We would like to thank the two anonymous reviewers for their time and effort in responding to our manuscript. Their comments were quite helpful in refocusing the paper to improve the clarity of its message. We hope that our responses to their comments about the definition of “critical times” in flux records and gap filling models have resulted in a stronger paper that will be a significant contribution both for HESS and the community at large.

Reviewer 1:

General:
The result of the different sampling strategy is to have relatively more data points in the extremes of the dataset range. Is this really useful in a gapfilling application considering that these data points will represent (by definition) conditions not or less common? The cost would be to have less data points extracted in the rest of the range where the probability to have gaps are higher (because more common situations). In addition one should consider that the extreme values could be spikes so a good despiking method must be applied before...

Response: The “standardized” distribution does not move data toward the extremes but rather generates a new distribution where those extremes are moved closer to a central tendency (see Figure 2 for an illustration). By doing this, we make those data points more likely to be sampled when running an ANN (or any other) model. Because data are lost at a higher rate during the night, using this altered sampling makes those (low) extreme values more available during model training. The outcome of this study shows that, despite this transformation, model performance is not substantially improved. The abstract and discussion sections are modified to clarify this point.

Par 2.2.2:
The idea to use the Shannon index to evaluate the amount of information in the training dataset is interesting but the results shown that it is not related to the usefulness of this additional information to improve the ANN performances. Why the authors didn’t use artificial gaps (simpler to evaluate and compare with others studies and more related to the real results obtained in the LE simulations) to compare the performances of the different approaches? Or at least both Shannon index and artificial gaps.

Response: Our purpose in applying the Shannon index was to quantify how the model “sees” the data after it is sampled for training. The index suits this purpose well by showing a quantifiable change in the information available by breaking down the low end of the ET values into several bins (Table 1).

Our results are still comparable to other studies, because the real rate of data loss is similar to other flux tower sites. Artificial gaps would be useful to determine the performance of different gap filling strategies under various gap “regimes”, i.e. more short gaps or more long gaps. We thank the reviewer for pointing out the need to clarify our thought process in this section, to emphasize that the goal of this study was to examine information extraction, rather than to test model performance across gap *regimes. In the most prominent example of the use of artificial gaps [Moffat et al., 2007], the various gap schemes are applied on top of the existing (gappy) data record. As such, those model results represent model performance when gaps are generally short, medium long. Our study here is focused on performance in the existing gap regime. We revised the discussion in section 2.2.2 to better reflect our goals regarding data gaps.*

Model Performances:
in general, it is not clear which data have been used in the model performances analysis. Only a validation set? Are really independent data? Also here artificial gaps would have been useful.

Response: We have clarified the training and validation discussion is clarified in revisions to the manuscript (see the end of section 2.2.2). The data for year 2003 were split into training and test data (the test data identifies when training is complete). The model validation is done on the preceding two years (2001-2002)

P6528 - L15-17:
This would be true only if the turbulence is directly related to the fluxes (explaining variable) but it is in contrast with the definition of the u* threshold that is based on the opposite assumption (ecosystem respiration independent of u* at given temperature).

Response: While the fluxes themselves are independent of friction velocity as a process, their measurement, the assumptions inherent to EC require sufficient turbulence, quantified as friction velocity, to produce valid data. As such, data which are near the friction velocity threshold are important because they may have information about ET during times when data were lost. We have clarified this point in the new manuscript (see second-to-last paragraph in introduction).

P6529 - L23-26:
it seems that the meteorological data used are not registered at the site (1 km for net radiation and precipitation could be a lot) but in particular these are gapfilled using the mean diurnal variation method and this could affect strongly the performances due to the limits of the method.

Response: While we recognize that filling the input data using the mean diurnal variation method may introduce some performance issues, any gap filling on the meteorological input data would influence model performance. Because the MDV gap filled data are used as input for both ANN models, there is a reasonable basis for comparison in this study. The only radiation data used as model input are shortwave radiation, which is likely stable across moderate spatial extent.

P6531 – EQ2:
where are the two equations different?

Response: The error in the equation is corrected in the new manuscript.

P6532 – L11-18:
it is not simple to relate the text to table 1, I would suggest to reorganize the text and better explain the table.

Response: In response to this comment, we have revised the text and table in the new manuscript for ease of interpretation (see text in section 2.2.2).

Par 2.2:
it is not specified how may data points have been used in training, test and validation and how these have been extracted. In fact, a stratified sampling without changing the data distribution could be a good compromise.

Response: We revised the text to describe training and validation data sets used for model identification (see "Model Performances" comment).

P6538 – L6-9:
I don't agree. Mismatch between model and data can easily indicate problems in the model that doesn't work properly… However the problem of the eddy covariance technique when turbulence conditions are low is well known. For this reason it would be important to better explain how the data have been processed (how the authors estimated the u* threshold? How the storage correction has been applied?)

Response: Here we state that, especially near-dawn, both the data and models fail to capture the ET process. Because this period is typified by the onset of instability in the boundary layer, measurements dependent on turbulence will be weakly representative of ET until turbulence is better established. At the same time, a model trained to represent flux during turbulent behavior may yield inappropriate estimates of ET, because it "expects" stronger mixing than actually occurs.

Regarding the authors comments about data processing: because the data were used in a separate study, we applied the same filter as in that study (u<0.2). No storage term was used due to the open canopy structure on the site. We have clarified the discussion of model/data mismatch as well as processing in the revised manuscript (see section 2.1 and 4.1).*

P6538 – L23-25
Based on the results in Moffat et al. 2007 where the authors shown errors in the gapfilling very close to the random component of the measurements, I'm not sure that new model structures may have large impact on gapfilling results…

Response: The intent of our remarks in the paper about model structure here are intended more toward using different models to represent daytime and nighttime ET. Novick et al [2009] explore the use of different models only for nocturnal ET, while the models in [Moffat et al., 2007] apply the same models to day and night data. Turbulence in the boundary layer is a necessary driver of ET and to validate EC measurements. When turbulence is weak at night, applying both new models and new measurement *techniques would improve flux records. We have inserted text in the new manuscript to clarify this matter (see section 4.4).*

Table 1: in the caption explain the difference between Hrsc and Hstd
Fig. 3: Y-right Axes label missing
Fig. 5: legend missing
Fig. 6: X axes label missing

Response: We have edited the caption and figures in the new manuscript.

Reviewer 2:
Specific comments:

2.1.
This was the main conclusion of the study – so addressing the flux behavior at critical times is of important to formulating this conclusion. Logically, it is imperative to commence the manuscript by rigorously establishing these ‘critical times’. Statements such as “…valid data from periods of low turbulence, that is just above the filter threshold, are particularly valuable as a result (validity often established by a criterion such as friction velocity…” actually miss a number of crucial issues pertinent to the interpretation of a turbulent flux as an ET value. Implicit here is the use of ‘weak turbulence mixing’ as the indicator of critical times. The authors then report that some hours during the day appear more problematic and critical. Perhaps a more rigorous definition of what *should be* labeled as ‘critical times’ (abstract, discussion, conclusion) can benefit from the derivation below. The mean continuity equation for water vapor is given as

$$\frac{\partial \bar{q}}{\partial t} + \bar{U}_j \frac{\partial \bar{q}}{\partial x_j} = -\frac{\partial}{\partial x_j}\left(K_m \frac{\partial \bar{q}}{\partial x_j} + \overline{u_j' q'} \right), \tag{1}$$

where $\bar{q}$ is the mean water vapor concentration, U are the three components of the velocity, K_m is the molecular diffusion coefficient of water vapor in air, and $\overline{u_i'q'}$ are the turbulent fluxes in all three directions. Let us explore under what conditions the vertical turbulent flux in the atmosphere, as measured by an EC system, represents ET.

[1] In stationary conditions, $\partial q / \partial t = 0$
[2] In planar-homogeneous flows $\partial q / \partial x1 = \partial q / \partial x2 = 0$.
[3] In conditions with no mean subsidence, $W = U = 0$.
[4] Strong turbulence mixing—

$$\left| K_m \frac{\partial \bar{q}}{\partial x_j} \right| << \left| \overline{u_j' q'} \right| .$$

For those 4 assumptions, the budget equation in (1) reduces to

$$\frac{\partial}{\partial x_3}\left(\overline{u_3' q'} \right) \simeq 0 , \tag{2}$$

where upon integration with respect to height (x3 or z) yields:

$$\overline{w'q'} = const = ET \, . \qquad (3)$$

Condition [1] is likely to be violated precisely during 'transition times' such as during sunrise and sunset even if the friction velocity is large. Also, condition (1) is likely to be violated when the forcing variables (e.g. solar or net radiation) is changing rapidly in time – at least on time scales commensurate with the averaging times of the EC system.

Condition [2] is difficult to test – but if the footprint is fluctuating significantly – and the source of water vapor is not uniform (soil-vegetation), then there are good reasons to suspect this condition is violated (or exhibit a 'directional' ET based on the prevailing wind even if the meteorological and soil moisture conditions are the same). Exploring whether ET inference is sensitive to wind direction for the same mean meteorological and edaphic conditions is needed here to demonstrate that this is not an issue.

Condition [3] is likely to be violated when the ABL initially grows (i.e. dawn). Perhaps looking at the pdf(*w*) around dawn may provide clues about how important the subsidence is, with the usual caveat that sonic anemometry cannot resolve mean velocity smaller than 0.05 m s-1. However, if the authors find that around dawn, *w* is more like 0.1 or 0.2 m s-1, then this unquestionably indicates that EC based measurements are basically unrepresentative of ET.

High friction velocity alone does not guarantee that assumptions [1]—[4] are satisfied. So, defining critical times as conditions in which assumptions [1]—[4] are violated makes sense. Recall that ANN is inferring ET from meteorological data – and these critical times are times that EC measurements are not appropriate approximations for ET. ANN modeled ET in gap-filling is being convolved with conditions that may be correct for inferring ET from EC measurements and may be wrong at other times.

Response: The mass balance equation used to analyze eddy covariance data indicates specific conditions when data loss may occur. At a practical level, however, testing the four conditions outlined by the reviewer are often difficult. The use of a friction velocity filter is well-established as a method to identify periods when turbulent conditions are likely sufficient to yield valid EC data. Of the four conditions described by the reviewer, the zero-subsidence is easiest to test. Examining nighttime vertical wind speeds, less than 4% of the data show wind speed less than 0.2m/s, and nearly all of these (>95%) fail the friction velocity filter criterion. The failure rate according to a u filter at vertical wind speeds over 0.3m/s is greater than 70%. When nocturnal data with W > 0.2 was removed, model performance was not substantially different from the results shown in the original manuscript.*

Homogeneous lateral flow is discussed widely in the literature associated with advective flow, and even spatially homogeneous, near-flat landscapes exhibit a degree of drainage flow [Goulden et al., 2006]. This site includes only a single tower and minimal corrections are available to address advection flow. The filtered LE data does not appear *to be biased with regard to direction of flow, although flow from the some directions (e.g. northeast) is relatively infrequent.*

The question of stationarity in condition [1] is at the crux of the discussion of the paper, since model-data disagreement is strongest around dawn. The discussion is centered around the fact that these times, perhaps more than mid-night periods, should be the focus of future investigations and the importance of applying the appropriate data (likely to come from a source other than EC) to models.

Moreover, when we speak of "critical periods" in EC data collection, we wish to draw attention to data which are near the filter margin. Because these points often occur at night, when environmental conditions (stability, energy input) are notably different compared to daytime, those nighttime data which may be valid (and may come around dusk and dawn transitions) are of particularly high value. We thank the reviewer for elaborating a more rigorous definition of EC filter criteria. A more complete and explicit discussion of the assumptions in EC data filtering is part of the revised manuscript.

2.2 *Energy balance closure at critical times*: The authors may want to discuss the energy-balance closure at those critical times. How off was it compared to more 'micro meteorologically ideal times'? This is important given that net radiation is a key driver for the ANN model as well.

Response: Energy balance closure is consistent at midday and middle-night times. Midday energy balance was slightly more than 65% on average, while mid-night energy balance was slightly less than 50%. Dawn and dusk energy balance varied dramatically; over and under-estimation of the energy balance were greater than twenty times the measured available energy.

2.3 Also, if the gap-filled ET is used to estimate sensible heat flux, how well does the approach work?

Response: While the model was run for both latent and sensible heat fluxes at the study site, the differences in sensible heating were omitted from this paper for the sake of clarity. The two models predict sensible heat similarly, with a mean absolute error of 47 and 53W/m^2 for the rescaled and standardized models, respectively. Determining H as the residual energy results in mean absolute error of 126 and 132 W/m^2, respectively; this reflects a tendency of both models to underpredict ET. Overall, we hoped to avoid complicating the current manuscript by minimizing discussion of sensible heat flux.

2.4 *ANN and conclusions:*
After reading this manuscript, I am left with the desire to know how well the two ANN approaches here differ from standard approaches to ET gap-filling. Novick et al. (2010) already presented 5 approaches to gap-filling ET and compared their performances – these approaches can be readily employed here and compared to the two ANN approaches. How different are the results on annual ET estimates? This is essential to illustrating whether ANN is effective over other approaches or not, especially for such

types of ecosystems. A summary table (as Table 3 in Novick et al., 2010) can be most helpful.

Response: The "rescaled" ANN described in the paper is functionally the same as the "Standard ANN" case described in Moffat et al. 2007. We include a table summarizing annual and seasonal ET and a discussion of the ET patterns in the new manuscript (see section 4.2).

References

Goulden, M. L., S. D. Miller, and H. R. da Rocha (2006), Nocturnal cold air drainage and pooling in a tropical forest, *Journal of Geophysical Research-Atmospheres*, *111*(D8), 14.

Moffat, A. M., et al. (2007), Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, *Agricultural and Forest Meteorology*, *147*(3-4), 209-232.

Novick, K. A., R. Oren, P. C. Stoy, M. B. S. Siqueira, and G. G. Katul (2009), Nocturnal evapotranspiration in eddy-covariance records from three co-located ecosystems in the Southeastern US: Implications for annual fluxes, *Agricultural and Forest Meteorology*, *149*(9), 1491-1504.