

I'd like to thank the reviewer for their thoughtful comments on the manuscript. Below, the reviewer's comments are displayed in blue and the author responses are in black.

"The manuscript claims to have designed a system for atmospheric chemical ensembles in general, but the actual scope of this work is much narrower: the model produced here is bound to the sole ammonia species in a north American domain. Furthermore, while the author have cleverly taken advantage of the short atmospheric lifespan of ammonia, the design of the model in this paper, i.e. a diagnostic model independent of the background concentration of ammonia (formula (5)) do not generalize easily to a broader scope such as other more long-lived chemical species, or, less even, chemical ensembles in general. The manuscript should better reflect the actual scope of this method."

I agree that the generalized language in the pervious version of the manuscript was a poor choice. The title of the paper has been changed to "EnsAI: An Emulator for Atmospheric Ammonia Ensembles". In the abstract,

"a new AI-based ensemble generation system for atmospheric chemical constituents"

has been changed to

"a new AI-based ensemble generation system for surface ammonia concentrations"

In the introduction, instances of "tropospheric chemical constituents" or "chemical constituents" were changed to "atmospheric ammonia" or "atmospheric ammonia surface concentrations".

"This model has been exclusively trained on one year of data on a northern American domain. This will prevent the model from successfully generalizing to any other spatial domain, and the manuscript should either display proof that the model is indeed able to generalize spatially, or clearly state that EnsAI is not a general emulator but restrained to this domain."

I have added this paragraph to the conclusions:

"EnsAI was trained on a year's worth of data from the GEM-MACH regional air quality forecasting model, which contained data only from North America. Although this exposed EnsAI to regions with a variety of different land-use categories, agricultural types, and meteorological conditions during training, the training data may have not included drastically different meteorological conditions or chemical regimes than those seen in North America during the year used for training. As such, until further tests can be conducted demonstrating its performance in different regions, EnsAI in its present form should be considered to be restricted to its current spatial domain."

"Furthermore, while training on a year long of data might capture seasonal variability, it will be insufficient to capture inter-annual variability of meteorological conditions that the model could encounter in its future uses."

See the response below regarding chemical precursors.

"The author should specify explicitly which of the network's hyperparameters has been tuned. Specifically, the first layer of the network takes a hard jump from 4 to 128 channels that appear somehow suspicious. The loss curves of figure S1 clearly displays strong signs of overfitting, the

validation loss stagnate while the training loss keeps decreasing, it seems the training of the model could have stopped much earlier than 50 epochs. The author should state if any early stopping was used, and if not, precise why.”

The weights with the minimum validation loss were used, which occurred at the 9th epoch. An explicit early stopping was not used. As the reviewer suggested, the training could have stopped at an earlier epoch, so not using an explicit early stopping was not an efficient use of computing resources, but this had no effect on the actual results.

In the previous version of the manuscript, details on the training were scattered throughout different places. I’ve reorganized the manuscript so that most of the details of the training are in a single subsection of the Appendix. This new subsection reads as:

“Training was done on a single NVIDIA A100 GPU. The network was trained using the Adam optimizer with a batch size of 32 and learning rate of 10^{-4} . The learning curve for the training is plotted in Figure S1 of the Supplement. The weights that produced the minimum validation loss were selected for the evaluation, which occurred at the 9th epoch. Selecting the weights that yielded the minimum validation loss reduces the risk of overfitting that might have occurred in the later epochs of the training.”

Also, in the learning curve plot, I’ve cut off the epochs past epoch 20 since it was not necessary to run the training for so many epochs.

In regards to the jump from 4 to 128 channels, the choice of expansion from 4 to 128 channels was guided by common U-net architectures used in image generation applications. For example, the DDPM model of Ho et al. (2020) and the NCSN network of Song and Ermon (2019) that start with an expansion of 3 to 128 channels. Some initial tests were performed varying the number of channels, but not a full hyperparameter tuning of this number.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Song, Yang, and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution." *Advances in neural information processing systems* 32 (2019).

“Since the model is trained with MSE error without physical constraint, the author should discuss how prone is the model to respect mass preservation or to predict physically impossible negative concentration.”

Neither the processes of creating the ensemble of concentration perturbations nor the inversion process conserve mass. One can instead compare the total mass of an ensemble member from the GEM-MACH ensemble to the total mass of an ensemble member from the EnsAI ensemble, but this comparison does not necessarily have anything to do with the MSE loss function. For example, there could be a mismatch in the global scaling of the concentrations (i.e. Eqs. (4) and (5) could be scaled by different amounts), which would cause a discrepancy in the total mass, but I don’t think that would have anything to do with the MSE loss function in particular.

With regard to predicting negative concentrations, the standard procedure used at ECCC for chemical data assimilation is to only enforce non-negative concentrations when the increment is

added to the background to form the analysis. When the analysis is computed, any negative values are clipped. Since this clipping step follows the inversion step, the physical constraint is applied post inversions.

I'm not sure what other physical constraint would be relevant to model error covariance that could be violated by using the MSE. I think the possible effect of using the MSE as the loss function that is relevant for this particular application is unwanted smoothing at smaller scales, which was already mentioned in the manuscript.

“Since the model is trained on the sole concentrations of ammonia without any information on its chemical precursors on a year of data, the emulator will implicitly assume a constant chemical regime based on that year, and fail to predict accurately ammonia in case of a chemical regime shift (ex: regulations on emissions or wildfire). This might break the physical consistency entirely. The author should discuss these limitations in the manuscript.”

I think this comment and the comment above about inter-annual variability pose a challenge regardless of how many years of data the model was (pre-)trained on. If in future years the chemical regime or meteorological conditions are drastically different, the accuracy of the model might decrease.

If the model had input fields for the chemical precursors of ammonia, the model might be able to handle a future year with very different conditions better than a model without these input fields if it can learn enough about the future chemical regime within the training data. However, if the future time period is very different, then the model might not be able to learn enough about the future conditions from the training period. Having the precursor fields as input to the model may decrease the degree of any degradation that might occur in future years with different conditions, but fundamentally if future conditions (either the chemistry or the meteorology) are drastically different than those in the training data, some form of continued training or fine tuning would be required to learn these new conditions.

To handle significant changes in either meteorological or chemical conditions, I think the most straight-forward solution is continued training on data from future years as they become available.

In addition to the new paragraph in the conclusions included above that dealt with the limitations of the model, this sentence has been added to the conclusions:

“Additionally, if future meteorological or chemical conditions change significantly from those in the training period, continued training (or fine-tuning) on data from future years might be necessary as this data becomes available to maintain EnsAI's accuracy.”

“The author present in figure 16 inversion comparison plots $|\Delta e^{\text{GM}} - \Delta e^{\text{static}}| - |\Delta e^{\text{EnsAI}} - \Delta e^{\text{static}}|$ which is a excellent choice since most of the increment are dominated by the static baseline, thus these plots show that the model is also sensitive to the dynamics meteorological correlations. The figure in S11 and S12 in the supplements showing the difference between the physical and the AI model bring the rest of the necessary information. While these plots are necessary, they can be hard to read and more representation of these maps showing different metric as normalized mean bias, correlations”

A new figure (what is now Fig. 17 in the updated manuscript, shown below) has been added to the manuscript. This figure contains six panels that show the mean and standard deviation of the emissions from the GEM-MACH inversions, and the bias and standard deviation of errors (STDE) of the emissions increments from the EnsAI and static covariance inversions. I tried plotting the correlations, but the plots were not very informative.

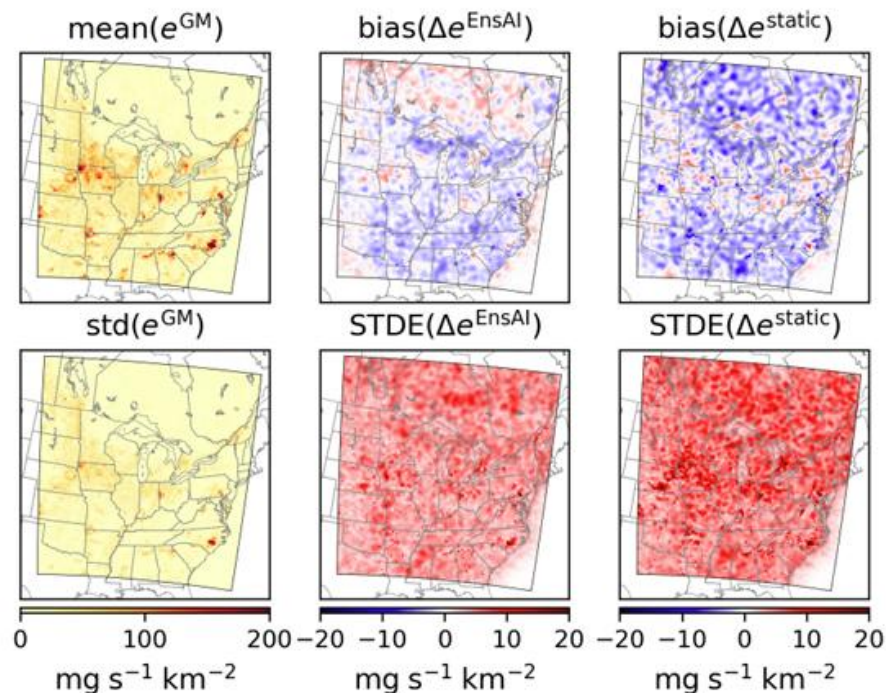


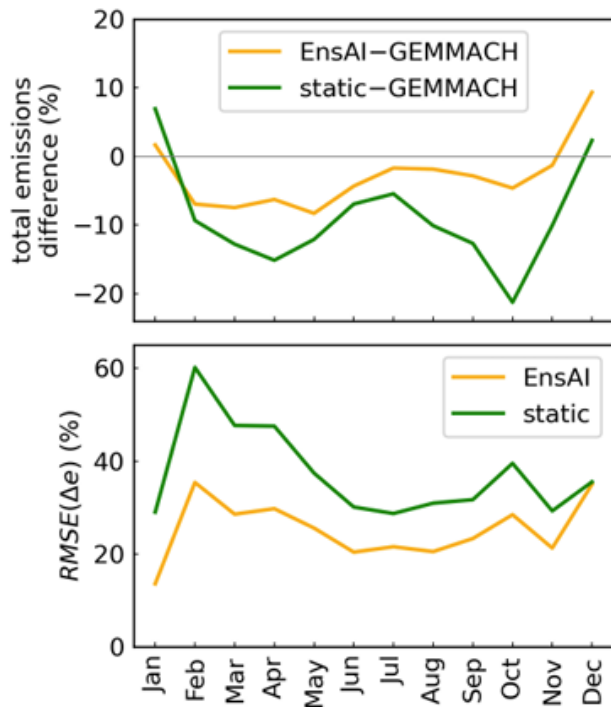
Figure 17. Statistics for the ammonia emissions inversions taken over 2015. The left column shows the mean (top panel) and standard deviation (bottom panel) of the emissions analyses from the inversions that used the GEM-MACH ensemble. The middle column shows the bias (top panel) and *STDE* (bottom panel) for the emissions increments from the EnsAI inversions. The right column displays the bias and *STDE* for the increments from the static covariances inversions.

Plots of $\Delta e^{\text{EnsAI}} - \Delta e^{\text{GM}}$ and $\Delta e^{\text{static}} - \Delta e^{\text{GM}}$ were also added to the Supplement to make it easier to see the difference between the various inversions.

Somewhat related to this comment, a new section was added to the manuscript examining the posterior surface concentrations. This new section contains a new figure showing the difference in biases in the posterior surface concentrations. Plots of the bias and STDE of the posterior surface concentrations were also added to the Supplement.

“... and some time dependent display would make the comparison easier to appreciate.”

A new panel was added to the figure that originally only showed the RMSE of the emissions increments for each month. The new panel in this figure (shown below) show the differences in total emissions within the model domain for each month.



“It would be particularly suited for this work to perform ablation study of the meteorological parameters. It would show in particular the ability of the model to predict accurately concentrations from meteorological context instead of from spatial statistical interpolation.”

From the comment above, I think the reviewer is suggesting I do an ablation study where EnsAI in its current form is compared to the network retrained without meteorological inputs (i.e. only δe as the input). The ‘static’ covariances that were used as a point of comparison in the manuscript is flow-independent, or in other words, it does not use the meteorological fields as input. The static covariances are two-month mean covariances (computed separately for each hour), so they reflect average meteorological conditions, but not the actual meteorology at a particular time.

The comparison to the static covariance is not technically an ablation study, since it does not use the neural network at all, but it answers the same question, namely, ‘what is the effect of not including meteorology / flow dependence’. So if I did interpret the comment correctly, I think that this particular ablation study might be redundant.

However, I’m not sure I’m interpreting the comment correctly. If I have misinterpreted the comment, I was hoping the reviewer could provide more details on the ablation study envisioned.

“Section 8.1, Paragraph 1: ‘...while the correlations are comprised by the univariate...’: are composed of

Corrected.

‘...the univariate emissions correlations ρ_{ee} , the univariate surface concentration ρ_{cc} , and...’: surface concentration correlations”

Corrected.