

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

1) "The justification for restricting the input variables to only emission perturbations, horizontal winds, and surface temperature is insufficient. Atmospheric ammonia concentrations are highly sensitive to boundary layer dynamics (e.g., PBLH), humidity, precipitation (which drives wet deposition), surface radiation, land-use categories (which govern dry deposition), and the background chemical state. By omitting crucial geographical and meteorological predictors, the neural network is forced to implicitly memorize location-specific mappings from the training data rather than learning generalizable physical relationships."

As the reviewer suggests, the effects of the variables not explicitly provided to the network may be modelled by the network indirectly through correlations with the supplied input fields. I interpret the assertion that the network is 'memorizing location-specific mappings' as meaning that the network is using positions within its grid to derive a large portion of its behaviour.

The longitude, latitude, and explicit grid position are not used at any point in network, so if location dependent behaviour is learned, the network would have had to indirectly learn positional information to bypass learning generalizable physical relationships.

Convolutional neural networks are in principle translationally invariant, but I am aware of at least one way in which specific architectural choices used in certain situations can break the translational invariance and positions can be deduced by the network. Kayhan et al. (2020) demonstrated that CNNs (at least for some problems/networks) can deduce locations within the grid by exploiting boundary effects. Kayhan et al. (2020) (reference below) mention that this is possible in CNNs with large receptive fields. However, as discussed in more detail below, EnsAI has a smaller receptive field than many U-Nets, which would presumably limit this effect (if it is a concern at all since this effect is application dependent).

To assess whether positional information is being used indirectly by the network to derive some significant part of its behaviour, I performed a test of EnsAI somewhat in the same spirit as the test illustrated in Fig. 1 of Kayhan et al. (2020). In this test, the field values (for all input fields) within the leftmost 90 grid points were fed into the rightmost part of the grid (akin to the image of the rook in Fig. 1 of Kayhan et al. (2020) moving from the top-left corner to the bottom right corner). In this test, the grid point at (97.62° W, 41.43° N) which is in Nebraska and near the midpoint of this translated patch is translated to the location in Pennsylvania that was examined in the manuscript and labelled as '(d)' in Fig. 1 of the manuscript. In other words, in this test, point '(d)' in Fig. 1 corresponds to a grid position of a point in Pennsylvania but the physical fields corresponds to conditions at a location in Nebraska.

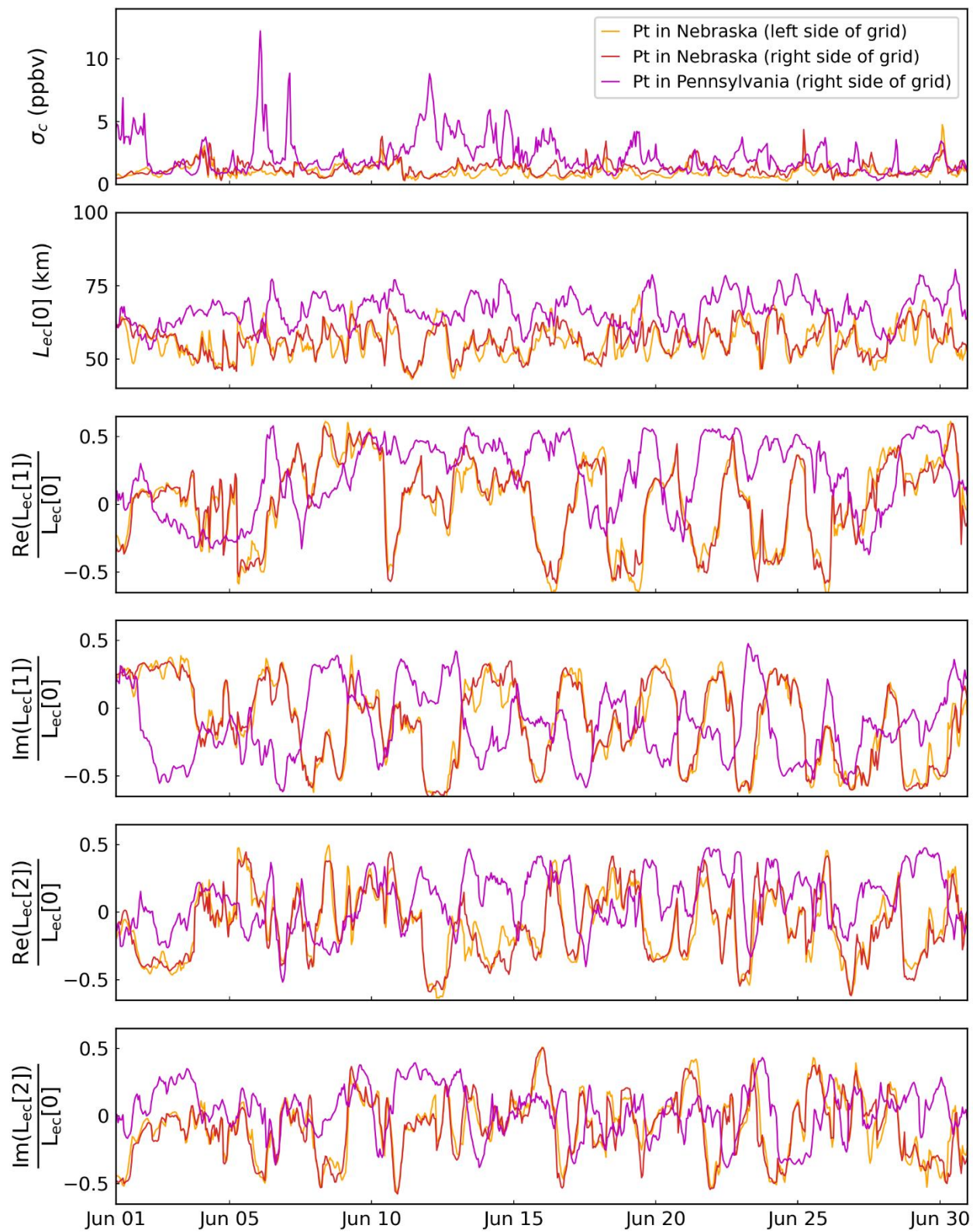
The figure below shows time series for the EnsAI results, similar to the time series in Fig. 5 to 9 of the submitted manuscript (the GEM-MACH and static covariance time series are omitted for clarity). The curve in orange labeled 'Pt in Nebraska (left side of grid)' corresponds to values in the original EnsAI run at the location in Nebraska (so this would be in the left side of EnsAI's grid). The curves in red labeled 'Pt in Nebraska (right side of grid)' are the results of this new EnsAI run at the translated position for Nebraska (i.e. at point (d) in Fig. 1 on the right side of EnsAI's grid). So the orange and red curves share the same physical conditions, but correspond to different grid

positions. The curve 'Pt in Pennsylvania (right side of grid)' in magenta shows the original EnsAI run at the locations in Pennsylvania (these curves are the same as shown in Fig. 8 of the submitted manuscript). So the orange and magenta curves share the same grid position, but correspond to different physical conditions.

While the orange and red curves are not identical, they are for the most part quite close to one another and have very similar behaviour. In contrast, the red and magenta curves are extremely dissimilar. This test provides evidence that EnsAI behaviour is primarily derived from the input field values fields rather than by absolute grid position, as this test shows that the same physical conditions, rather than the same grid location, produces very similar results.

Although the orange and red curves are very close to one another, they are not exactly the same, so EnsAI may not be completely translationally invariant. So it is possible that implicit positional information might explain the small differences between the orange and red curves. But these differences are small, so any possible positionally dependent behaviour would be small as well.

The network does need to implicitly learn about the factors mentioned above (PBLH, deposition, etc...), but the translational invariance test shows this is (at least primarily) not done using positional information.



Kayhan, Osman Semih, and Jan C. Van Gemert. "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location." *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

"The authors must conduct and present sensitivity experiments (or ablation studies) justifying that this minimal set of four predictors is fully adequate to represent the complex atmospheric chemistry involved, or explicitly discuss the limitations this imposes on the emulator's generalizability to different domains or vastly different meteorological regimes."

I've added a new section 'Translational Invariance of the Network' to the Appendix of the manuscript discussing the test done above (and is referenced in the main part of the manuscript).

While I think this test demonstrates that the network is learning location independent physical relationships, if a different domain has 'vastly different meteorological regimes', then there is no guarantee that EnsAI would have a similar performance in this domain. Accordingly, I have added these sentences to the conclusions

"EnsAI was trained on a year's worth of data from the GEM-MACH regional 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."

2) "Expanding directly from 4 input channels to 128 channels in the first convolutional layer leads to a massive surge in network parameters."

The expansion from 4 to 128 channels was selected because similar initial channel widths are used in established U-net architectures. 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.

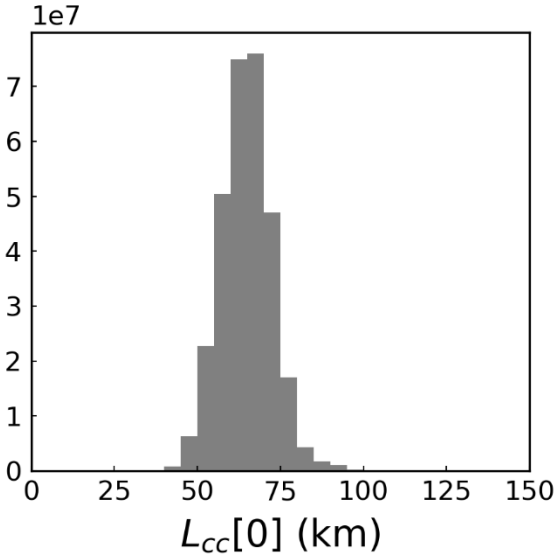
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

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

"Furthermore, utilizing only two downsampling layers (bottlenecking at a  $64 \times 64$  grid from a  $256 \times 256$  grid) restricts the model's receptive field, which might limit its ability to capture larger scale regional transport phenomena given the  $\sim 10$  km resolution... A study justifying the chosen channel depth and downsampling levels would also significantly strengthen this section."

The justification for only using two downsampling layers was not adequately explained in the previous version of the manuscript. Ammonia has a relatively short atmospheric lifetime ( $\sim$ hours to a day or two), which limits the scale of its advection. Consequently, its background error covariance matrix is sparse and concentrated within  $\sim 100$  km of each location.

Below is a histogram of the horizontal (isotropic) correlation length scale of surface ammonia ( $L_{cc}[0]$ ) for the GEM-MACH ensemble for all grid points within the domain for 2015. From this histogram, we can see that there are almost no horizontal correlation lengths larger than 100 km. The network used in EnsAI has a receptive field of ~320 km (so 160 km in either direction), so its network should have an sufficiently large receptive field to properly capture the spatial correlations of the error covariance matrix.



Adding more downsampling layers would not only increase the number of tunable model parameters beyond what is required for this particular application, any larger-scale features that would be created by adding further downsampling layers would just be removed by localization (see Eq. (3) in the manuscript), so would not be present in the error covariance. As stated in Appendix A, the localization used had a half-width at half-maximum of 120 km. Adding another identical downsampling layer into the network would expand the receptive field to ~680 km (340 km in either direction), which is far beyond the correlation length scales in this application.

The following paragraph has been added to the end of Section 4:

“The U-Net used for EnsAI has two downsampling layers, in contrast to many other U-Net designs that have three. Ammonia has a relatively short atmospheric lifetime which limits the scale of its advection. Consequently, its background error covariance matrix is sparse and concentrated within ~100 km of each location. After EnsAI’s two downsampling layers, at the bottleneck of the network, the network’s receptive field is about 320 km (or 160 km in either direction), which should be adequately large to properly capture the spatial correlations of the error covariance matrix. Adding additional downsampling layers would increase the number of tunable model parameters beyond what is required for this particular application, although we note that longer lived species would likely require additional downsampling layers. As mentioned in Section 2.3, localization is applied to the ensemble covariance to remove spurious long-distance correlations due to the limited ensemble size. As the horizontal localization length scale for the ammonia emission inversions had half-width at half-maximum of 120 km (see Section A of the Appendix), the

additional longer-scale features created by adding more downsampling layers would be removed by the localization in any case.”

“Crucially, Figure S1 reveals a classic signature of overfitting: the validation loss plateaus after approximately 3-5 epochs, while the training loss continues to decrease steadily. The authors must clarify if an early stopping strategy was employed to save the model weights at the optimal validation loss”

Weights were saved at the epoch with the minimum validation loss (the 9<sup>th</sup> epoch) and used for evaluation of the test data set. The training was just run through 50 epochs instead of having an explicit early stopping. An explicit early stopping strategy would have been better since more epochs were performed than was necessary. So not having an explicit early stopping was not an efficient use of computing resources, but did not impact the evolution of the model.

“specify the optimizer used”

The ‘Adam’ optimizer was used.

“and any regularization techniques applied (e.g., weight decay, dropout)”

No explicit regularization methods were used.

”and explicitly state the activation function utilized at the final output layer.”

No activation function is applied after the 2D convolution in the final layer. A final linear output layer is commonly used when the final output field should be unbounded, since the activation function often bounds the output. Although the full (i.e. background + perturbation) emissions and concentration fields are non-negative, which would give a lower bound to the perturbation fields, in practice it is much easier to impose this bound after the increment is produced when the increment is added to the background field to produce the posterior field.

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.”

Figure S1 has been limited to the first 20 epochs since it was not necessary to run the training for so many epochs.

3) “The comparison in Section 9 demonstrates that the EnsAI-based inversion closely matches the GEM-MACH-based inversion increments. However, reproducing the physical model's output validates the emulation quality, but does not inherently guarantee a successful inversion in terms of real-world physics. While the manuscript heavily evaluates the emissions increments, the state vector naturally involves concentrations (Eq. 1), and the manuscript currently lacks a discussion on the concentration optimization effect.”

In the previous version of the manuscript, I chose not to include evaluations of the posterior concentrations and comparisons to observations for brevity, but I agree that these should have been included.

A new subsection named 'Posterior Surface Concentrations' has been added to the manuscript that discusses optimized surface concentrations. This is accompanied by a new figure in the main body of the manuscript that shows a comparison of the surface concentration biases between the different model error covariance methods. Additionally, six new figures have been added to the Supplement that show biases and standard deviations of differences of the various surface concentration fields.

[“The authors should discuss how well the posterior concentrations match the CrIS satellite retrievals post-inversion.”](#)

As the CrIS retrievals were used in the inversions, comparison to those retrievals does not provide an independent evaluation of the inversions. As the reviewer suggests below, an independent evaluation of the posterior surface concentrations using observations from the AMoN and NAPS networks were added instead (see below).

[“Additionally, to prove true inversion utility, the authors are strongly encouraged to validate the posterior results against independent observational data \(e.g., ground-based NH<sub>3</sub> monitoring networks\) to confirm that the EnsAI emulator maintains or improves actual forecast skill compared to the baseline.”](#)

Evaluations of the posterior concentrations to surface observations from the AMoN and NAPS networks has been added to the manuscript. The AMoN and NAPS surface observations are introduced in a new section named 'Surface Observations' that is right before the results sections.

Another new subsection named 'Comparison to Surface Observations' has been added to the manuscript directly after the new 'Posterior Surface Concentrations' subsection. A new figure showing the RMSE values for these stations has been added to the main body of the manuscript. In the Supplement, three new figures have been added that show the biases of the posterior concentrations relative to these observations, as well as a new table that lists various comparison statistics for these surface observations.

Because these additions are lengthy, I will refer the reviewer to the updated version of the manuscript for these additions.

4) [“The manuscript notes that the ensemble of emissions is constructed from a Gaussian random field and that “EnsAI uses the same emissions ensemble as the GEM-MACH ensemble”. If the spatial fields are static across time and the exact same 60 spatial perturbation modes were used for both training \(2016\) and testing \(2015\), the neural network may have simply memorized the spatial response to those specific 60 random fields. The authors must clarify if the Gaussian random fields are dynamic \(redrawn temporally\) or static.”](#)

The inversion system that EnsAI was designed for performs inversions with a one-month time window to estimate emissions parameters within this time window. This set of emissions model parameters are then used in the model within this month. In this sense, these model parameters

are 'static' in that the same emissions parameters are used at each time within this month. While these emissions parameters are 'static', the actual emissions at each hour changes, as the actual emissions are meteorologically dependent. So in this sense the hourly emissions are not static. But I think this comment is about whether or not the Gaussian random fields are redrawn. To avoid ambiguity, instead of discussing if a field is static or not, I'll instead discuss whether perturbations are redrawn or not.

In this inversion system (regardless of whether EnsAI is used or not), the Gaussian random fields do not get redrawn. These Gaussian random fields are used to create the emissions perturbations, and these emissions perturbations change in time, but the underlying Gaussian random field for each member does not change in time. The perturbations are emissions model parameters that are used within the month-long inversion window, so they cannot be redrawn within the inversion window.

[“If they are static, the model must be tested on a completely newly generated set of 60 random fields to prove that EnsAI can generalize to unseen perturbation patterns, rather than just unseen meteorological conditions”](#)

New 'unseen' perturbations can be constructed by simply taking linear combinations of these ensemble members (this is exact for a linear model and approximate to first order for a non-linear model). This property is fundamental to inversion/data assimilation algorithms, and the analysis increment itself is such a linear combination.

The 60 ensemble perturbations form a basis that span a subspace of the covariance matrix that it is attempting to approximate. In this sense, the product of the EnsAI is this subspace, and its basis can be used to construct any number of new perturbations within the subspace. So the issue of generalization in this context is quite different than in other applications. The relevant question is are 'unseen' perturbations in this subspace.

From previous work with this inversion system, a 60 member ensemble provides error covariances of sufficient rank and the subspace that it spans is adequately large. If it were the case that our ensemble-based error covariances were significantly rank deficient, then it would be interesting to see if EnsAI was able to draw perturbations that were outside of the subspace spanned by the ensemble it was trained on, so that the rank of the error covariance could be increased. But that would require drawing more than 60 members (not 60 as suggested in the comment). A short note about this was made at the beginning of the 'Direct Comparisons of the Covariances' section of the manuscript.

[5\) “The authors are strongly encouraged to present spatiotemporal distribution maps to demonstrate whether the concentration simulation by the EnsAI can fully reproduce the concentration fields generated by the physical model.”](#)

As mentioned in the response to comment (3), one new figure has been added to the main body of the manuscript and six new figures have been added to the Supplement to show this.

[6\) “The Title and Introduction heavily frame this work as an emulator for generic "atmospheric chemical ensembles" and "tropospheric chemical constituents". However, the study exclusively trains, tests, and evaluates the model on surface ammonia \(NH<sub>3</sub>\). Ammonia is heavily surface-bound and short-lived, making it uniquely suited to a surface-only emulator, whereas other](#)

tropospheric chemicals (e.g., ozone, carbon monoxide) have longer lifetimes, complex multi-species chemical pathways, and require 3D vertical inputs. The language throughout the manuscript should be tempered to reflect this specific application, and the title and abstract should explicitly mention that the current framework is demonstrated specifically for surface ammonia emissions, avoiding aggressive generalizations to all chemical constituents.”

Looking back at the manuscript, I agree that some of the language in the title, abstract, and introduction was overly ambitious. The title of the paper has been changed from “EnsAI: An Emulator for Atmospheric Chemical Ensembles” 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”.

“Section 9, Page 21: The concluding sentence of Section 9 states: “Therefore, overall, the increments produced using the EnsAI ensemble are much closer to the increments produced by the static covariances...”. This is a direct contradiction of the manuscript's findings. It must be corrected to state that the EnsAI increments are much closer to the GEM-MACH increments.”

Corrected. Thank you for catching this mistake.

“The model is trained on 2016 data, tested on 2015 data, and validated on July 2014 data. Using future data to predict past data can introduce hidden biases, especially if there are long-term trends in background emissions. A brief justification for this specific temporal split should be added to Section 5.”

I’ve added this sentence to Section 5:

“As the change in total atmospheric ammonia in North America is small between these years (Van Damme et al., 2021), evaluating with the year prior to the year used for training is unlikely to be an issue.”