Review - AMV Error Characterization and Bias Correction by Leveraging Independent Lidar Data: a Simulation using OSSE and Optical Flow AMVs

Overall the proposed scheme is interesting and a useful contribution to the field. Up to and including the bias correction scheme it is well described. After that once we get into (I think) a proposed error estimation scheme I found it difficult to understand. I've picked out a few things that were unclear to me, if these things are explained in the text I'm happy to be corrected.

Quality of English is good, it would benefit from another look through by the authors to fix a few grammatical errors. Examples (not an exhaustive list):
Line 51 an bias-correction should be a bias correction
Line 88 AMV should be AMVs
Line 171: due -> due to
Line 193: track -> track of

Comments by section:

Section 2.1

In section 2.1 it would be worth explaining how the heights were chosen – traditional AMVs are concentrated around 850 and 200 hPa, so 300 hPa is a little low in the atmosphere and there are normally few quality AMVs at 500 hPa.

In the same section, line 115-6, why only conational obs + brightness temperatures to initialise, and not all observation types? And on line 114-5, why the 3rd most powerful member?

Line 121: I don't know a great deal about tropical cyclones but a drop of 0.4 hPa seems very small and not a 'plunge'

Section 2.2:
Although you refer to another paper, you could state here which traditional feature tracking algorithm you're referring to and AMV people will know its main strengths and weaknesses.

Line 128: what do you mean by dense AMVs? If they are derived for every pixel that already implies they are spatially dense, so are they dense in some other way?

What part of the NatureRun is used to derive the AMVs – the clouds, as in traditional AMV feature matching? Or the humidity surfaces, or something else?

The main source of error in AMVs is generally thought to be height assignment. If the optical flow is done on WRF humidity fields, then the error is only coming from the tracking and likely to be smaller than typical AMV errors

On figure 1 a scale would be helpful, to know how fast the winds are, and how large the differences are. The factor by which the differences are magnified should also be stated.

Section 3.1

I was a bit confused here, are we treating the WRF winds as truth and trying to bias correct towards WRF? Or are we simulating some Aeolus-style u-winds from WRF and bias correcting towards them? How are the lidar winds simulated? There should be some error associated with such lidar winds.

What is meant by wind-moisture gradient? Also, I've never heard the word heteroskedasticity and had to look it up, could you define it in the text or explain what you mean without using it? What is 'error residual'?

Figure 2 caption – mention that the labels are defined in Table 2

Table 3 caption, mention that the numbers in arrays are being searched through.

Section 3.2

Line 238-245: I appreciate that there is an attempt to explain MLP but there are lots of new terms that are not fully explained for example nodes and neurons, backpropagation, gradient descent

Line 250: what is the standard Euclidean distance?

Figure 3 – what is NR?

Table 4 – what is MarginalSTD and why are some numbers red?

Lines ~264-281: by this point we're talking about the bias of the optical flow winds (with or without bias correction applied) vs the WRF simulation truth? Is that right? Perhaps worth re-mentioning the first time 'bias' is mentioned.

Line 274: what is meant by state-of-the-art ensemble methods and neural networks? Do any of the methods being tested fit this description?

Table 5 – perhaps I missed this in the text, but why are they biases so much larger with the spatial validation than with the temporal validation

Section 3.3

Table 6 – looks like a caption for a non existent table

Table 7: does it say in the text somewhere which data is Marginal?

Normally, AMV quality indicators are given as an integer from 0 to 100. Why is a different scale used for these quality indicators?

Page 19: I found this page quite a struggle, I get that you are trying to come up with error estimates, but what are prediction intervals, nodes, 'leaf' etc? Is QRF quantile random forest or quantile regression forest?

Table 7 is quite a few pages away from where it is referred to, it would be good to move it closer.

Section 3.3.1

Line 398: Here it is suggested that the error estimation would only work when lidar data is coincident. I though the idea was to train a model using lidar data, that can then be used any time, anywhere? If the proposed scheme would only work when, for example, Aeolus data happens to be in the same place as AMVs then the scheme would only apply to a very small fraction of AMVs were it used operationally.

Section 4 (Conclusions)

Line 450: true AMVs = WRF NatureRun, correct? Better to say so