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

Teixeira et al.

Responses to Referee 1

We would like to thank the referee for the careful read of the paper and for the detailed comments. Please see our responses below:

1. *I have a problem with the present title. Reading it the first time I thought that paper was about improving error/quality of AMVs during the extraction process, and not during the assimilation process. From my understanding a title like: 'Use of Machines Learning to improve Uncertainty Quantification of Atmospheric Motion Vectors assimilated in NWP models', would certainly match better the real content of the paper and be less confusing.*

We certainly understand the reviewer's outlook on this; uncertainty quantification can often be a confounding term with different interpretations across subject areas. The title has been modified to "Using Machine Learning to Model Uncertainty for Water-Vapor Atmospheric Motion Vectors" to reflect this.

2. *The test presented in this paper is limited to water vapour AMVs extracted on specific layers. This potentially corresponds to extraction of 3D winds from hyperspectral sounders, as mentioned in the introduction. However, there is actually no evidence that the results can be generalised to the common AMVs extracted from clouds tracking in infrared or visible channels. If the method is limited to hyperspectral winds, this must be clearly specified in the text and probably also in the title of the paper, and not let the reader supposed that it works for all types of AMVs. If the method is not limited to hyperspectral AMVs authors have to present results also with common cloud motion winds extracted from satellite imagery. I understand from the text that another paper is upcoming (line 325), but there is no description or information that can actually let me assume that common AMVs have been used, and that the results are positive.*

Referee #2 expressed similar concerns, and we can understand the reviewers' perspective. We have qualified our statement that the approach may be globally applicable to any measurements, and have stated more specifically that it is likely to be useful for other sources of AMVs (especially those obtained by tracking gradients in trace gases). In paragraph 3 of the introduction we have included additional mention of the height assignment errors known to be an issue in tracking cloud features from radiances. This source of error is expected not to be as great of an issue when tracking retrieved trace gases (as shown in Posselt et al. 2019), as it is when tracking cloud features or radiance images. In addition, there are sources of error that are expected to be common to any feature tracking algorithm (e.g., regions without strong gradients in the field being tracked, or regions in which the wind is oriented parallel to contours in the field being tracked). We have modified our conclusions to include this in the last paragraph of discussion.

3. *The algorithm seems to be too dependent on the user's choice of the number of clusters, and the paper does not discuss the dependence of the algorithm on the chosen training dataset. It is also very unclear if the different clusters identified could refer to kind of physical or geographical AMVs properties, or if they are only blindly resulting out of the numerical tests. Authors must clarify/discuss if the results may depend on the AMV extraction model used (Mueller 2017). It is not clear if the same clusters can be used for operational AMV extracted from other schemes too (NOAA, EUMETSAT, C2 JMA. . .Etc). If it is not the case I guess this study must be repeated individually for every different AMV extraction schemes and maybe after every releases of these codes, which should represent an important limitation for operational use in NWP models. Although the authors promise the possibility to distinguish different geophysical regimes, the application ultimately presented by the paper comes down to discriminating the AMVs that are null because they are tracking the ground radiance, which is much too simple to showcase the real benefits of the algorithm.*

This study is meant to be a proof of concept – to show how a combination of random forest, plus a Gaussian mixture model, can be used to learn error structures found via comparison of simulated measurements with a reference "truth" dataset (as was done in our previous work). Naturally, the particular algorithm developed in this paper is wholly dependent both on the nature run and the AMV extraction method. However, it is not intended to be an algorithm that can be immediately used in NWP models. Instead, we aim to present a model that can be reproduced (and tuned) for use in specific contexts of AMV methods and data assimilation frameworks. The computational costs of training the algorithm (~1 day on a single processor, per pressure level) and even the computational costs of running the AMV extraction on the nature run (an average of 3 days per pressure level, on a non-optimized cluster network), are not outside the usual demands when updating parameters of NWP models.

In regards to the physical and geographical properties of the identified clusters, we have added a section in lines 329-345 and Figures 8-11 discussing this. They illustrate that the clustering algorithm manages to generally discriminate among geophysical regimes. Regarding the choice of number of clusters, this is a tuning parameter that is highly specific to application. We note that having one or more tuning parameters is not uncommon in many data analysis methods (e.g., k-means, PCA, self-organizing network, random forest, neural nets, regularized regression, smoothing splines, wavelets, etc.). Here, our method requires only 1 major tuning parameter (the random forest model also has tuning parameters, but that process, being a supervised regression, can be guided by cross validation). We note that the search for the 'optimal' number of clusters should be guided by expert knowledge, although this process should be greatly simplified by including an information criterion (e.g., the Bayesian Information Criterion) in the Gaussian Mixture Modelling algorithm. We have updated the end of the last paragraph of Section 3.4 to include this discussion.

Specific Comments:

1) *Everywhere I would change the denomination "true wind" to "G5NR wind" throughout the text. No matter the quality of any dataset relating to physical quantities, it does not deserve to be called "true".*

We understand that the term ‘true’ can often be controversial even when referencing a simulation. The denomination has been changed to ‘Nature Run Wind’ throughout the text. Thank you for the comment.

2) *Line 144 It would be good to recall that this Figure relates to the first 1.5 months of the dataset, in the caption of the Figure.*

Thank you. The distinction between training and test dataset has been made throughout the figure captions.

3) *Lines 144-145 This is disappointing. Given the use of a powerful tool like GMM and the possibility of identifying "geophysical regimes" (line 132), I expected far more than just discriminating two groups, one being functional AMVs, and the other merely being the AMVs tracking the ground radiance, when the water vapour layer is too thin.*

Figure 8-11, and lines 329-345 show that the clustering algorithm performs adequately in capturing consistent geophysical regimes. We focus in this paper on the ‘skillfull’ vs ‘unskillfull’ distinction because it is the most straightforward analysis for our purposes. More specific regime dependent uncertainties (as discussed in response to reviewer 2) is certainly a forward step after scaling this methodology beyond proof of concept.

4) *Line 270 This parts misses a "is" between "xi" and "the".*

Thank you for catching this. It was been corrected.

5) *Section 4 The term Continuous Ranked Probability Score should be mentioned at least once before the formula at line 278. The two acronyms CPRS and CRPS are used in this section. Please correct.*

The typo has been corrected. We mentioned the full name for CRPS immediately preceding its equation in (4), and we added a reference to a paper (Gneiting and Katzfuss, 2014) immediately after the equation.

6) *Line 309 You are referring to Figure 13, and not Figure 12 as written.*

Thank you for catching this. It has been corrected.

7) *Lines 329-330 I find your conclusion a little daring, knowing that you had to try different numbers of clusters before actually managing to discriminate the null AMVs.*

We apologize for the ambiguity. Our intention in these lines was different from what came across. We meant to say that our algorithm is able to ‘find’ or separate geophysically meaningful clusters without requiring domain knowledge expertise or prior information on the distribution of the variables. Granted, the algorithm requires the users to slide the number of clusters across some scales, but this process is vastly simplified since there is only 1 scalar parameter to vary. As we noted before, having tuning parameters is par-the-course for the majority of data analysis methods such as k-means, PCA, self-organizing network, random forest, neural nets, regularized regression, smoothing splines, wavelets, etc.

We understand the referee's concern, however. Therefore we have removed the aforementioned lines in the Conclusion, and we have included a note about the need to optimize over the number of clusters in 2nd paragraph of the Conclusion.

"Using Machine Learning to Model Uncertainty for Water-Vapor Atmospheric Motion Vectors"
Teixeira et al.

Responses to Referee 2

We would like to thank the referee for the careful read of the paper and for the detailed comments. Please see our responses below:

> *1. My main criticism of the study is that I am unsure about the practical applicability of the results. The study relies on the "truth" being available from a nature run to train the algorithm in the first place (e.g., to derive the clustering, to derive the random forest). It is unclear to me how this will be circumvented for real-life applications, without introducing other problems that may jeopardise the performance of the algorithm. I am not convinced that the algorithm could be applied "as is" on Motion Vectors derived from humidity fields retrieved from real sounding data, and indeed no attempt is presented in the paper to investigate this. The paper should discuss how it is envisaged that the algorithm can be applied to real-life situations and what the potential problem areas are.*

This study is meant to be a proof of concept – to show how a combination of random forest, plus a Gaussian mixture model, can be used to learn error structures found via comparison of simulated measurements with a reference "truth" dataset (as was done in our previous work). As such, we would not expect the results to be applicable "as is". However, we do expect that there are certain errors endemic to AMVs that are captured by our algorithm, and as such are also applicable to other scenarios. We have revised our conclusions to contain a discussion of this issue, but in summary we expect that current practice in numerical weather prediction may provide guidance here. While we never know "truth" in any practical application, there are ways to approximate errors without having exact knowledge of the true field. This is done routinely to characterize errors in any observation used in any data assimilation system. Typically, error estimation involves comparison with respect to an independent dataset, and in the case of our machine learning algorithm, a similar procedure could be followed.

Furthermore, we note that in this paper we are primarily interested in the distribution of a retrieved quantity versus the hidden 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($\hat{Y}_i - Y$))). We model our uncertainty *relative to the truth*, and therefore we cannot avoid the need to have some instances of the true data, or proxies thereof. This is a departure from much of the literature on uncertainty modelling with machine learning (e.g., Coulston et al., 2016; Tripathy et al., 2018; Tran et al., 2019; Kwon et al., 2020), which primarily define the uncertainty of a prediction as var($\hat{Y}_i$), or how sensitive that prediction is to tiny changes in the models/inputs. Our methodology allows for error estimates that fit naturally within the data assimilation framework, and, unlike the sensitivity estimate var($\hat{Y}_i$), also enable hypothesis testing and risk determination in support of decision making. To address the referee's concern, we have expanded on this in the 2nd paragraph of Section 3.1 and the 4th paragraph of the conclusion.

> *2. In several areas the manuscripts appears to suggest that the method would be generally applicable, ie to other AMVs and possibly beyond (e.g., p3 L80 ". . . our methodology in principle could be used to quantify uncertainty in any measurements..."). I think this should be qualified. Subject to the point above, the algorithm may offer some value for AMVs derived from sounder retrievals; I suspect the value for the cloud-tracked AMVs is very limited - though these are currently the most widely used AMV datasets. There may be applicability beyond this, but the authors should explain more clearly how they expect the algorithm to be applied to "any measurement".*

We have qualified our statement that the approach may be globally applicable to any measurements, and have stated more specifically that it is likely to be useful for other sources of AMVs. There are sources of error that are expected to be common to any feature tracking algorithm (e.g., regions without strong gradients in the field being tracked, or regions in which the wind is oriented parallel to contours in the field being tracked). We have modified our conclusions to include this discussion.

3. It would be useful if the authors took a critical look at the physical basis or motivation of their algorithm. The algorithm attempts to provide an uncertainty estimate for a derived wind vector with the derived wind vector and water vapour as the only inputs. I would expect other factors to play a considerable role, such as predictors describing the texture of the scene (to characterise the likely success of the tracking step), or C2 predictors that describe more the meteorological conditions (to characterise how likely humidity features are passive tracers). Spatial consistency measures such as the ones typically used in the formulation of the Quality Indicator (Holmlund 1998) may also be relevant. The predictor choice used in the study appears ad-hoc to me, and it could almost certainly be improved.

The predictor choice is indeed constrained and could almost certainly be improved in implementation. However, the limits on the input variables are a specific decision and not an oversight. The framework presented in this paper is not to necessarily intended to produce the best possible AMV uncertainty algorithm but to show, in a proof of concept, what a purely data-driven approach can lead to. In particular, we based our approach around the state-dependent errors characterized in Posselt et al. (2019), and sought to build an error characterization model that is itself state-driven. Including other parameters towards improving the algorithm is certainly interesting, and would most likely occur when implementing this methodology at scale, but is beyond the specific intentions of this paper. We address this in L250-257. Furthermore, we see in Figures 8-11 that even these limited inputs can produce physically recognizable regimes.

Specific points:

1. Title: I find the title misleading, as the authors only address the uncertainty in the wind estimates, not the height assignment uncertainty, which is a leading contributor of uncertainty for the most commonly used AMVs. The use of "Atmospheric Motion Vectors" may also lead readers to believe they will read about cloudy-tracked winds, when the links to these in the manuscript are very weak. I suggest to be more specific in the title, maybe "Estimation of uncertainty in wind retrievals derived from tracking humidity structures using Machine Learning".

We understand the reviewer's viewpoint on this. With additional consideration of the comments from reviewer 1 on the subject matter, the title has been rewritten to reflect that the paper concerns water-vapor AMVs. With vapor-vapor AMVs, height assignment uncertainty is less of a concern (we address this in our response to specific point #3), and should guide the reader towards a better interpretation of what the paper covers.

2. p2, L34: Nguyen et al (2019) is referred to quite extensively in the paper (here and elsewhere), but is listed as a comparatively inaccessible report from the National Institute for Applied Statistics Research Australia. A journal paper with a similar title has recently been published, and I wonder whether this could be referred to instead.

The reference noted has been replaced with the most updated reference to this paper.

3. p2, L 44-45 "However, height assignment is not the dominant portion of the error. . .": This is a strong claim to make, and I think it needs to be backed up with a suitable reference. Retrievals from infrared or microwave sounders do not represent radiosonde-like profiles. For a given level in the retrieved profile, the averaging kernel will describe the characteristics in the vertical represented by the retrieval - and these are not Diracdelta functions. Height characteristics of AMVs derived from such retrievals will hence be rather complex, and interpreting them subsequently as single-level winds may well be a considerable contribution to the error budget. I am not aware that this aspect has been thoroughly investigated in the literature yet. It should at least be mentioned in the present study.

This statement has been rephrased and expanded upon in lines 54-57. We acknowledge that height assignment error due to misspecification of height in the water vapor profiles could be impactful on the uncertainties for the extracted AMVs. However, this uncertainty cannot be directly assessed through analysis of the AMV extraction algorithm alone. Instead, it necessitates quantified uncertainties on the water vapor profiles themselves, something which is well beyond the scope of this paper.

4. p2, L51 "The Expected Error . . . to correct AMV observation error.": The EE aims to provide an estimate of the statistical characteristics of the observation error, but does not try to correct any errors in the AMVs. Please rephrase.

Thank you for noticing this. This has been rephrased as recommended.

5. p3, L90/91: It would be useful to provide an idea of the spatial scales used in the tracking step, ie what is the typical size of the target used.

The tracking step size is a 33km grid box for a sigma level of 4.2. More details can be found in Posselt et al (2019).

6. p 3, L 100/101, Fig. 1: The authors emphasise the poorer performance in drier regions. While it is a little harder to see, my impression is that there is also poorer performance near frontal features (e.g, positive biases East of South America or East of North America). Poorer performance around frontal regions seems physically plausible, as single-level humidity may not be a passive tracer in these regions. I think it would be worth commenting on this in the main text. This could also motivate a predictor other than water vapour in the scheme developed later.

We also suspect that vertical motion may be part of the reason behind the large errors near fronts, although a portion is also certainly due to the features identified in Posselt et al. (2019) (winds oriented along lines of constant water vapor). This paper aims to model uncertainties that are both regime dependent and state dependent. Obviously, these are intertwined: we see in figure 11 that the unskillful cluster 6 has a large representation on the east coasts of North and South America, indicating that it is at least partially capturing this frontal dynamic. When optimizing the methodology at scale, special consideration for more specific regime types (that are not purely state dependent) is a positive way of improving the uncertainty modelling approach for specific needs.

7. p 5, L139-142: It is not quite clear to me whether the description of the training/testing dataset in this paragraph is effectively referring to the same datasets described later (p8 L248/249). I got the impression here that all data for the 1.5/0.5 months were used, but later it sounds as if the dataset was subsampled. I suggest making this clearer to avoid confusion.

The text has been rewritten to make it clearer that data has been subsampled from the training and testing datasets.

8. p 6, L187-191: It would be good if the authors could motivate further how they chose 9 clusters in the Gaussian mixture model. The text sounds as if it was a subjective choice, but maybe there was an objective component as well? Given the very limited inputs to characterise the conditions, and the lack of clear distinctions between some clusters, the chosen number of clusters appears high.

As the reviewer suspected, there was a combination of quantitative and qualitative reasoning in determining the number of clusters. We address this in lines 329-345. New figures 8-11 also show greater clarity of the distinction between clusters.

9. p 7, L224/225 "Relative to . . . entire dataset.": I am unsure about what is meant here. I suggest rephrasing.

This redundant sentence has been removed.

10. p 8, first paragraph: It looks to me as if the clustering algorithm performs significantly more poorly once the true wind value has been substituted. Contrary to what is said in the text, clusters 4 and 5 shown in Fig. 9 appear relatively unskilful, certainly in comparison to the same clusters shown in Fig. 6. Also, it looks as if the population in clusters 6 and 8 (referred to as the "unskilled" regimes) is very low, and much lower than what was found in Fig. 6. It appears that the assignment into these clusters is very different to what was possible before. This may not be too surprising, as the previous assignment had the benefit of the truth being available, but the aspect is not addressed much in the text.

This is certainly a chief concern we have with the approach. There is substantial degradation in the clustering algorithm's performance when the model is not given the true winds. An implementation of this methodology at scale could benefit from an improvement in the random forest (or its replacement with a better performing emulator). This is addressed in lines 261-266. We must note, however, that our ultimate intention is not to create a machine learning emulator for the wind-tracking algorithm, but simply to employ it to model uncertainties.

11. p8, second paragraph/Fig. 11: Are the differences in standard deviation or bias between the clusters statistically significant? Also, what is the relative population of each cluster? Judging by Fig. 9 and 10, the clusters with the most different standard deviation (clusters 6 and 8) appear to have relatively small populations, whereas the variation in standard deviation in the remaining clusters is smaller.

We have over 800,000 observations in the dataset, and their relative population is listed below

Regime	Count	Percent
1	42308	4.95%
2	77545	9.08%
3	49187	5.76%
4	231268	27.07%
5	190543	22.31%
6	311	0.04%
7	206353	24.16%
8	41223	4.83%
9	15491	1.81%

To address the question of whether the differences in standard deviation (std) or bias between cluster is statistically significant, we opted to construct confidence intervals for the bias and std within each regime using the bootstrap (Efron and Tibshirani, 1993). The procedure of our bootstrap is as follows

1. Subset the data to retain only observations with regime index j. Let's assume that we have N_j observation within this data subset
2. Sample *with replacement* N_j observations from this subset. This forms a bootstrap sample
3. From 2., compute an estimate of the bias and std.
4. Repeat step 2-3 for 1000 times, giving us 1000 estimates of the bias and 1000 estimates of the std within regime j.
5. Compute 95% confidence intervals from the 1000 estimates of bias and std from 4.

The results for the confidence intervals (in graphical forms) are listed below:

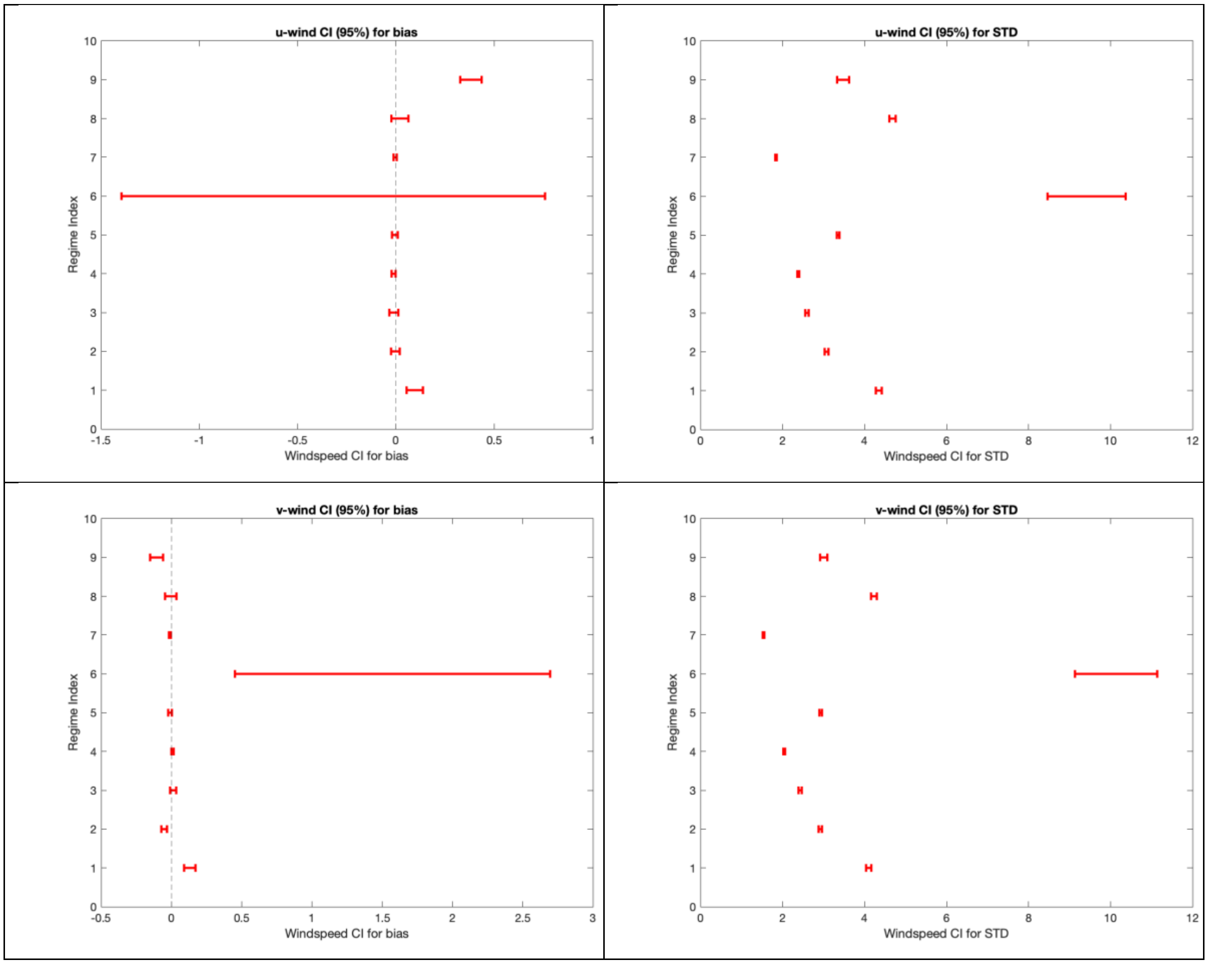


Figure: Top rows (bias and std confidence intervals for u-wind), bottom rows (bias and std confidence intervals for v-winds). The interval represent a 95% confidence interval.

We note that the Figure above indicates that for many of the biases, they can be considered unbiased since their confidence interval includes 0 (e.g., regimes 2-8 for u-wind). However, the plot also clearly indicates that two regimes are statistically different from 0 (regime 1 and 9). We also note that for the standard deviation maps, the CI's indicate that they are fairly stable (small narrow range) and that most of the regimes have statistically different standard deviation (denoted here visually as CI's that do not overlap one another). We also note that u and v wind direction tend to have very similar patterns, indicating that our regime classification is persistent across u and v.

To summarize, the CI plot above indicate that the differences in std between different regimes are highly statistically significant (as evidenced by the small confidence intervals and their spacing). For the biases, 3 of the regimes are statistically significantly different from the rest (i.e., regimes 1, 6, and 9), while the rest are likely relatively unbiased (i.e., bias = 0).

12. p 8, L248/249: The authors mention that they use a training set of 1,000,000 points, and a testing dataset of the same number of points. How have these been chosen within the available data? It looks as if many more points were available, at least for the training dataset. Also, the link to p 5 L139-142 was not quite clear to me.

We apologize for the lack of clarity on line 239-242. What we meant was that we used the NatureRun data from Posselt et al. (2019), which applied an AMV algorithm to outputs from the NASA Goddard Space Flight Center (GSFC) Global Modeling and Assimilation Office (GMAO) GEOS-5 Nature Run (G5NR; Putman et al. 2014). The

Nature Run is a global dataset with ~7 km horizontal grid spacing that includes, among other quantities, three-dimensional fields of wind, water vapor concentration, clouds, and temperature. The AMV algorithm is applied on four pressure levels (300hPa, 500hPa, 700hPa, and 850hPa) at 6-hourly intervals, using three consecutive global water vapor fields spaced one hour apart, and for a 60-day period from 07/01/2006 to 08/30/2006. In this paper, we make use of this dataset, although we focus only on the data at 700 hPa. We updated the manuscript on line 123-124 to refer to the data description in Section 2.1 and to make clear that we are using the data at 700 hPa.

Regarding the full dataset, it uses a 5758 x 2879 grid for longitude and latitude, with 240 time steps (60 days at 6 hours intervals). This forms a 5758 x 2879 x 240 = 3978547680 data points, which is simply too large for us to feasibly train a model. Therefore, we subsampled 1,000,0000 data points from this dataset uniformly where each of the 3978547680 data point has an equal chance of being selected.

Thank you for bringing this point to our attention. We have clarified the paper about the sampling process on the 1st bullet point of Section 3.7, and we have added information about the lon, lat, time grid at the bottom of the 1st paragraph in Section 2.1.

13. p 9, formula 4 and elsewhere: Typo: CPRS should be CRPS.

Thank you for catching this. The typo has been fixed.

14. p 9, L279-283: The "≤" in L282 appears to be inconsistent with what is said about CRPS earlier in the paragraph.

Thank you for catching this. The mistake was earlier in the paragraph, and has been addressed.

15. Fig. 12 and 13: Are these showing results for the test dataset? I assume they do (based on what is said on p 5, L141/142), but I think it would be clearest if this information was provided in the caption (a similar comment could be made for Fig. 6- 11).

The distinction between the training and test dataset has been made throughout the figure captions.

16. p 10, L306-311: The authors point to the finding that the residuals normalised with the estimated error have a standard deviation close to 1. It's a useful cross-check, but I suspect this finding primarily reflects that the training and testing data has similar standard deviations of AMVs vs true winds. I suspect it would have been obtained by assigning one constant observation error equal to the standard deviation of the whole population together. It would be more meaningful to consider other metrics that measure the Gaussianity of the distribution.

The reviewer's assessment is correct in that assigning a constant observation error equal to the standard deviation of the whole population together would also produce normalized residuals with standard deviation close to 1. However, this test is designed to show that our error predictions are actually consistent with the variability in validation data (this is termed 'validity' in the statistical literature).

As a thought experiment, consider the case of optimal estimation uncertainty estimates. Optimal estimation (Rodgers, 2000) purports to make estimates of the distribution $[\hat{Y}_i – Y]$ by making assumptions about the data structures, distributions, and/or forward models, and the robustness of these uncertainty estimates are usually only valid if these assumptions are correct. It is well-known in remote sensing that retrievals from optimal estimations tend to produce uncertainty estimates that are too low relative to validation data (Hobbs et al., 2017). For example, the uncertainties from OE for the Orbiting Carbon Observatory-2 (OCO-2) instrument tend to be too small (relative to validation data) by a factor of two. If we applied the same z-score test to the OCO-2 data, we would have obtained standard deviations of z-scores that is probably around 2, indicating that there is something awry with their error estimates.

The referee has noted that an error estimate can be 'valid' without being useful (this is the case with using the population standard deviation). This is why we also included the discussion on the CRPS, which gives a comparative assessment of skill (or usefulness) between two different predictions, and we have shown in this paper that our regime-based method is more skillful than using the population-based mean, and at the same time its error predictions are also valid.

17. p 11, L326-333: Given the points 10, 11, and 16, I'm not fully convinced by the claim that the algorithm produces "accurate error estimates" and that it is as skilful as the authors claim in identifying areas where the derived Motion Vectors are less skilful. There is some skill improvement compared to assigning a single value, but that is a very low baseline to compare the results with. Quality Indicator values are, for instance, used at some NWP centres to assign situation-dependent observation error values to AMVs. How would the present algorithm compare to such a scheme? Also, the algorithm appears to perform not particularly convincingly in a situation where the truth was available for training and no measurement noise or retrieval errors further complicate the situation. How much skill will remain if it has to deal with these issues?

We understand the reviewer's concerns in this regard. The uncertainties presented in this paper are not, in it of themselves, a marked improvement from state of the practice AMV uncertainty modelling. But neither are they intended as such; this paper is a proof of concept for the methodology it entails. As discussed in previous comments, there is no doubt that the algorithm itself can be tuned and enhanced for specific use cases. This would involve some reckoning with retrieval error of the water vapor features. However, we do believe that the paper demonstrates that even a bare-bones implementation of the approach can produce uncertainties that are valid and, to some degree, useful. We further note that they are produced in a physics-agnostic framework with no underlying assumptions and, critically, with a data-driven analysis of only the state elements. The research presented in Posselt et al. (2019) is fundamental in driving the analysis in this paper: state-dependent errors provide the context for a purely state-dependent uncertainty modelling approach. Ultimately, we hope to add to the literature and understanding of AMV uncertainty modelling, not supplant existing approaches. To the extent that the specific uncertainties produced in this paper are useful, that will be exhibited in an upcoming paper.

Fig. 7 and Fig. 10: The scale of the y-axis is rather large. The region of interest is probably confined to values < 20 m/s.

Thank you for this note. The axis on the figures have been changed accordingly.

References

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

Hobbs, J., Braverman, A., Cressie, N., Granat, R. and Gunson, M., 2017. Simulation-Based Uncertainty Quantification for Estimating Atmospheric CO _2 from Satellite Data. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), pp.956-985.

Kwon, Y., Won, J.H., Kim, B.J. and Paik, M.C., 2020. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, *142*, p.106816.

Rodgers, C.D., 2000. *Inverse methods for atmospheric sounding: theory and practice* (Vol. 2). World scientific.

Tran, D., Dusenberry, M., van der Wilk, M. and Hafner, D., 2019. Bayesian layers: A module for neural network uncertainty. In *Advances in Neural Information Processing Systems* (pp. 14660-14672).

Tripathy, R.K. and Bilionis, I., 2018. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of computational physics*, *375*, pp.565-588.

List of relevant changes in manuscript (in order as they appear):

1. Abstract and Title:
    a. Title Change
2. Introduction:
    a. L55-57: Additional discussion of height assignment error
    b. L95-103: New paragraph detailing paper's intention to present itself as proof of concept.
3. Section 2:
    a. L117-118: Mention of the use of the term 'Nature Run' to describe the simulated true wind.
    b. L132-133: Mention of the dimension size of the dataset.
    c. L136-142: Further discussion of height assignment error
4. Section 3:
    a. L194-203: New paragraph discussing the paper's attempt to characterize retrieved quantity versus the truth, as opposed to variance of the retrieved quantity itself.
    b. L213-214: Specification of the fact that the training data is subsampled.
    c. L223-228: Additional discussion on the definition of 'skillfull' and 'unskillful' regime.
    d. L276-283: New paragraph discussing the choice of inputs to the clustering model, as well as a discussion of implementing the methodology on larger-scale use.
    e. L335-351: New paragraph discussing the choice of number of clusters and detailing the geophysical dynamics of the clusters themselves.
    f. L392-398: New paragraph discussing the degradation of results with the addition of the random forest emulator.
5. Conclusion:
    a. Rewritten conclusion (about 5 new paragraphs) which includes augmented discussion of the challenges of and approaches to implementing the methodology at scale. In particular, we discuss at greater length the necessity of a 'truth' dataset and how to deal with this challenge in custom implementations.
6. References:
    a. Five new references: Coulston et al. (2016), Gneiting and Katzfuss (2014), Kwon et al. (2020), Tran et al. (2019), Tripathy and Bilionis (2018)
    b. Updated reference for Nguyen et al. (2019)
7. Figures
    a. Four new figures (8-11) which visualize the geography and physical conditions of the clusters determined by the Gaussian mixture model.

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

Deleted: Quantification

Deleted: with Machine Learning

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.

Deleted: (also known as uncertainty quantification)

Deleted: quantification

Deleted: input into

Deleted: methods

Deleted: supplement

Deleted: This paper proposes an

Deleted: that

Deleted: quantification

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 remote sensing applications. Nguyen et al. (2019) proved analytically that when the input bias is incorrect in Bayesian methods (specifically, optimal estimation retrievals), then the posterior estimates would also be biased. Moreover, they proved that when the input standard error is 'correct' (that is, it is as close to the unknown truth as possible), then the resulting Bayesian estimate is 'efficient'; that is, it has the smallest error among all possible choices of prior standard error. Additionally, multiple active and passive technologies are being developed to measure 3D winds, such as Doppler wind lidar (DWL), radar, and infrared/microwave sensors that derive AMVs using feature-tracking of consecutive images. Therefore, an accurate and robust methodology for modeling uncertainty will allow for more accurate assessments of mission impacts, and the eventual propagation of data uncertainties for these instruments.

Deleted:) and

Deleted: quantification methodology

Velden and Bedka (2009) and Salonen et al. (2015) have shown that height assignment contributes a large component of uncertainty in AMVs tracked from cloud movement and from sequences of infrared satellite radiance images. However, with AMVs obtained from water vapor profiling instruments (e.g., infrared and microwave sounders), height assignment error cannot be directly assessed purely through analysis of the AMV extraction algorithm. Height assignment is instead an uncertainty in the water vapor profile itself. Unfortunately, without the quantified uncertainties on the water vapor profile necessary to pursue such a study, that is well beyond the scope of this paper. As such, this study will focus on errors in the AMV estimates at a given height. Previous work has demonstrated several different approaches for characterizing AMV vector error. One common approach is to employ quality indicator thresholds, as described by Holmund et al (2001), which compare changes in AMV estimates between sequential timesteps and neighboring pixels, as well as differences from model predictions, to produce a quality indicator to which a discrete uncertainty is assigned. The Expected Error approach, developed by Le Marshal et al. (2004), builds a statistical model using linear regression against AMV-radiosonde values to estimate the statistical characteristics of AMV observation error.

Deleted: height assignment is not the dominant portion of the error in…

Deleted:).

Deleted: with

Deleted: correct

In this study, we outline a data-driven approach for building an AMV uncertainty model using observing system simulation experiment (OSSE) data. We build on the work by Posselt et al. (2019) in which a water vapor feature-tracking AMV algorithm was applied to a high-resolution numerical simulation, thus providing a global set of AMV estimates which can be compared to the reference winds produced by the simulation. In this case, a synthetic "true" state is available with which AMVs can be compared and errors are quantified, and it is shown that errors in AMV estimates are state dependent. Our approach will use a conjunction of machine learning (random forest) and unsupervised parametric clustering (Gaussian mixture models) to build a model for the uncertainty structures found by Posselt et al. (2019). The realism and robustness of the resulting uncertainty estimates depend on the realism and representativeness of the reference dataset. This work builds upon the work of Bormann et al. (2014) and Hernandez-Carrascal and Bormann (2014), who showed that wind tracking could be divided into distinct geophysical regimes by clustering based on cloud conditions. This study supplements that approach with the addition of machine learning, which, compared with traditional linear modeling approaches, should allow the model to capture more complex non-linear processes in the error function.

Deleted: detail

Deleted: tool

Deleted: tracking

Deleted: by

Traditional techniques for modeling uncertainty through machine learning have focused on characterizing bias but often struggle when estimating standard error. By pairing a random forest algorithm with unsupervised parametric clustering, we propose a data-driven, cluster-based approach for quantifying both bias and standard error from experimental data. According to the theory developed by Nguyen et al. (2019), these improved error characterizations should then lead to improved error characteristics (e.g., lower bias, more accurate uncertainties) in subsequent analyses such as flux inversion or data assimilation.

Deleted: quantification

This paper does not purport that the specific algorithm detailed here should supplant error characterization approaches for all AMVs; indeed, most commonly assimilated AMVs are based on tracking cloud features, not water vapor profiles. In addition, this algorithm is trained and developed for a specific set of AMVs extracted from a water vapor field associated with a particular range of flow features. As such, application of our algorithm to modeled or observed AMVs will be most appropriate in situations with similar dynamics to our training set. However, we intend in this paper to demonstrate that the methodology is successful in characterizing errors for this set of water vapor AMVs and suggest that this approach— that is, capturing state-dependent uncertainties in feature-tracking algorithms through a combination of clustering and random forest— could be implemented in other feature-tracking AMV extraction methods and situations.

The rest of the paper is organized as follows: In Section 2, we give an overview of the simulation which provides the training data for our machine learning approach. We then motivate and define the specific uncertainties this study aims to characterize. In Section 3, we describe the error characterization approach with the specifics of our error characterization model, including both the implementation of and motivations for employing the random forest and Gaussian mixture model. In Section 4, we provide a validation of our methods, attempting to assess the bias of our predictions. In Section 5, we discuss the implications of our error characterization approach, both on AMV estimation and data assimilation more broadly.

Deleted: and

2. Experimental Set-up

2.1 Simulation and Feature-Tracking Algorithm

Formatted: Font: Not Bold, Font color: Text 1

We trained our model on the simulated data used by Posselt et al. (2019), which applied an AMV algorithm to outputs from the NASA Goddard Space Flight Center (GSFC) Global Modeling and Assimilation Office (GMAO) GEOS-5 Nature Run (G5NR; Putman et al. 2014). The Nature Run is a global dataset with ~7 km horizontal grid spacing that includes, among other quantities, three-dimensional fields of wind, water vapor concentration, clouds, and temperature. Note that throughout the text we will use the term ‘Nature Run wind’ to refer to reference winds in the simulation dataset used to train the uncertainty model. The AMV algorithm is applied on four pressure levels (300hPa, 500hPa, 700hPa, and 850hPa) at 6-hourly intervals, using three consecutive global water vapor fields spaced one hour 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 local-area pattern matching algorithm that approximates wind speed and direction from movement of the matched

Deleted: While our methodology in principle could be used to quantify uncertainties in any measurements used in data assimilation, in this paper we devote special emphasis to the use case of wind-tracking algorithms. In particular, we

Deleted: in

Deleted: they

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

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

Deleted: The

Deleted: (or, at the very least, the pressure level uncertainty

Deleted: captured by the satellite measurement uncertainty rather than the AMV estimate).

A snapshot of the dataset at 700hPa is given in { REF _Ref29398327 \h * MERGEFORMAT }, where we display the water vapor from Nature Run (top left panel), the wind speed from Nature Run (top right panel), the tracked wind from the AMV-tracking algorithm (bottom right panel), and the difference between the Nature Run and tracked wind (bottom left panel). Note that the wind-tracking algorithm tends to have trouble in region where the Nature Run water vapor content is close to zero. It is clear that while the wind-tracking algorithm tends to perform well in most regions (we can classify these regions as areas where the algorithm is skilled), in some regions the algorithm is unable to reliably make a reasonable estimate of the wind speed (unskilled). We will examine these skilled and unskilled regimes (and their corresponding contributing factors) in section 3.{ **INCLUDEPICTURE "https://lh4.googleusercontent.com/Bxx2AuV0Yv_LyfydtO30hkD9PeGug6p_AMNp7hKH4ZIU9SY-rZBzlPepaT-fAG51TilWVrFM0KlfBkBZLjfWQbubq8aSFsxKvRu0LGALEH-cNQpJeJ1qvxE6Dimat5t6hP2UFfCK" * MERGEFORMATINET** }

Deleted: true

Deleted: true

Deleted: true

Deleted: true

Deleted: are

Deleted: the

2.2 Importance of Uncertainty Representation in Data Assimilation

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

$$\mathbf{J} = (\mathbf{x} - \mathbf{x}_\mathbf{b})\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_\mathbf{b}) + \left((\mathbf{y} - \mathbf{a}) - \mathbf{H}[\mathbf{x}]\right)^{\mathrm{T}}\mathbf{R}^{-1}\left((\mathbf{y} - \mathbf{a}) - \mathbf{H}[\mathbf{x}]\right) \quad (1)$$

where $\mathbf{x}$ represents the analysis value, $\mathbf{x}_\mathbf{b}$ represents the background field (first guess), $\mathbf{B}$ represents the background error covariance, $\mathbf{y}$ represents the observation, and $\mathbf{H}$ represents the forward operator that translates model space into

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Bold

observation space. This translation may consist of spatial and/or temporal interpolation if **x** and **y** are the same variable (e.g., if the observation of temperature comes from a radiosonde), or may be far more complicated (e.g., a radiative transfer model in the case of satellite observations). **R** represents the observation error covariance, and **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 (**y** – **H**[**x**] - **a**) and the contribution from the prior (**x** – **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 $\breve{y}$ is a biased and noisy estimator of the true wind **y**, and might be assumed to follow the model Eq. (2):

$$\breve{y} = y + \epsilon \tag{2}$$

where ϵ is an error term, commonly assumed to be Gaussian with mean **a** and covariance matrix **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 **a** and the standard error matrix **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 { REF _Ref29398351 \h * MERGEFORMAT }. Given a set of training predictors X, training responses Ŷ, 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, Ŷ, 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 Ŷ. This step produces an estimate of the true response (we call this $\breve{Y}$) using only the training predictors and response. We then employ the Gaussian mixture model to estimate the clusters which the set of X, Ŷ, and $\breve{Y}$ pertain to. Subsequently, we compute the error characteristics of each cluster of X, Ŷ, and $\breve{Y}$ in the training dataset. Thereafter, given a new point consisting solely of X and Ŷ, 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

Formatted: Font: Bold
Formatted: Font: Bold
Formatted: Font: Bold
Formatted: Font: Bold
Deleted: -most part
Formatted: Font: Bold
Formatted: Font: Bold
Deleted: $\breve{y}$
Formatted: Font: Bold
Deleted: -1
Deleted: the
Deleted: Therefore
Deleted: This forms the basis for our error characterization model.…

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)).

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

3.2 Error Regime

When examining the relationship between AMVs and Nature Run winds in { REF _Ref29398366 \h * MERGEFORMAT }, it is clear that there are two distinct 'error-regimes' present in the dataset. The majority of AMV estimates can be categorized as 'skilled', wherein their estimate lies clearly along a one-to-one line with the Nature Run wind. However, there is also clearly an 'unskilled' regime, for which the AMV estimate is very close to zero when there are actually moderate or large Nature Run wind values present. Our goal is to provide unique error characterizations for each error regime, because the error dynamics are different within each regime. Furthermore, when we analyze this error and its relationship to water vapor, we see that 'unskilled' regime correlates highly with areas of low water vapor in { REF _Ref29398395 \h * MERGEFORMAT }. This matches the error patterns discussed in Posselt et al. (2019). We note that the division between skilled and unskilled regimes does not need to be binary. For instance, in some regions the wind-tracking algorithm might be unbiased with high-correlation with the true winds, and in other regions the algorithm might still be unbiased relative to the true winds, but with higher errors. The second situation is clearly less skilled than the first, although it might still be considered 'skilled', and this separation of the wind-tracking estimates into various 'grades' of skill forms the basis of our error model.

3.3 Gaussian Mixture Model

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

1. Define each location containing Nature Run winds, water vapor, and AMV estimates as a random variable x_i
2. Define θ as the population that consists of all x_i in the training dataset

Deleted: sample

Deleted: sample

Deleted: simulated true

Formatted: Font color: Auto

Deleted: simulated true

Formatted: Font color: Auto

Deleted: high

Deleted: mid-level true

Formatted: Font color: Auto

Deleted: is capable of determining whether any

Deleted: belongs in the 'skilled' or 'un-skilled' cluster

Deleted: , since we see the relationship between the error-regimes and water vapor content.

Deleted: a

Deleted: simulated true

3. Model the distribution of the population $P(\theta)$ as:

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

Where $N(\mu_j, \Sigma_j)$ is the normal distribution with mean μ_j and covariance Σ_j of the *j*-th cluster, K is the number of clusters, and π_j is the mixture proportion.

4. Determine π_j, μ_j, Σ_j for K clusters using the Expectation–Maximization Algorithm
5. From 3. and 4., estimate the probability of a given x_i belonging to the j-th cluster as $P(x_i \in k_j) = p_{ij}$
6. Assign point x_i to the cluster with the maximum probability p_{ij}

The mixture model clustering is based on the R package 'Mclust' developed by Fraley et al. (2012), which builds upon the theoretical work of Fraley and Raftery (2002) for model-based clustering and density estimation. The process uses an Expectation-Maximization algorithm to cluster the dataset, estimating a variable number of distinct multivariate Gaussian distributions from a sample dataset. Training the Gaussian mixture model on this dataset provides a clustering function which outputs a unique cluster for any data point with the same number of variables.{ INCLUDEPICTURE "https://lh5.googleusercontent.com/hTyZLCheUtO1nax6OCT-PrH_M3_pNyiVsopj3jyTJr8CNcKeTT_GkQa80dAOLFIWvxtdQ9EApaE5c8G2WyjlYfk1ihxPSxClrO5xtAz0LgG2ToHP00myCbV6YGIEMXnwpn1FE5n6" * MERGEFORMATINET }

In one dimension, a Gaussian mixture model looks like the distributions depicted in { REF _Ref29398417 \h * MERGEFORMAT }: instead of modeling a population as a single distribution (Gaussian or otherwise), the GMM algorithm fits multiple Gaussian distributions to a population. One key aspect of this algorithm is the capability of assigning a new point to the most likely distribution. For example, in the 1-D figure, a normalized AMV estimate with a value of 10 would be more likely to originate from the broad cluster '2' than the narrow cluster '4'. In this case, we model the population as a Gaussian mixture model in five-dimensional space, which consists of two Nature Run wind vector components (u and v), two AMV estimates of these wind components (u and v), and the simulated water vapor values, all of which have been standardized to have mean 0 and standard deviation of 1. Each cluster has a 5-dimensional mean vector for the center and a 5x5 covariance matrix defining their multivariate Gaussian shape. The estimation of a covariance matrix allows for the characterization of the relationships between the different dimensions within each cluster, and as such the gaussian mixture model approach provides greater potential for understanding the geophysical basis of error regimes than other unsupervised clustering approaches.

We note that the choice of inputs to the clustering methodology is limited, and that a more successful clustering may be achieved by including additional meteorological or geographic information. However, the intention of this paper is to study the ability of a purely data-driven approach, where no additional information or assumptions are passed to the machine learning model outside of the inputs and outputs to the AMV algorithm itself. Posselt et al. (2019) showed

Deleted: An Expectation–Maximization Algorithm determines…

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Tab after: 0.5" + Indent at: 0.5"

Deleted: Density estimation gives us

Deleted: Maximum p_{ij} is the assigned cluster for

Deleted: modelling

Deleted: A

Deleted: is that

Deleted: has

Deleted: simulated true

Deleted: .

Formatted: Font color: Black

Deleted: In { REF _Ref29338668 \h * MERGEFORMAT }, we applied the Gaussian mixture model to true u and v wind data using 9 clusters. Although { REF _Ref29398395 \h * MERGEFORMAT } indicates that the data tends to separate roughly into 'skilled' and 'unskilled' regimes, we opted to choose 9 clusters in the Gaussian mixture model after several sensitivity tests across all pressure levels found 9 to be the minimum number of clusters needed to ensure the separation into these separate regimes, as well as allowing for further stratification of sub-regimes within the skilled and unskilled regimes¶

that state dependent uncertainties are a major source of error in water vapor AMVs; introducing further information may cloud our ability to discern these specific uncertainties. While scaling this methodology to other applications may incentivize tailoring to specific conditions, this paper aims to demonstrate that modifications are encouraged for improvement, but not necessary for success.

Having trained the Gaussian mixture model on the 1.5 month training dataset, we applied the clustering algorithm to a testing dataset sampled from the subsequent 0.5 months of the nature run. By re-analyzing the AMV estimate in relation to the Nature Run winds within each cluster ({ REF _Ref29338668 \h * MERGEFORMAT }), we find that the clustering approach successfully separates the AMV estimates according to their 'skillfulness'. Essentially, we repeat Figure 3 but divide the AMV estimates by cluster. We see that, for example, clusters 4, 5, and 7 clearly represent cases in which the feature-tracking algorithm provides an accurate estimate of the Nature Run winds, with very low variance around the one-to-one line (i.e., low estimation errors). Clusters 1, 2, 3, and 9 are somewhat noisier than the low-variance clusters, with error characteristics similar to those of the entirety of the dataset. That is, they are considered less skilled, but their estimates still lie on a one-to-one line with respect to the true wind. Clusters 6 and 8, on the other hand, are clearly unskilled in different ways. Cluster 6 is a noisy regime, which captures much of the more extreme differences between the AMV estimates and the Nature Run winds. Cluster 8, on the other hand, represents the low AMV estimate, high Nature Run wind regime. This cluster is returning AMVs with values of zero where the Nature Run wind is clearly non-zero because of the very low water vapor present. We further see the stratification of the regimes when analyzing the absolute AMV error in relation to the water vapor content (Figure 7). We see that clusters that have similar behaviors in the error pattern (such as 1, 2, and 3) represent different regimes of water vapor content.

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

Deleted: simulated true

Deleted: that each point has been assigned to

Formatted: Font color: Auto

Deleted: , separated into the

Deleted: true

Formatted: Font color: Auto

Deleted: estimate

Deleted: simulated true

Formatted: Font color: Auto

Deleted: true

Formatted: Font color: Auto

Deleted: true

Moved down [1]: We see that the clustering algorithm succeeds in capturing physically interpretable clusters without having any knowledge of the underlying physical dynamics. We further see

Moved (insertion) [1]

of this parameter in their modeling, although this tuning process should be greatly simplified by the inclusion of an information criterion (e.g., BIC) in the GMM algorithm.

3.5 Random Forest

The clustering algorithm requires the Nature Run wind vector component values (u and v) in order to classify the AMV error. When applying the algorithm in practice to tracked AMV wind from real observations, the true winds are unknown. To represent the fact that we will not know the true winds in practice, we develop a proxy for the Nature Run winds using only the AMV estimates and the simulated water vapor itself. This is an instance in which the application of machine learning is desirable, since machine learning excels at learning high-dimensional non-linear relationships from large training datasets. In this case, we specifically use random forest to create an algorithm which predicts the Nature Run wind values as a function of the tracked wind values and water vapor.

Random forest is a machine learning regression algorithm which, as detailed by Breiman (2001), employs an ensemble of decision trees to model a nonlinear relationship between a response and a set of predictors from a training dataset. Here, we chose random forest specifically because it possesses certain robustness properties that are more appropriate for our applications than other machine learning methods. For instance, random forest will not predict values that are outside the minimum and maximum range of the input dataset, whereas other methods such as neural networks can exceed the training range, sometimes considerably so. Random forest, due to the sampling procedure employed during training, also tends to be robust to overtraining in addition to requiring fewer tuning parameters compared with methods such as neural networks.

We trained a random forest with 50 trees on a separate set of tracked winds and water vapor values to predict Nature Run winds using the 'randomForest' package in R. While the random forest estimate as a whole does not perform much better than the AMV values in estimating the Nature Run wind (2.89 RMSE for random forest vs 2.91 RMSE for AMVs), as shown in { REF _Ref29394704 \h * MERGEFORMAT }, it does not display the same discrete regimentation as the AMV estimates in Figure 3. As such, the random forest estimates can act as a proxy for Nature Run wind values in our clustering algorithm — they remove the regimentation which is a critical distinction between the AMV estimates and the Nature Run wind values.

3.6 Finalized Error Characterization Model

The foundation of the error characterization approach is to combine the random forest and clustering algorithm. We apply the Gaussian mixture model, as trained on the Nature Run winds (in addition to the AMVs and water vapor), to each point of water vapor, AMV estimate, and associated random forest estimate. This produces a set of clusters which, when implemented, require no direct knowledge of the actual Nature Run state ({ REF _Ref29394987 \h * MERGEFORMAT }).

Deleted: true
Formatted: Font color: Auto
Formatted: Font color: Auto
Deleted: Therefore
Deleted: true
Formatted: Font color: Auto
Deleted: true
Formatted: Font color: Auto
Deleted: certainty
Deleted: true
Formatted: Font color: Auto
Deleted: true
Formatted: Font color: Auto
Field Result Changed
Deleted: Relative to the AMV estimates, the error in each of the random forest estimates is closer to the mean of error of the entire dataset.
Deleted: true
Formatted: Font color: Auto
Deleted: true
Formatted: Font color: Auto
Deleted: true
Formatted: Font color: Auto
Deleted: true
Formatted: Font color: Auto
Deleted: We

Naturally, the clustering algorithm performs better when applied to the dataset with the Nature Run winds, as opposed to winds generated from the random forest algorithm. The former is created with direct knowledge of the Nature Run winds, and any approximation will lead to increased uncertainties. In practice, the performance of the cluster analysis can be improved by enhancing the performance of the random forest itself. As with any machine learning algorithm, the random forest contains hyperparameters that can be optimized for specific applications. In addition, performance could be improved by including additional predictor variables. Our intent is not to use the random forest as a wind tracking algorithm; rather, the random forest is presented in this paper as a proof of concept.

Nonetheless, we see in Figure 13 and Figure 14 that the error characterization still discretizes the testing data set into meaningful error regimes. The algorithm manages to separate the AMV estimates into appropriate error clusters. Once again, clusters 6 and 8 manage to capture unskilled regimes, and cluster 7, and to a lesser extent clusters 4 and 5, remain skillful. By taking the mean and standard deviation of the difference between AMV estimates and Nature Run winds in each cluster, we develop error characteristics for each cluster ({ REF _Ref29395022 \h * MERGEFORMAT }); these quantities are precisely the bias and uncertainty that we require for the cost function J in Eq (1). We see that the unskilled clusters have very high standard errors and they correspond roughly to the areas of unskilled regimes in Figure 3. Similarly, skilled clusters 5, 4 and 7 have standard errors below that of the entire dataset. Since each cluster now has associated error characteristics (e.g., bias and standard deviation), it is then straightforward to assign the bias and uncertainty for any new tracked wind observation by computing which regime it is likely to belong to.

Deleted: that the

Deleted: extremely

Deleted: While there is some degradation in the performance relative to the classification algorithm on the training set, we see in Figure 9 and Figure 10 that the error characterization still discretizes the testing data set into meaningful error regimes.¶

Deleted: true

3.7 Experimental Set up

In this section we will describe our experimental setup for training our model on the GEOS-5 Nature Run data and testing its performance on a withheld dataset. We divide the dataset into two parts: a training set consisting of the first 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 training/testing procedure for the simulation data and tracked wind is as follows:

1. Divide the simulation data and tracked wind into two sets: training set of 1,000,000 points from the first 1.5 months of the Nature Run and a testing set of 1,000,000 points from the final 0.5 months of the Nature Run.
2. We train a Gaussian Mixture Model on a normalized random sample of observations from the training dataset of Nature Run winds (u and v direction), tracked winds (u and v direction), and water vapor with n=9 clusters.
3. We train two separate random forests on a different random sample of 750,000 observations from the training dataset. We use tracked wind (u and v direction) and water vapor to model, separately, Nature Run winds in both the u and v directions.
4. We apply the random forests to the dataset used for the Gaussian Mixture Model. This provides a random forest estimate for each point, which is used as a substitute for Nature Run wind values in the next step.
5. We predict the Gaussian mixture component assignment for each point of water vapor, tracked winds, and random forest estimate using the GMM parameters estimated in Step 2.

Deleted: Using the ‘density.Mclust’ function, we

Deleted: true

Deleted: true

Deleted: true

Deleted: Using the ‘predict.Mclust’ function, we

6. We compute the mean and standard deviation of the difference between the tracked winds and the Nature Run winds, per direction, for each Gaussian mixture model cluster assignment. This provides a set of error characteristics that are specific to each cluster.
7. We can apply the random forest, and then the cluster estimation, to any set of water vapor and tracked AMV estimates. Thusly, any set of tracked AMV estimates and water vapor can be mapped to a specific cluster, and therefore its associated error characteristics.

4 Results and Validation

In this section, we compare our clustering method against a simple alternative, and we quantitatively demonstrate improvements that result from our error characterization. Recall that in Section 3, we divided the wind-tracking outputs into 9 regimes, which range from very skilled to unskilled. For the *i*-th regime, we can quantify the predicted uncertainty estimate as a gaussian distribution with mean m_i and standard deviation σ_i, which has a well-defined cumulative distribution function which we denote as F_i. To test the performance of our uncertainty forecast, we divide the dataset described in Section 2 into a training dataset (first 1.5 month) and a testing dataset (last 0.5 month). Having trained our model using the training dataset, we apply the methodology to the testing dataset, and we compare the performance of the predicted probability distributions against the actual wind error (tracked winds - Nature Run winds). This is a type of probabilistic forecast assessment, and we assess the quality of the prediction using a scoring rule called continuous ranked probability score (CRPS), which is defined as a function of a cumulative distribution function F and an observation x as follows:

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

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 is positive or zero, and it is equal zero if the argument is negative (Gneiting and Katzfuss, 2014) . The continuous rank probability score here is strictly proper, which means that the function $\mathrm{CRPS}(F, x)$ attains the minimum if the data x is drawn from the same probability distribution as the one implied by F. That is, if the data x is drawn from the probability distribution given by F, then CRPS(F, x) < CRPS(G, x) for all G ≠ F.

The alternative error characterization method that we test against is a simple marginal mean and marginal standard deviation of the entire tracked subtract Nature Run wind dataset. This is essentially equivalent to an error characterization scheme that utilizes one regime, where m and σ are given as the marginal mean and the marginal standard deviation of the residuals (i.e., tracked wind minus Nature Run winds). Here, we use a negatively oriented version of the CRPS (i.e., Eq.(4) without the minus sign), which implies that lower is better. A histogram evaluating the performance of our methodology against the naive error characterization method is given in { REF _Ref29398184 \h * MERGEFORMAT }.

Deleted: true

Deleted: each

Commented [MOU1]:

Deleted: via a 95% confidence interval, which in the Gaussian case can easily be constructed

Deleted: [x_i - 2 σ_i, x_i+ 2 σ_i], where x_i the predicted

Deleted: σ_i is the predicted

Deleted: of the i-th cluster.

Deleted: bias

Deleted: confidence interval

Deleted: confidence intervals

Deleted: true

Deleted: in this paper

Deleted: ,

Deleted: probabilistic forecast F (here represented by our confidence interval)

Deleted: CPRS(F, x)

Deleted: .

Deleted: CPRS

Deleted: maximum

Deleted: probabilistic forecast

Deleted: ≤

Deleted: track - true

Deleted: datasets

Deleted: and its confidence interval similarly could be constructed as [x – 2 σ, x + 2 σ], where x

Deleted: true

Deleted: , to evaluate

Deleted: . We plot the histogram of the scores

The relative behavior of the CRPS is consistent between u and v winds. The CRPS tends to have to wider distribution when applied to the regime-based error characterization. Compared to the alternative error characterization scheme, our methodology produces a cluster of highly accurate predictions (low CRPS scores), in addition to some cluster of very uninformative predictions (high CRPS scores). These clusters correspond to the highly skilled cluster (e.g., Cluster 3) and the unskilled clusters (Cluster 6 and 8), respectively. Overall, the mean of the CRPS is lower for our methodology than it is for the alternative method, indicating that as a whole our method produces a more accurate probabilistic forecast.

Thus far we have shown that our method produces more accurate error-characterization than an alternative method based on marginal means and variance. Now, we assess whether our methodology provides valid probabilistic prediction; that is, we test whether the uncertainty estimates provided are consistent with the empirical distribution of the validation data. To assess this, we construct a metric in which we normalize the difference between the Nature Run wind and the tracked wind by the predicted variance. That is, for the *i*-th observation, we compute the normalized values for u_i and v_i using the following equations:

$$z_{u,i} = \frac{u_i - ʉ_i}{\sigma_{u,i}}$$

$$z_{v,i} = \frac{v_i - ʋ_i}{\sigma_{v,i}} \quad (5)$$

Where u_i is the *i*-th Nature Run u wind from the Nature Run data, $ʉ_i$ is the tracked-wind, and $\sigma_{u,i}$ is the error as assessed by our model (recall that it is a function of the regime index to which $ʉ_i$ has been assigned). The values for the v-wind are defined similarly. The residuals in Eq (5) can be considered as a variant of the z-score, and it is straightforward to see that if our error estimates are valid (i.e., accurate), then the normalized residuals in Eq. (5) 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 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 of $z_{u,i}$ and $z_{v,i}$ should be larger than 1. In *{ REF _Ref45710459 \h * MERGEFORMAT }*, we display the histogram of the normalized residuals z_u and z_v. 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 that our error characterization model is highly accurate when forecasting uncertainties.

5 Conclusion

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

Deleted: likely

Deleted: 5

Deleted: true

Deleted: u

Deleted: v

Deleted: $z_u = \frac{u - ʉ}{\sigma_u}$ z_v

Deleted: u

Deleted: true

Deleted: ʉ

Deleted: σ_u

Deleted: .

Deleted: In { REF _Ref29398184 \h * MERGEFORMAT },

Deleted: methodology produces

Deleted: (std = 1.003 and 1.009 for u and v, respectively).

Formatted: Font color: Black

Deleted: Uncertainty quantification, which is the quantification of an imperfect or incomplete state of knowledge within a model,

Deleted: a

Formatted: Font: Bold

mischaracterization of bias will systematically offset a tracked wind, while an erroneous standard error could incorrectly weigh the cost function.

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

At its most general, our methodology consists of two parts: an emulator and a clustering algorithm. In this implementation, random forest and Gaussian mixture modelling are the approaches; in theory, these two steps could be accomplished using other algorithms belonging to the appropriate class. Indeed, improvements to the methodology could surely be made with further research in both areas. Given the degradation in the uncertainty estimates between those produce with and without Nature Run wind values, an improvement of the emulator could yield the most efficient returns. This could either take the form of improving or replacing the random forest algorithm. As previously discussed, improvements could also be made in both the inputs to and nature of the clustering algorithm.

We note that our methodology requires knowledge of the true field of interest (here u and v), or some proxy thereof. This makes our methodology a natural supplement to OSSE-based studies where the true fields of interest are provided by numerical weather models. Such studies are important components of algorithm validation (e.g., Posselt et al., 2019), and our proposed methodology provides a framework for characterizing the error within different geophysical regimes. In practice, we envision that the lack of true fields could be addressed by either using independent validation data or reanalysis model data. Therefore, there would be an additional component of error due to the usage of proxy data in lieu of the true field, but this error should be inversely proportional to the quality of the proxy data.

The error estimation algorithm presented in this paper is a proof of concept. While the methodology is expected to be generally applicable to other AMV extraction methods (e.g., cloud-track winds or tracking of other trace gases), our specific error functions are only valid for our specific training dataset. That said, the state-dependent errors identified by Posselt et al. (2019) are also expected to apply to other water vapor AMVs. This is because, in general, AMV algorithms have difficulty tracking fields with very small gradients, and will produce systematic errors in situations for which isolines in the tracked field (e.g., contours of constant water vapor mixing ratio) lie parallel to the flow. To the extent that our algorithm represents a general class of errors, the results may be applicable to other geophysical scenarios and other AMV tracking methodologies. As mentioned in the introduction, robust estimates of uncertainty

Deleted: assimilate an incorrect

Deleted: weight

Deleted: ,

Deleted: develop an error-characterization scheme based on random forest and mixture model clustering. Here, the mixture of a parametric approach and a

Deleted: method allows us to combine the flexibility of machine learning with the interpretability of mixture modelling in an entirely

Deleted: . In theory, the fidelity of our method should scale with the number of training data observations, making the methodology well-suited for the massive datasets that are typical within remote sensing applications. Our error function has been applied to an AMV OSSE study using GEOS5…

Deleted: its impact will be reported in a forthcoming paper.¶ We demonstrate that our methodology produces accurate error estimates (also called validity), and that it is able to identify and remove the biases within the wind-tracking algorithm's …

Deleted: . Particularly, the methodology is able to identify unskilled regimes that are physically meaningful — in our case, unskilled regimes related to regions of near-zero water vapor content. We note that our methodology is able to find this dependence between unskilled regimes and low water content without any prior knowledge or specification from the user…

Deleted: true

Deleted: While we position the methodology

Deleted: an error characterization tool, this property also makes it useful as an exploratory tool to aid

Deleted: understanding the distribution of multivariate and potentially complex data.

Deleted: Our algorithm

Deleted: Future research includes replacing random forest with other machine learning methods such as neural networks or support vector machines, and investigating other methods of clustering, such as self-organizing networks. We note that the issue of bias removal in data assimilation and in remote sensing is certainly not limited to atmospheric motion vectors. The methods we have used to characterize uncertainties in AMVs are general, and can be applied to other inverse problems as well are important for data assimilation, and we expect that our methodology could be used to provide more accurate uncertainties for AMVs used in data assimilation for weather forecasting and reanalysis. To do so, we recommend training on a dataset that is large enough to encompass the full range of features likely to be seen by the assimilation and forecast system. To the extent that errors may be seasonally and regionally dependent, it may be more effective to train the error estimation algorithm on data that is expected to represent the specific flow regimes and water vapor features valid for a particular forecast or assimilation period. We have tested the error function described in this paper in an AMV weather forecast and data assimilation OSSE study using GEOS5, and its impact will be reported in a forthcoming paper.

Author Contribution

Teixeira conceived of the idea with inputs from Nguyen. Teixeira performed the computation. Wu provided the experimental datasets along with data curation expertise. Posselt and Su provided subject matter expertise. All authors discussed the results. Teixeira wrote the initial manuscript and updated the draft with inputs from co-authors.

Deleted: and Nguyen

Deleted: .

Deleted: .

Deleted: . All authors contributed to

Deleted: subsequent

Competing Interest: The Authors declare no conflict of interest.

Funding Acknowledgment: The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). © 2020. California Institute of Technology. Government sponsorship acknowledged.

Formatted: Font: Bold, Font color: Auto, Pattern: Clear

References

Deleted: ¶
¶

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

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

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

Formatted: Font color: Black

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

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

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

Formatted: Font color: Custom Color(RGB(34,34,34)), Pattern: Clear (White)

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

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

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

Kawa, S.R., Erickson, D.J., Pawson, S. and Zhu, Z.: Global CO2 transport simulations using meteorological data from the NASA data assimilation system, Journal of Geophysical Research: Atmospheres, 109, { HYPERLINK "https://doi.org/10.1029/2004JD004554" }, 2004.

Kwon, Y., Won, J.H., Kim, B.J. and Paik, M.C., 2020. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics & Data Analysis*, *142*, p.106816.

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

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

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

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

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

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

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

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

Tran, D., Dusenberry, M., van der Wilk, M. and Hafner, D., 2019. Bayesian layers: A module for neural network uncertainty. In *Advances in Neural Information Processing Systems* (pp. 14660-14672).

Tripathy, R.K. and Bilionis, I., 2018. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of computational physics*, *375*, pp.565-588.

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

Formatted: Font color: Text 1

Formatted: Font color: Text 1

Formatted: Left, Space Before: 0 pt, After: 0 pt

Deleted: H.,

Formatted: Font color: Text 1, Pattern: Clear

Formatted: Font color: Text 1, Pattern: Clear

Formatted: Font color: Text 1, Pattern: Clear

Deleted: N.,

Formatted: Font color: Text 1, Pattern: Clear

Deleted: , J.:

Deleted: retrievals: Implications

Formatted: Font color: Text 1, Pattern: Clear

Formatted: Font color: Text 1, Pattern: Clear

Deleted: consequences when the prior's mean and covariance are misspecified, National Institute for Applied Statistics Research Australia Working Paper Series,

Formatted: Font color: Text 1, Pattern: Clear

Formatted: Font color: Text 1, Pattern: Clear

Formatted: Font: 12 pt, Font color: Text 1

Formatted: Font color: Text 1

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

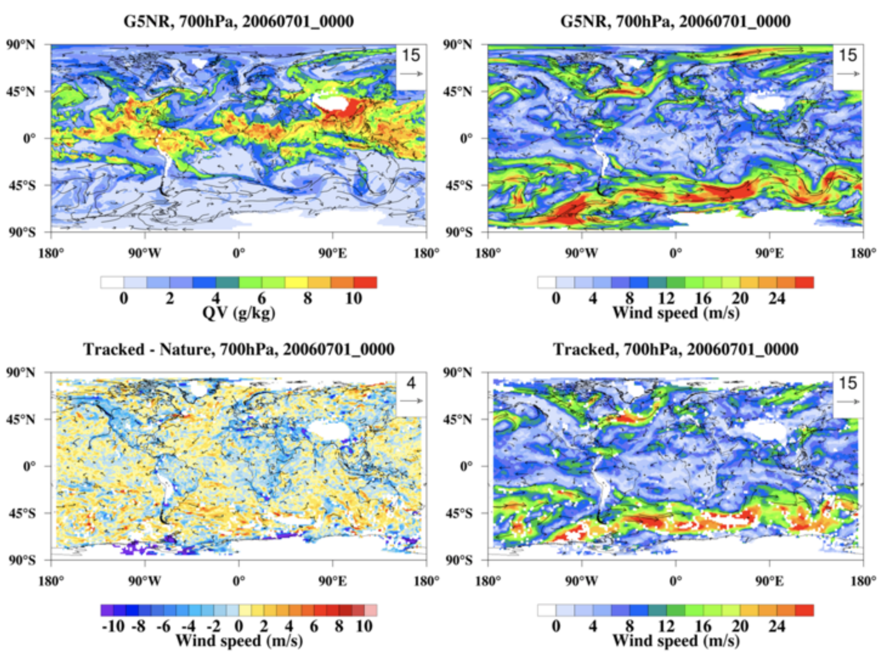


Figure { SEQ Figure * ARABIC }: Map of Nature Run at one timestep at 700hPa (A): Water Vapor (B): Nature Run Wind Speed (C): Difference between Nature Run Wind Speed and AMV Estimate (D): AMV Estimate.

Deleted: True

Deleted: True

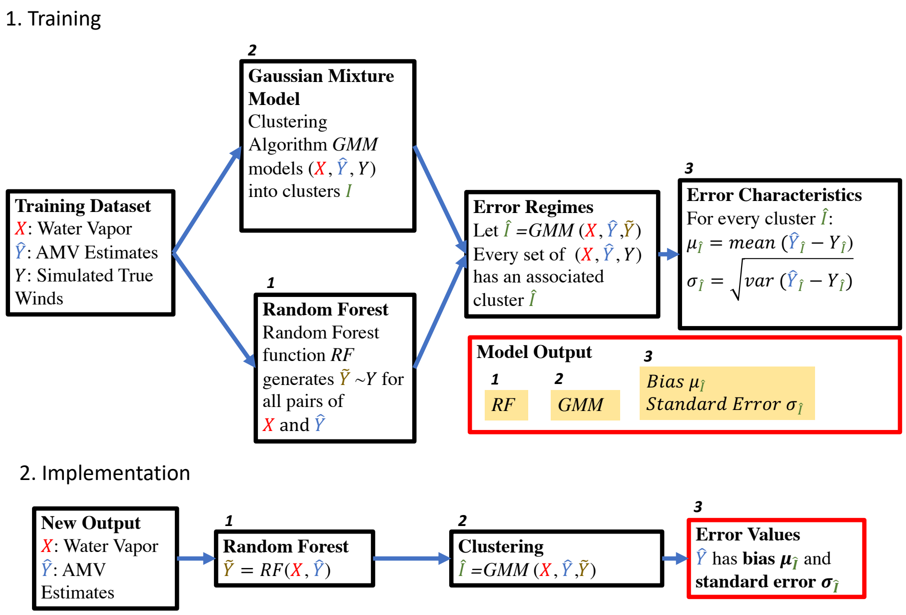


Figure { SEQ Figure * ARABIC }: Diagram of Training Approach and Diagram of Implementation steps.

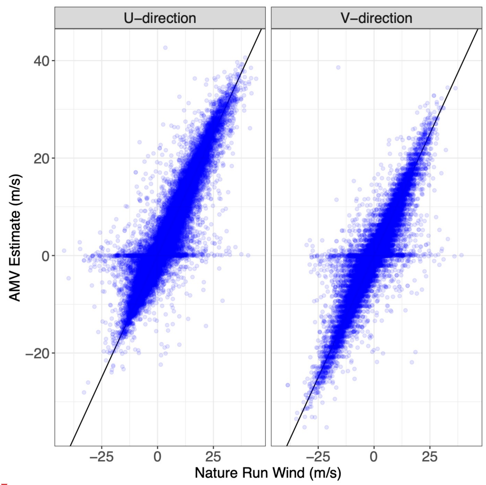


Figure { SEQ Figure * ARABIC }: Scatter plot of the simulated Nature Run wind vs AMV estimates for u and v wind in the training dataset.

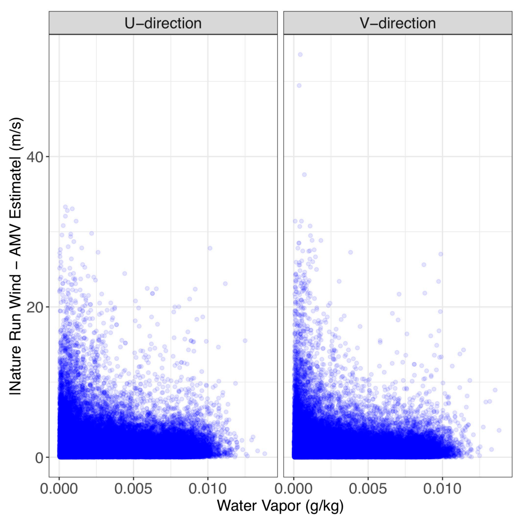
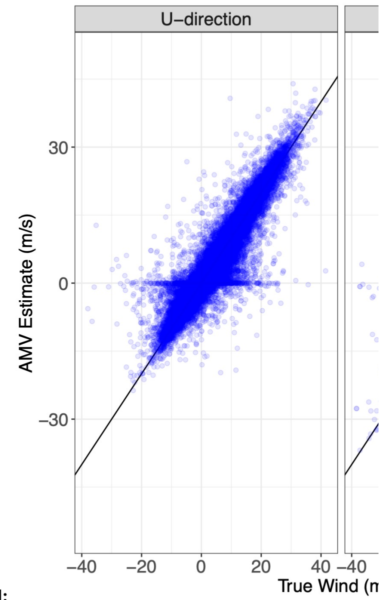


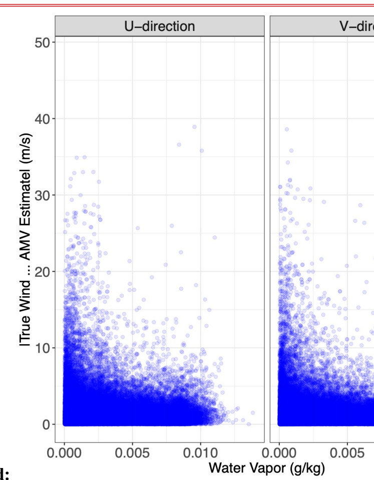
Figure { SEQ Figure * ARABIC }: Simulated water vapor vs the absolute value of the difference between Nature Run and tracked winds in the training dataset.

Deleted:



Deleted: true

Deleted:



Deleted: true

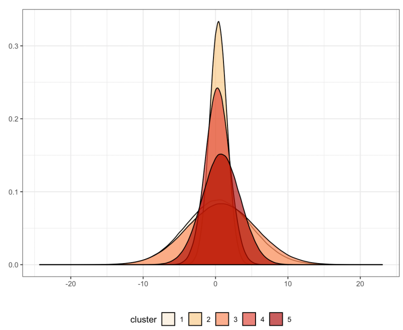


Figure { SEQ Figure * ARABIC }: Example of Gaussian Mixture Model in one dimension. Density Figures for the U-Direction AMV Estimate dimension of fitted Gaussian mixture.

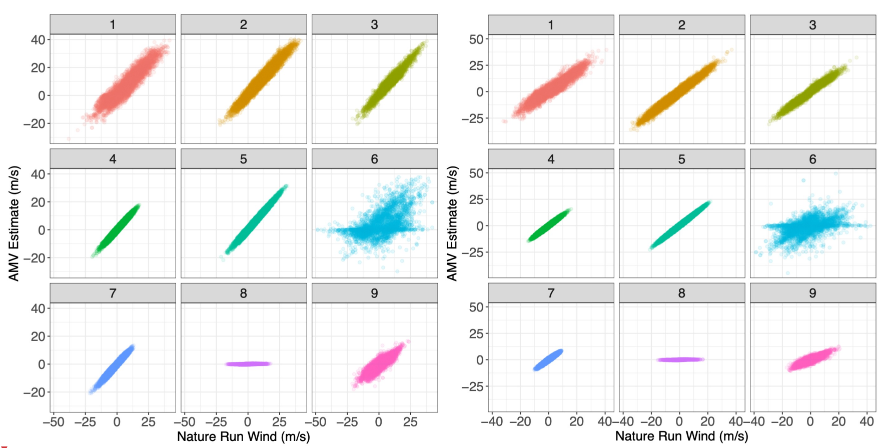
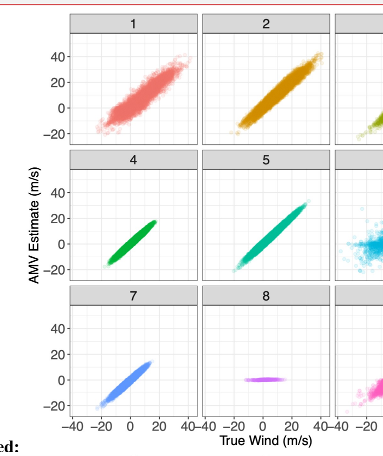
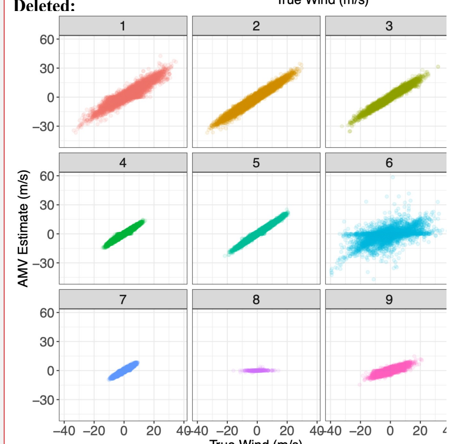


Figure { SEQ Figure * ARABIC }: Scatterplot of simulated Nature Run wind vs AMV Estimates, each sub-panel corresponding to the specific Gaussian mixture component to which each point in the testing set has been assigned. (A): U-Direction Wind (B): V-Direction Wind.



Deleted:



Formatted: Space Before: 0 pt, After: 10 pt, Line spacing: single

Deleted: true

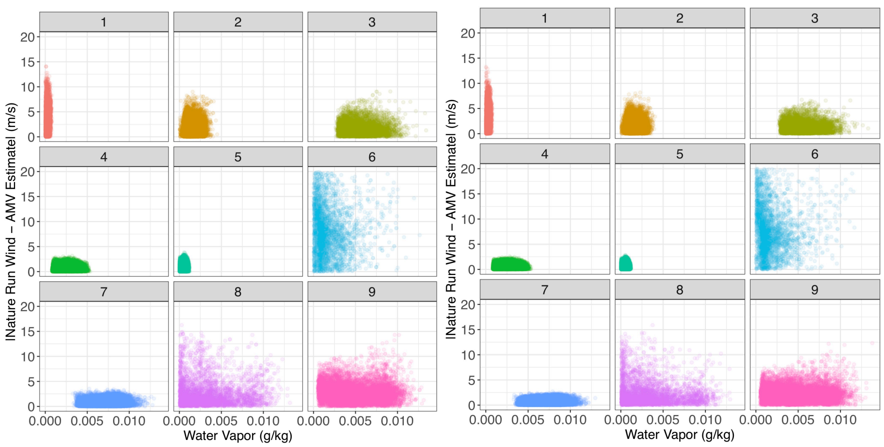


Figure { SEQ Figure * ARABIC }: Scatterplot of Water Vapor vs Absolute Tracked Wind Error, each sub-panel corresponding to the specific Gaussian mixture component to which each point in the testing set has been assigned. (A): U-Direction Wind (B): V-Direction Wind.

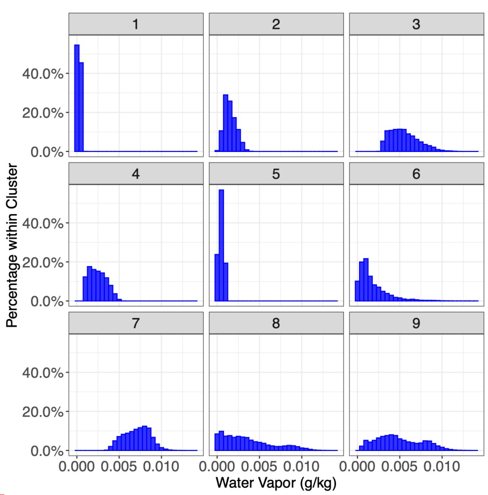
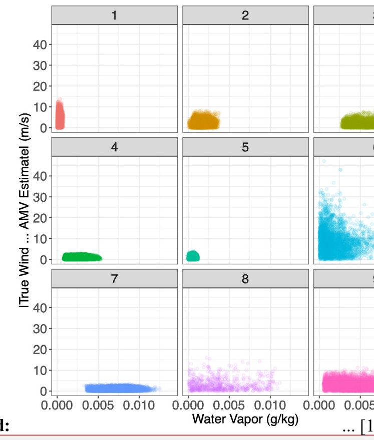
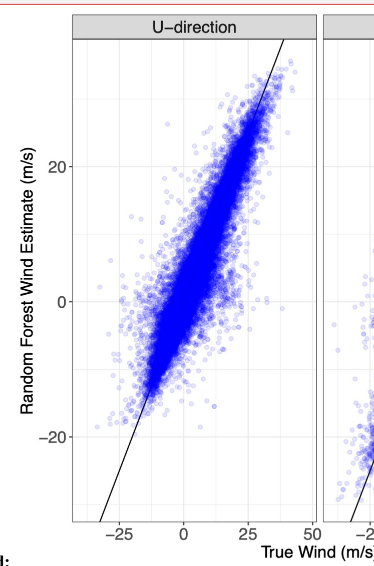


Figure { SEQ Figure * ARABIC }: Histogram of Nature Run water vapor for each cluster identified by the Gaussian mixture model, applied to the testing set. Each sub-panel represents the cluster each point was assigned to.



Deleted: ... [1]

Formatted: Space Before: 0 pt, After: 10 pt, Line spacing: single



Deleted:

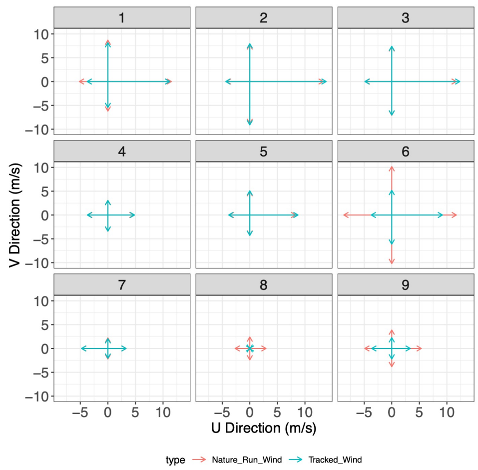


Figure { SEQ Figure * ARABIC }: Mean tracked winds and Nature Run winds, in each direction, for each cluster applied to the test set. Each sub-panel represents the cluster each point was assigned to.

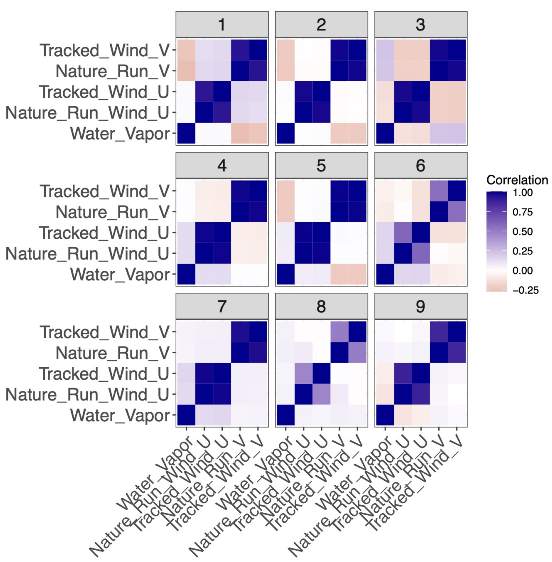


Figure { SEQ Figure * ARABIC }: Correlation matrix between each clustered element for each identified cluster in the original training dataset. Each sub-panel refers to a specific cluster.

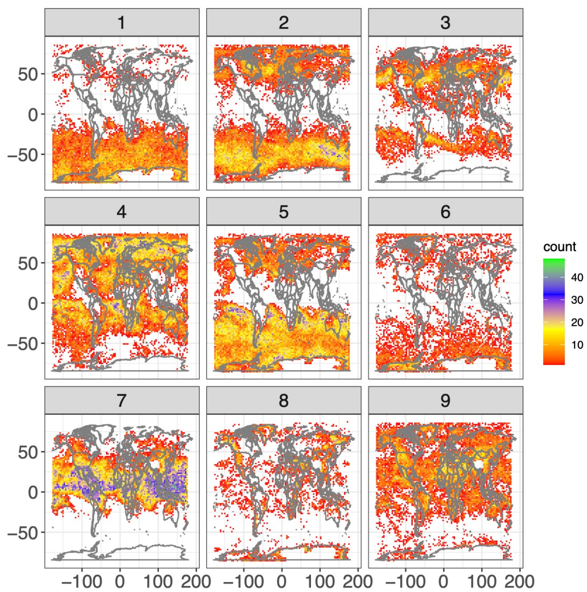


Figure { SEQ Figure * ARABIC }: Geographic distribution by cluster of AMV retrieval locations in the testing dataset. Each sub-panel represents one cluster.

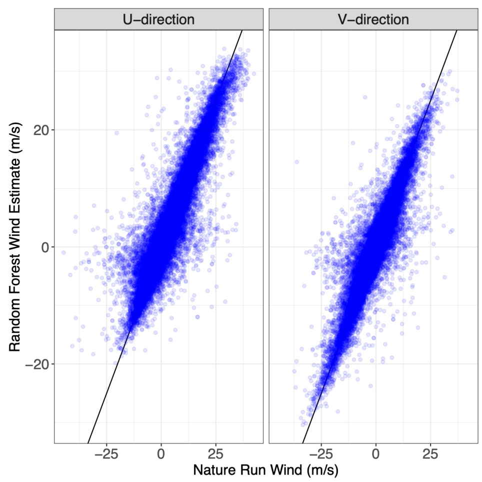
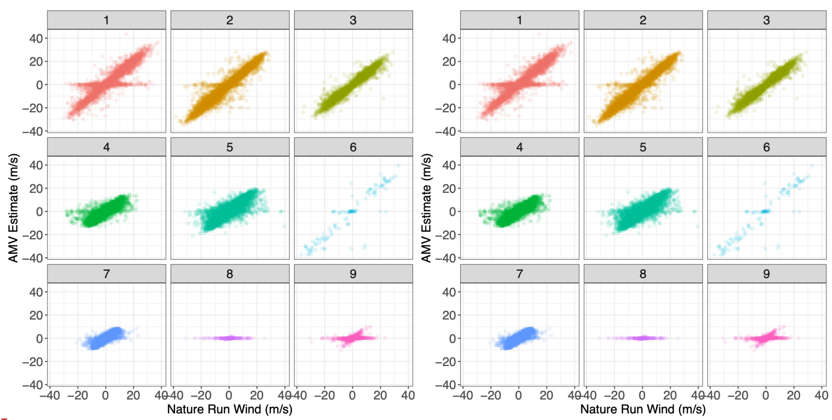
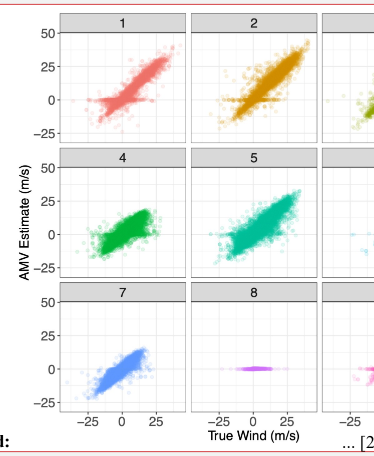


Figure { SEQ Figure * ARABIC }: Scatterplot of Nature Run wind estimate vs random forest produced estimate. (A): U Direction (B): V Direction

Deleted: true



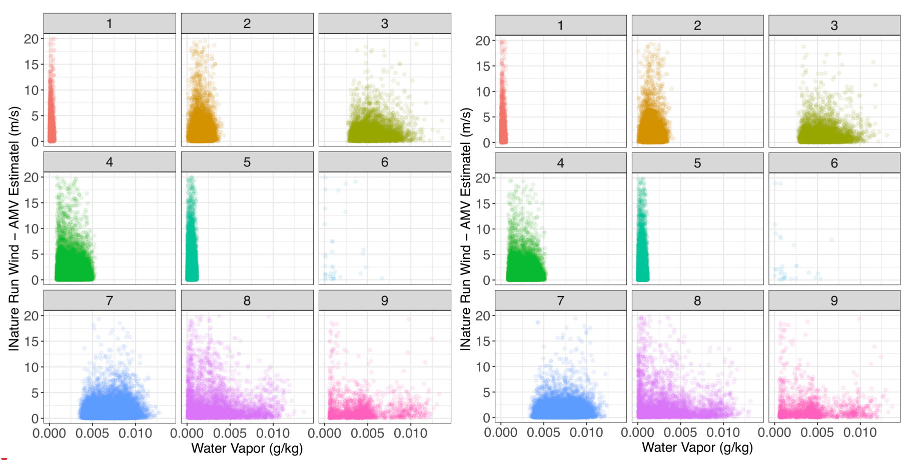


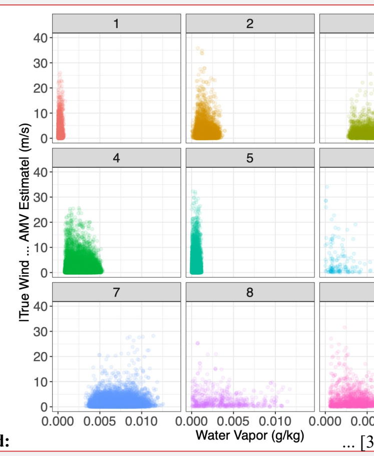
Deleted: ... [2]

Figure { SEQ Figure * ARABIC }: 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

Deleted: simulated true

Deleted: true





Deleted: ... [3]

Figure { SEQ Figure * ARABIC }: 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

Deleted: true

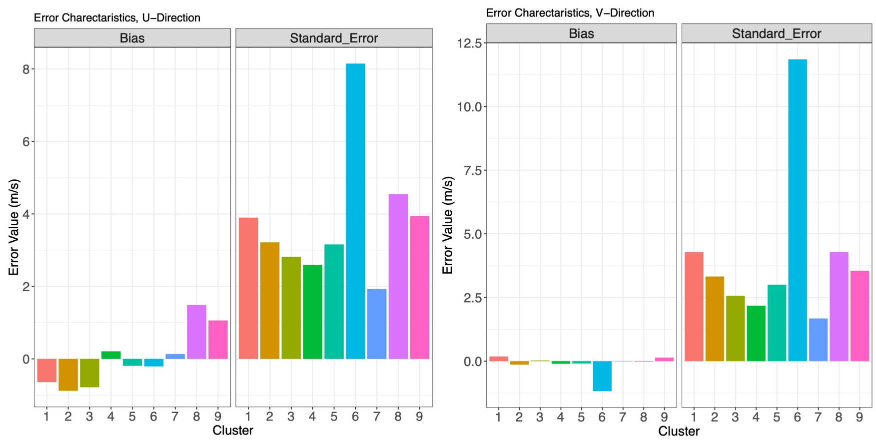


Figure { SEQ Figure * ARABIC }: (A): Bias (Left Panel) and Standard Error (Right Panel) for each Gaussian mixture cluster in figure 6, U direction. (B): Same as (A) for V-direction

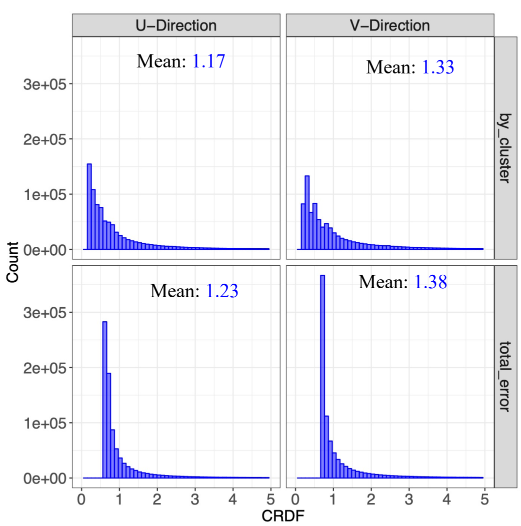
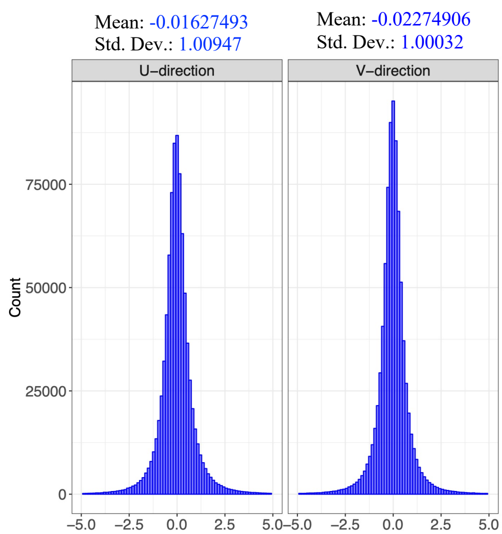


Figure { SEQ Figure * ARABIC }: CRSP applied to different error approaches. (A): Cluster Errors for U Winds (B): Total Errors for U Winds (C): Cluster Errors for V Winds (D): Total Errors for V Winds.



$$z_u = \frac{u - \hat{u}}{\sigma_u} \qquad z_v = \frac{v - \hat{v}}{\sigma_v}$$

Figure { SEQ Figure * ARABIC }: U and V winds normalized using the error characteristics developed by our methodology.

Deleted: Error Clusters

Page 20: [1] Deleted	**Microsoft Office User**	**8/11/20 5:22:00 PM**
Page 24: [2] Deleted	**Microsoft Office User**	**8/11/20 5:22:00 PM**
Page 24: [3] Deleted	**Microsoft Office User**	**8/11/20 5:22:00 PM**