Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2020-95-AC3, 2020

Atmospheric Measurement Techniques Discussions

Interactive comment on "Uncertainty Quantification for Atmospheric Motion Vectors with Machine Learning" *by* Joaquim V. Teixeira et al.

Joaquim V. Teixeira et al.

joaquim.p.teixeira@jpl.nasa.gov

Received and published: 11 August 2020

"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(Ŷi), 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(Ŷi), 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

a. Subset the data to retain only observations with regime index j. Let's assume that we have Nj observation within this data subset

b. Sample with replacement Nj observations from this subset. This forms a bootstrap sample

c. From 2., compute an estimate of the bias and std.

d. Repeat step 2-3 for 1000 times, giving us 1000 estimates of the bias and 1000 estimates of the std within regime j.

e. Compute 95% confidence intervals from the 1000 estimates of bias and std from 4.

The results for the confidence intervals in the attached Figure.

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 $\sim$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 "$\leq$" 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 surro-

gate models for high dimensional uncertainty quantification. Journal of computational physics, 375, pp.565-588.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2020-95, 2020.

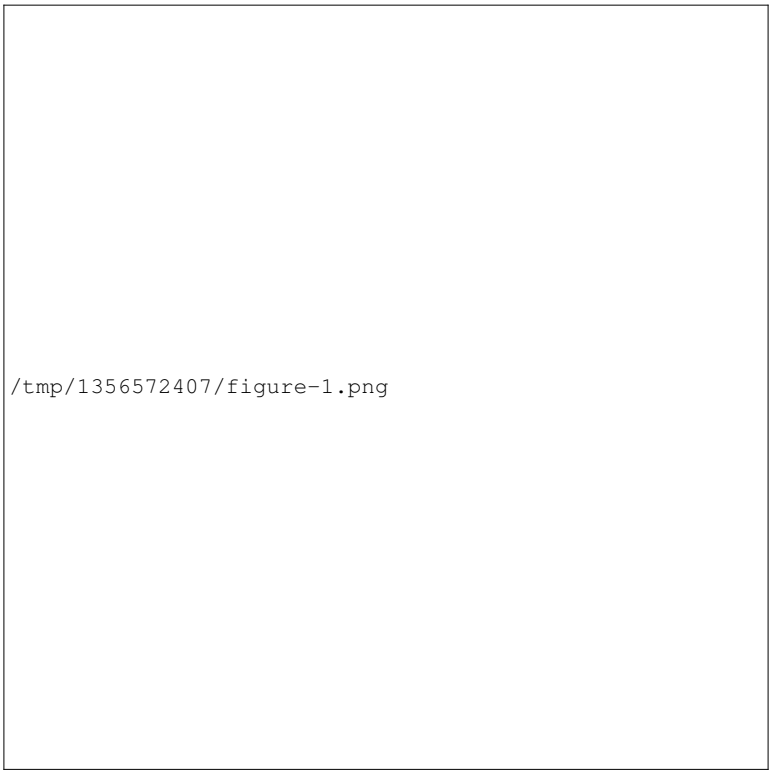

Fig. 1. Figure: Top rows (bias and std confidence intervals for u-wind), bottom rows (bias and std confidence intervals for v-winds). The interval represent a 95% confidence interval.

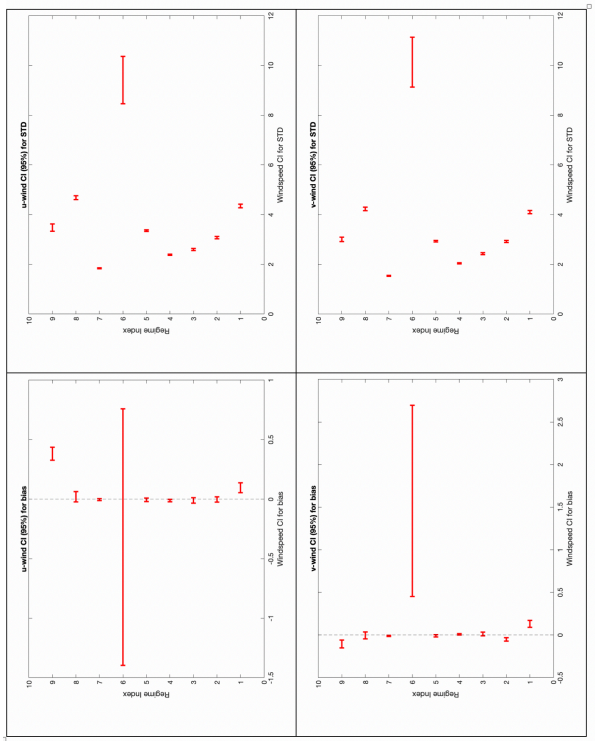

Fig. 2.