Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2020-95-RC2, 2020

Atmospheric Measurement Techniques Discussions

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

Anonymous Referee #2

Received and published: 25 June 2020

The paper summarises a study that uses a Gaussian Mixture Model to describe error characteristics of Motion Vectors derived from single-level model humidity fields. The Gaussian Mixture model is trained with the values of the "truth" available. To be able to apply this without the true wind field available, a random forest approach is applied to estimate the true wind from the derived Motion Vectors and an estimate of water vapour. The approach aims to provide situation-dependent estimates of typical errors in the Motion Vectors, and this is highly relevant for data assimilation applications. The manuscript is clearly written and mostly well-structured. While I'm not fully convinced about the practical applicability of the results for real-life observations, I feel the general methodology described may indeed offer novel approaches and hence merits publica-

tion, but the limitations and shortcomings need to be more critically assessed, as well as the physical motivation for some of the choices. Subject to addressing the points outlined below, I think this could be achieved in a major revision.

General points:

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.

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".

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

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.

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".

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.

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 Dirac-delta 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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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

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

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).

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.

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?

18. 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.

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