Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-263-RC1, 2019

Geoscientific Model Development Discussions

Interactive comment on "Calculating the turbulent fluxes in the atmospheric surface layer with neural networks" *by* Lukas Hubert Leufen and Gerd Schädler

Anonymous Referee #1

Received and published: 4 January 2019

In my view this is overall an interesting manuscript that introduces a promising novel approach based on neural networks for calculating turbulent fluxes with MOST. With this novel approach, the authors aim amongst others to reduce the computational effort involved in current applications of MOST by circumventing the need for iteratively solving non-linear equations at each time step. Consequently, I believe that many weather and climate models could benefit from the proposed approach in this manuscript because MOST is often applied at the lowest vertical levels to calculate the turbulent fluxes. It is therefore promising that the authors showed that ANNs were able to produce results comparable to the standard method based on MOST.

It is important to note though that an actual decrease in computational time has not been demonstrated yet in this manuscript (as the authors acknowledged), because it required an additional extensive analysis of the proposed approach in at full 3d weather/climate model. Since to my knowledge this is one of the first attempts to apply ANNs to MOST, in my opinion the analysis and results currently presented in the manuscript are nonetheless sufficient for publication. The authors may consider to submit later on a second publication including this additional analysis.

I do have additional specific comments and suggestions regarding amongst others the methodology, citations and clarity of the text that I think should be addressed before publication (especially the comments related to Table 1 and the data used for training). First I will list my specific comments including the line and page number (note that the line numbers were erroneous in the submitted manuscript, and therefore I included the page numbers). After this, I give some grammatical/typographical errors I came across.

Specific comments and suggestions:

Page 1, line 3-7: the first part of this motivation for the ANN is in my view not convincing. The NNs rely on the same small amount of field experiments for training and consequently vary as well, just as the standard empirical MOST equations. NNs trained on 2 different sets of field experiments are unlikely to be the same, to a large extend because of the large uncertainty in the observational data. In short, I think this is more a problem related to the available data rather than the standard MOST method. Also, 'avoiding explicit formulas' can be considered to be actually a disadvantage since it makes it harder to interpret the outcomes from the NN.

Page 2, line 1: see comment above, NNs will likely vary as well depending on the training data.

Page 2, line 7: I think this goal is still a bit too vague. What do you precisely want to improve about the 'overall model performance'? Reducing the computational effort and potentially increasing the accuracy?

Page 2, line 10: I suppose this is observational data, which is not immediately apparent from the text.

Page 2, line 11-12: ANNs are actually non-linear regression models, so I think this statement needs to be rephrased. Furthermore, with 'lagged data' I suppose you mean 'time lagged data'.

Page 2, line 14-15: I think that using different architectures for different years is not feasible in practice as you seem to suggest here, because you do not know for years not present in the training set what architecture will work best. Which architecture performed best overall?

Page 2, line 16-17: See comment earlier, ANN is actually a non-linear regression model. To which (non-)linear regression models did they compare it?

Page 2, line 22-25: This seems to imply that the reduction of inherent variance by the ANN is an advantage. However, Gentine et all. call this an 'unintended side effect' and thus consider this actually to be a disadvantage. Furthermore, it is not clear here what is meant with 'inherent variance': inherent to what? Additionally, Gentine et all did not use one submodel, but rather thousands of cloud resolving models embedded within a climate model. Please rephrase these lines accordingly.

Page 2, line 8-25: I think it is worth mentioning some additional papers here that experiment with the application of ANNs in RANS/LES. LES relies often on MOST at the surface and therefore could be a potential application of your method. Furthermore, the analysis presented in these papers are to some extent in line with the suggestions you make at the end of the conclusions section (i.e using high-resolution model data like LES for training the NNs, which in this case inform lower-resolution models). Examples:

1. Ling, J., Kurzawski, A., & Templeton, J. (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. Journal of Fluid

Mechanics, 807, 155-166.

2. Vollant, A., Balarac, G., & Corre, C. (2017). Subgrid-scale scalar flux modelling based on optimal estimation theory and machine-learning procedures. Journal of Turbulence, 18(9), 854-878.

3. Sarghini, F., De Felice, G., & Santini, S. (2003). Neural networks based subgrid scale modeling in large eddy simulations. Computers & fluids, 32(1), 97-108.

Page 3, line 8: Replace 'quantity' with 'stability parameter' to make it consistent with the terminology used in chapter 1. Furthermore, θ should be θv.

Page 3, line 15: I suggest to clarify here immediately that these stability functions can be used to provide u* and θ* once the stability functions are known. This will connect it better to the previous paragraph.

Page 4, line 1-7: In these lines, you motivate again your approach. Similar to my comments before, I am not convinced that using NNs can reduce the uncertainty: it arises in my view mainly from the large scatter in the data. You can use of course, as you mention here, newer and larger datasets to reduce that uncertainty. However, I don't see the need for ANNs solely for this purpose: the standard method can also 'learn' from newer and larger datasets. I do agree that NNs are a much more flexible method than the standard one, requiring less assumptions about the functional form of the relationship. I see this as an important advantage of NNs, which you may also mention more explicitly in the abstract and introduction.

Page 4, line 12-13: This sentence is not fluent, please rephrase it. Also, you don't mention that the neurons are located in different layers (input layer, one or more hidden layers, output layer), which makes the text sometimes hard to follow. I think a visualization as in Figure 7 would be of great help here for the reader.

Page 4, line 14 equation: You forgot here to include the bias. You can either introduce a separate bias term in the equation, or explicitly specify in the text below that the

summation over the neurons includes the bias neurons. Also, I think you should remove the subscript k: I don't see the added value. Note also that the equation number is missing.

Page 4, line 15: Consider here to replace 'inputs' with something like 'number of neurons in the preceding layer (which includes the bias neurons)'. If you have more than one hidden layer, the input is the preceding layer, not the input layer as 'inputs' implies.

Page 4, line 20 'consistency with the activation functions': In what way? Do you mean that the inputs have an appropriate scale (0-1) for the activation functions?

Page 4, line 25 equation: remove superscript Ω, it seems to me that it has no specific meaning. If it has, please explain it.

Page 5, line 1-9: You could consider to move this to section 2.5, since you discuss it there in more detail. In this section (which focuses on explaining the NN), to me it seems to be out of place. Furthermore, be aware of the distinction between validation and testing (line 2). Rather, you have a training/validation phase and a separate testing phase. Later on, you do make this distinction correctly.

Table 1: Here you list a couple of parameters that are fixed during training based on recommendations from 2 books (please include those references also in the caption). Of those, the early stopping criterion is fixed at 50 training iterations (or do you actually mean 50 epochs?), meaning to me that the training is always stopped after 50 epochs regardless of the complexity of the network. Although this can help to prevent overfitting, it may also cause the training to stop too early, especially for the more complex networks that take much longer to train (and thus need more epochs). A common other approach is that early stopping is done automatically based on a certain stopping criterion which takes into account the training error and validation error. I suggest that you either implement such a 'automatic' early stopping, or that you show some additional plots with the training/validation error during training. The latter should make clear whether training is indeed stopped too early for the more complex networks or

not. Furthermore, I don't completely get what you mean with 'maximum number of training steps'. Does this mean that you stop training after 1000 steps/iterations, in a similar way as the already implemented early stopping criterion? If so, why did you include it besides the already implemented stopping criterion?

Page 5, line 18-20: Are NL-Cab and DE-Keh actually left out of the data used for the cross-validation? I am asking because you mention in line 17 that you wanted to test the ability of the NN to generalize to new stations not seen during training. DE-Keh is however mentioned in figure 1,2 and Table 2, which implies that it is part of the training data used. If De-Keh is in fact used during the cross-validation with random split (and thus seen during training), you cannot claim that you test the ability of the NN to generalize to new stations as done in Section 3.2. Or do you retrain the ANNs presented in Section 3.2 with a training set not including DE-Keh, without cross-validation? I think this needs to be clarified because it can otherwise undermine the analysis presented in Section 3.2.

Table 2: Indicate in the header that these are the stations used for training and validation (and hence not all of them) because only those were compatible with MOST, referring to Section 3.4 for further explanation and Table A1 for all stations. This was not clear to me at first.

Table 2 and A1: indicate also the observation period, such that the reader has a better idea about what data you are using.

Figure 1: Does not include all stations used. You may use separate sub-figures to show all of them. Note that Figure 2 does show all of them.

Page 5, line 21 'observation sites': I suggest that you specify here already that these sites consist of meteorological towers with varying height, such that the reader knows better from the start what kind of data you are using.

Page 5 (where counting from one starts), line 27-28: I suggest that you don't mention

here the input for the ANN, but focus more on the data description itself. It is also redundant because you mention it already in Section 2.5.

Section 2.3 & 2.5: I suggest to put Section 2.5 within 2.3 or vice versa. Both are part of your strategy to select the 'best' ANN, which is now 'interrupted' by the data description. In general, I think the clarity of Chapter 2 will benefit from some reorganization of the content. See also other suggestions.

Chapter 2: What batch size are you using? I cannot find it in this chapter.

Chapter 3, Page 7, line 16: This implies that early stopping is done automatically, while in Table 1 you mention it always stops after 50 iterations (or epochs?).

Figure 3: please indicate in the caption that these results are valid for an ANN with six inputs and one hidden layer under six-fold cross-validation.

Figure 4b: Indicate at bottom axis more clearly that the different numbers represent the two different hidden layers.

Page 7 (where counting from 1 starts), line 27: I think you can, based on Figure 4, make this statement even more general by replacing 'smaller error minima' with 'a lower MSE'. Not only the error minimum, but also for instance the median, doesn't always decrease.

Page 7 (where counting from 1 starts), line 28: I think you can better leave this statement out. When looking at the figures, it is questionable whether this is the case: for higher numbers of neurons, the minima in Figure 4 are around 0.002, while in Figure 3 they are around 0.0075. Also, I think the error minima are less relevant than other features shown in the plots (e.g. the median) since the minima are strongly influenced by outliers and are therefore not representative for the 'common' ANN.

Page 8, line 1 'These results show': In fact, only the observations made in the beginning of the paragraph support this claim. Therefore, I suggest to start here another paragraph and rephrase the sentence in a way like this: 'All in all, the comparison be-

tween Figure 3 and 4 shows that the station-wise data split reduces the performance substantially.'

Page 8, line 3: Replace 'portability' with 'generalization', this term is much more common in this context.

Page 8, line 7: Replace 'to transfer knowledge' with 'to generalize', see also comment above

Page 8, line 8: Replace 'transferability' with 'generalization', see comments above.

Table 3: The first column contains a lot of redundant information. I suggest to organize the table differently (grouping 'overall best' and 'best simple' together such you only need to mention it twice)

Page 8, line 9-23: I think it is also worth mentioning that the network with 7 inputs perform worse on the test set than the ones with 6 inputs.

Page 8 line 25-26: This sentence is not fluent and contains too much information. Consider to split this sentence into 2 or more sentences.

Table 4/5: Mention in the caption these results are from DE-Keh

Page 8, line 28-29: Remove second part sentence after ';' : it is redundant because it basically repeats the first part of the sentence.

Page 9, line 2-3: This sentence may suggest to some readers that the 7-5-2-2 ANN is favourable because MAE is lowest, even considering the next sentence. You probably don't intend to write this, so I think it is good to put here more emphasis on the fact that the discrete classifier behaviour is unwanted and thus the 7-5-2-2 ANN is not favourable to use.

Page 9, line 7: This is I think to a large extent caused by the lack of (high-quality) training data, which is worth mentioning here.

Section 3.2: In my view this section lacks a summary/conclusion (that deals with generalization as the header implies) and a coherent message the reader should remember. In Section 4 you argue that the 6-3-2 ANN is best in terms of accuracy and performance, so I think it would be good to mention that here already as well. Note that you did properly include such an intermediate summary in Section 3.1.

Page 9, line 13-15: If I interpret this sentence literally, it says to me that the reference version contains the ANN rather than the standard MOST method. I advise you therefore to rephrase this sentence.

Page 9, line 19 " 30" ": you probably mean here 30 seconds in time, not 30 arcseconds. To avoid any possible confusion, I suggest to simply write " 30s ".

Page 9, line 23: Note that DE-Fal is mentioned to be part of the training data (Table 2). Similar to my comment at Page 5, line 18-20, it is again a bit confusing to me what data is used for training and testing. Try to be more specific and less ambiguous about this. Furthermore, I also note that DE-Tha was not included in the previous analysis because it was not compatible with MOST (Section 2.4). This deserves an additional explanation why you can use this station in this analysis while you couldn't use it in the previous analysis.

Figure 7: With the purple points you probably indicate the bias neurons, please indicate this more explicitly

Page 10, line 10: In your paper I did not see any sensitivity study with respect to the training method. In fact, Table 1 mentions that the used training algorithm was fixed. Therefore, you cannot claim that you did this.

Page 10, line 14-15: Sentence is not very fluent. My suggestion is to say something like: "In view of the trade-off between ... and .... , ..."

Page 10, line 26-29: Despite that a couple of papers already attempted this to some extent (e.g. Gentine et al. 2018 as cited in your manuscript, see previous comment at

Page 2, line 8-25 for more papers), I agree that still a lot of work needs to be done in this regard and therefore your (future) contributions can be very relevant. Maybe you can cite some of these papers here (again) as they can make your suggestions more credible (since it shows that researchers are already attempting it).

Grammatical/typographical errors and suggestions:

Overall: use either 'nonlinear' or 'non-linear', be consistent in present/past tense and active/passive form (at least within individual sections)

Page 1, line 3: remove '(resp Earth system) models', add something like 'climate and weather forecast models'

Page 1, line 18: remove brackets, write something like 'climate and weather forecast models'

Page 1, line 19: replace 'fluxes respective covariances' with 'fluxes'

Page 2, line 2: replace 'resp' with 'and/or'

Page 2, line 5-25: Be consistent with present and past tense.

Page 2, line 20: is –> is given in

Page 2, line 28: Consider to put the parts between brackets as a footnote.

Page 4, line 16: A network becomes non-linear –> In a network non-linearity is included by

Page 4: equation number missing

Page 5, line 17: 2x on –> to

Page 6, line 2: was to be seen –> was present

Page 6, line 9: as follows –> as follows : , or remove 'as follows'.

Figure 3 and 4 label: MSE[1] –> MSE

Figure 4 caption: whiskers have –> whiskers indicate, remove second ‘each’

Page 7, line 21: of of –> of

Page 7, line 23: results of ANNs trained ... vary –> the quality of the results from ANNs trained ... varies

Page 7, line 25: remove first ‘,’

Page 8, line 4: overfitting –> overfit

Table 3 caption: too –> Also, ...

Section 3.2: sometimes in activate form, mostly in passive form. Try to be consistent.

Page 8, line 30: Also for heat flux, –> Regarding the heat flux,

Page 9, line 16-18: Remove all ‘,’ except the third one.

Section 4: active and passive form used interchangeably, try to be consistent.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-263, 2018.