Review of 'Using wavelet transform and dynamic time warping to identify the limitations of the CNN model as an air quality forecasting system' by Eslami et al.

In this paper, Eslami et al. present a method based on wavelet transform and dynamic time warping (DTW) to characterize the quality of a machine-learning (ML) algorithm (convolutional neural network, CNN) for air quality forecasting (AQF). Using the example of two AQF applications, they show how wavelet transform and DTW can provide new insights into the strengths and weaknesses of the CNN model.

Better understanding the potential and limitations of ML algorithms for AQF applications is a topic that is rapidly gaining importance given the explosion of ML applications in this area. This paper makes a valuable contribution to this discussion by presenting a powerful analytical tool that can effectively highlight conditions under which the employed ML algorithm fails to produce satisfactory results. As such, the manuscript is highly suitable for publication in GMD. However, in its current form there are still some issues regarding the main message of the paper and how wavelet transform and DTW can be used to improve error characterization of ML applications.

For instance, the authors simultaneously say that the tested CNN models have 'significant limitations' and 'show promising accuracy', and generally seem to switch between the view that the ML model is either 'bad' or 'good'. In reality, the CNN models – like chemical transport models – perform very well under some conditions and poorly under others. One of the powerful elements of the discussed statistical analysis tools is that they offer a method to identify these conditions and thus help the model developers better understand the strengths and limitations of the ML algorithms. This information also helps identify how the ML model might be improved, which is very powerful. The authors should stress this more clearly.

Another point that needs more discussion is the time dimension. The used CNN models seem to use snapshots of time-series data as inputs (rather than a window of the time-series) and are thus not designed to learn temporal relationships. This should be stated more clearly, as it means that the wavelet transform and DTW offer an assessment of a feature that is not directly optimized by the ML algorithm (which is a good thing).

Minor comments:

- Page 4, line 100: 'general inability of the machine learning model' seems a bit too harsh. I suggest to rephrase this.
- Page 5, line 124. Should be Figure 1, not Figure 3.
- Page 6, line 201: Please provide the definition of index of agreement
- Page 6, line 213: I'd be careful with the statement that NOx and VOC emissions are constant in time. These emissions have large diurnal and seasonal cycles.
- Page 7, line 251ff: maybe worth mentioning here the potential of long short-term memory (LSTM) algorithms to incorporate time dependency in the training?