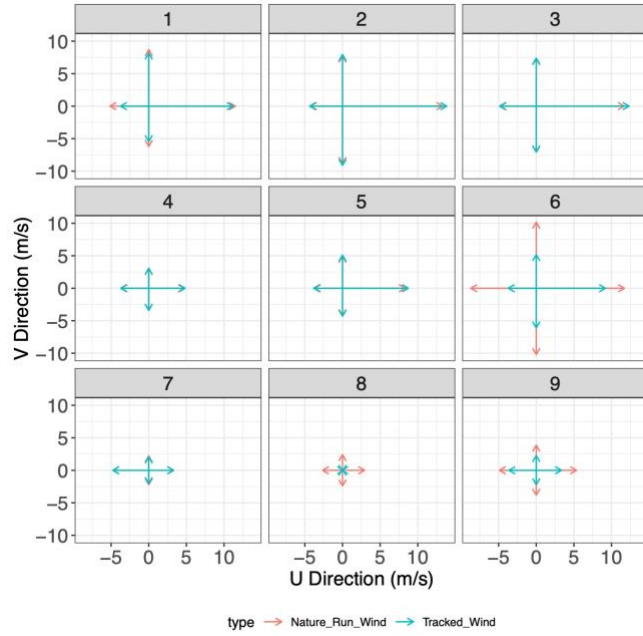
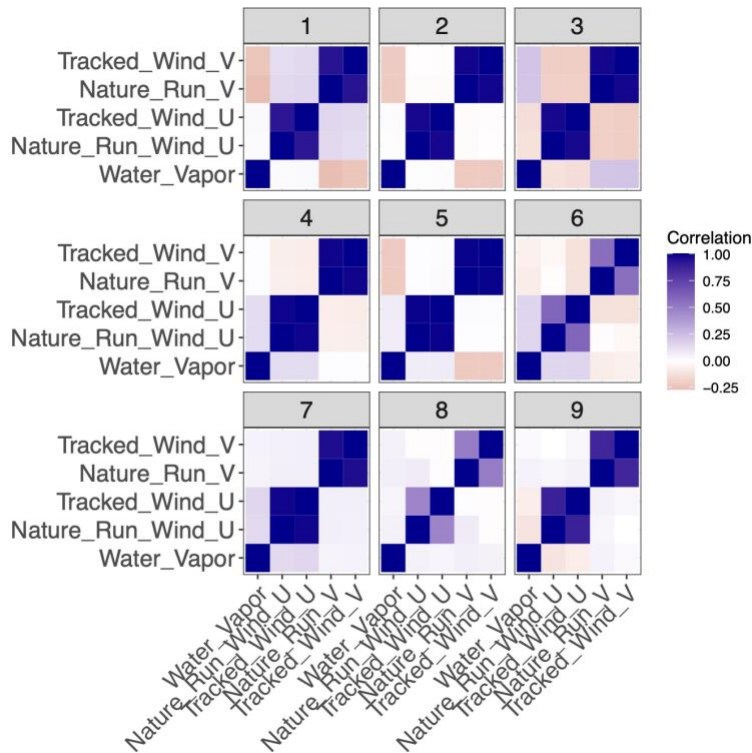
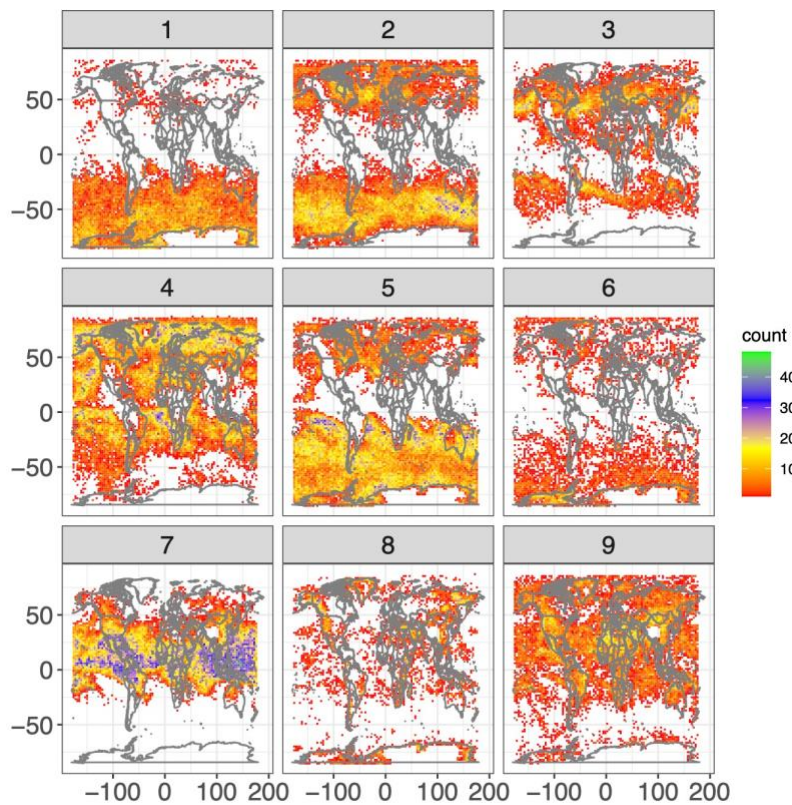



Figure 7: Scatterplot of Water Vapor vs Absolute Tracked Wind Error, each sub-panel corresponding to the specific Gaussian mixture component to which each point in the testing set has been assigned. (A): U-Direction Wind (B): V-Direction Wind.

516
517 **Figure 8: Histogram of Nature Run water vapor for each cluster identified by the Gaussian mixture model,**
518 **applied to the testing set. Each sub-panel represents the cluster each point was assigned to.**

519
520 **Figure 9: Mean tracked winds and Nature Run winds, in each direction, for each cluster applied to the test**
521 **set. Each sub-panel represents the cluster each point was assigned to.**

522

523
524 **Figure 10: Correlation matrix between each clustered element for each identified cluster in the original**
525 **training dataset. Each sub-panel refers to a specific cluster.**

526
Figure 11: Geographic distribution by cluster of AMV retrieval locations in the testing dataset. Each sub-panel represents one cluster.

529

530

531

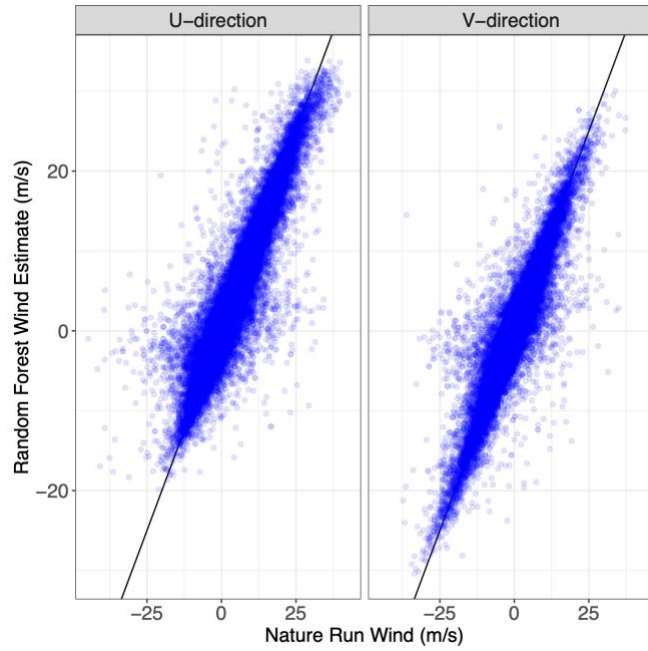



Figure 12: Scatterplot of Nature Run wind estimate vs random forest produced estimate. (A): U Direction (B): V Direction

532
533

534

535

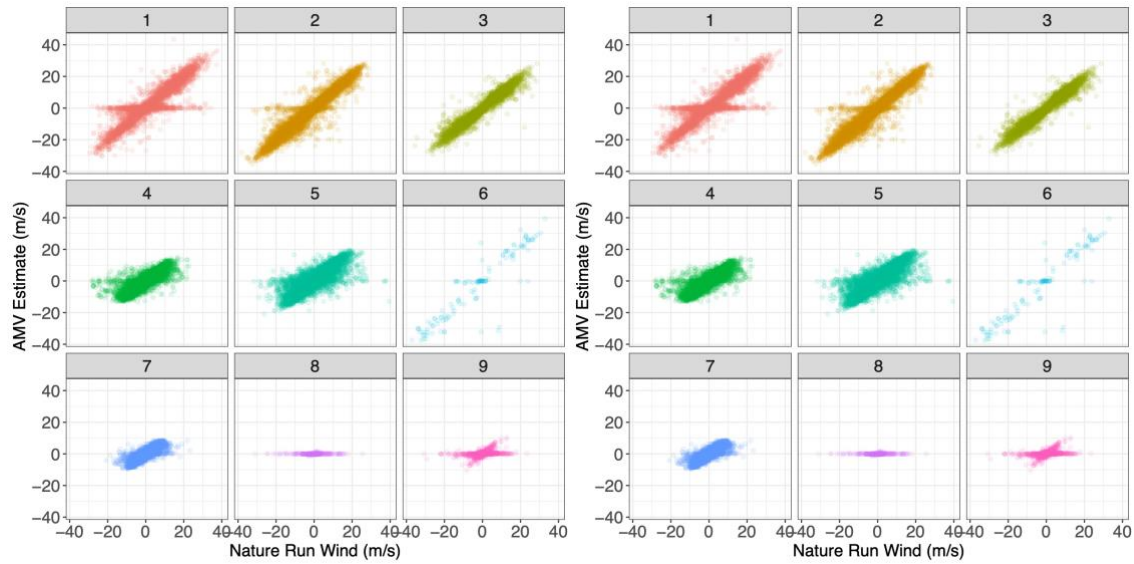



Figure 13: Scatterplot of Nature Run wind vs AMV Estimates, each sub-panel corresponding to the specific Gaussian mixture component to which each point in the testing set has been assigned when the Nature Run wind value has been substituted by the random estimate. (A): U-Direction Wind (B): V-Direction Wind

539

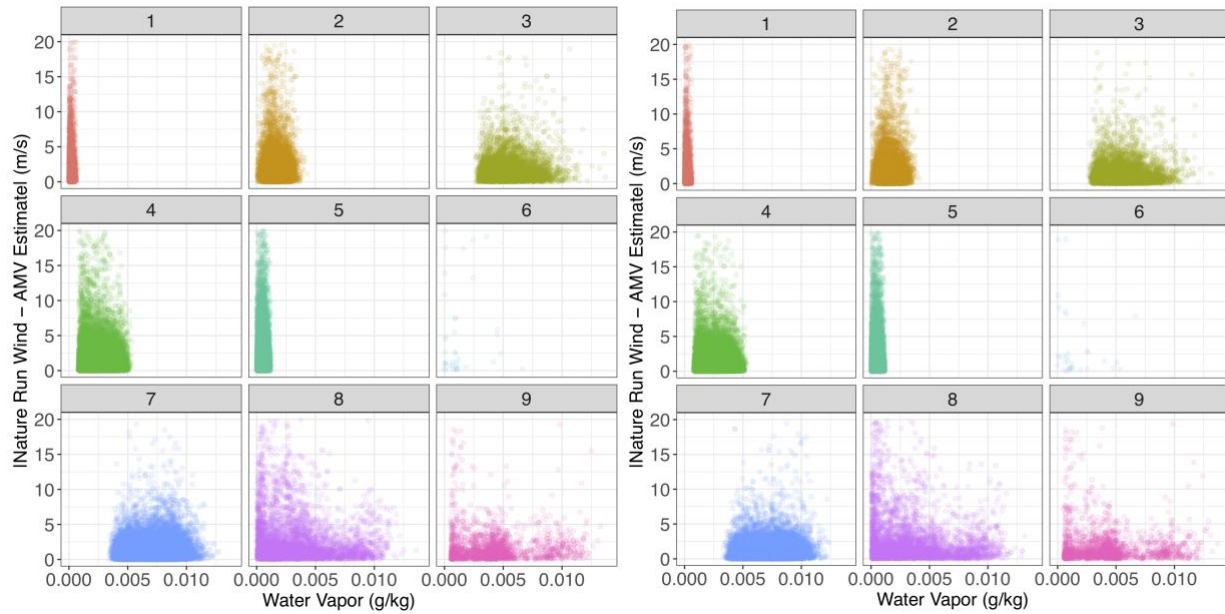



Figure 14: Water Vapor vs Absolute Tracked Wind Error, each sub-panel corresponding to the specific Gaussian mixture component each point in the testing set has been assigned when the Nature Run wind value has been substituted by the random estimate. (A): U-Direction Wind (B): V-Direction Wind

543

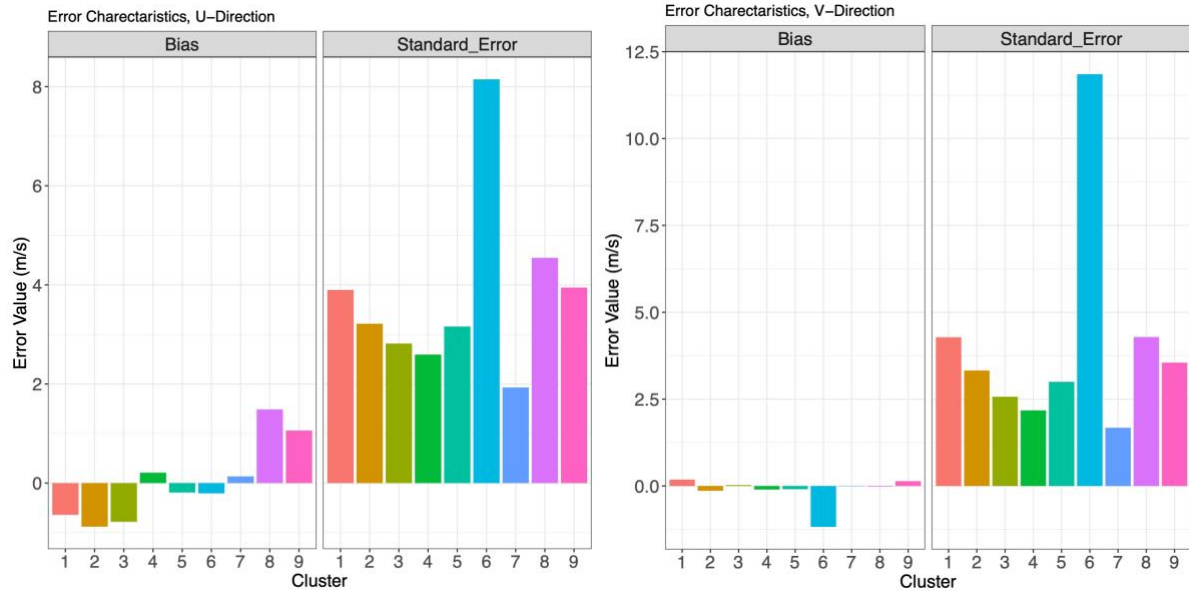



Figure 15: (A): Bias (Left Panel) and Standard Error (Right Panel) for each Gaussian mixture cluster in figure 6, U direction. (B): Same as (A) for V-direction

546

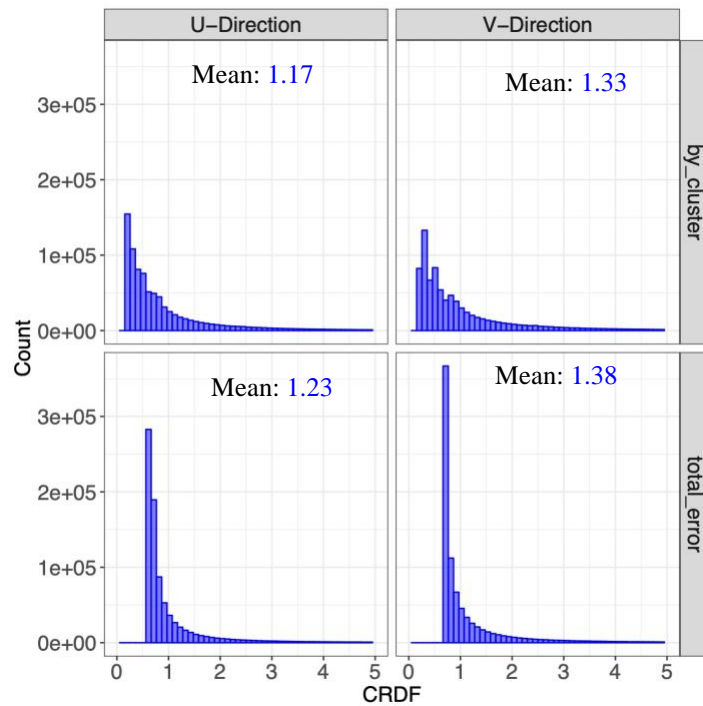



Figure 16: CRSP applied to different error approaches. (A): Cluster Errors for U Winds (B): Total Errors for U Winds (C): Cluster Errors for V Winds (D): Total Errors for V Winds.

549

550

551

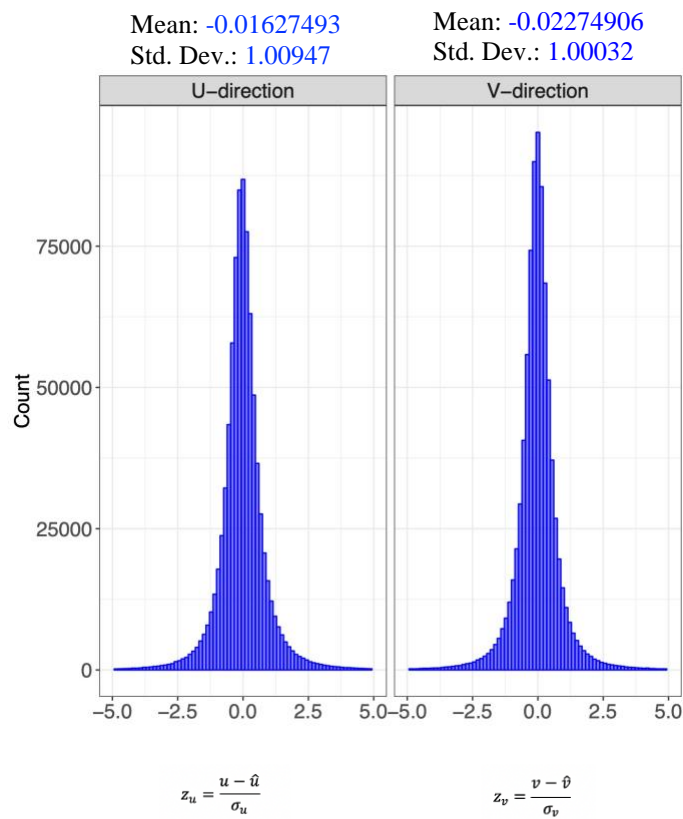



Figure 17: U and V winds normalized using the error characteristics developed by our methodology.

553

554

555

556

557

558