Review #1

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.

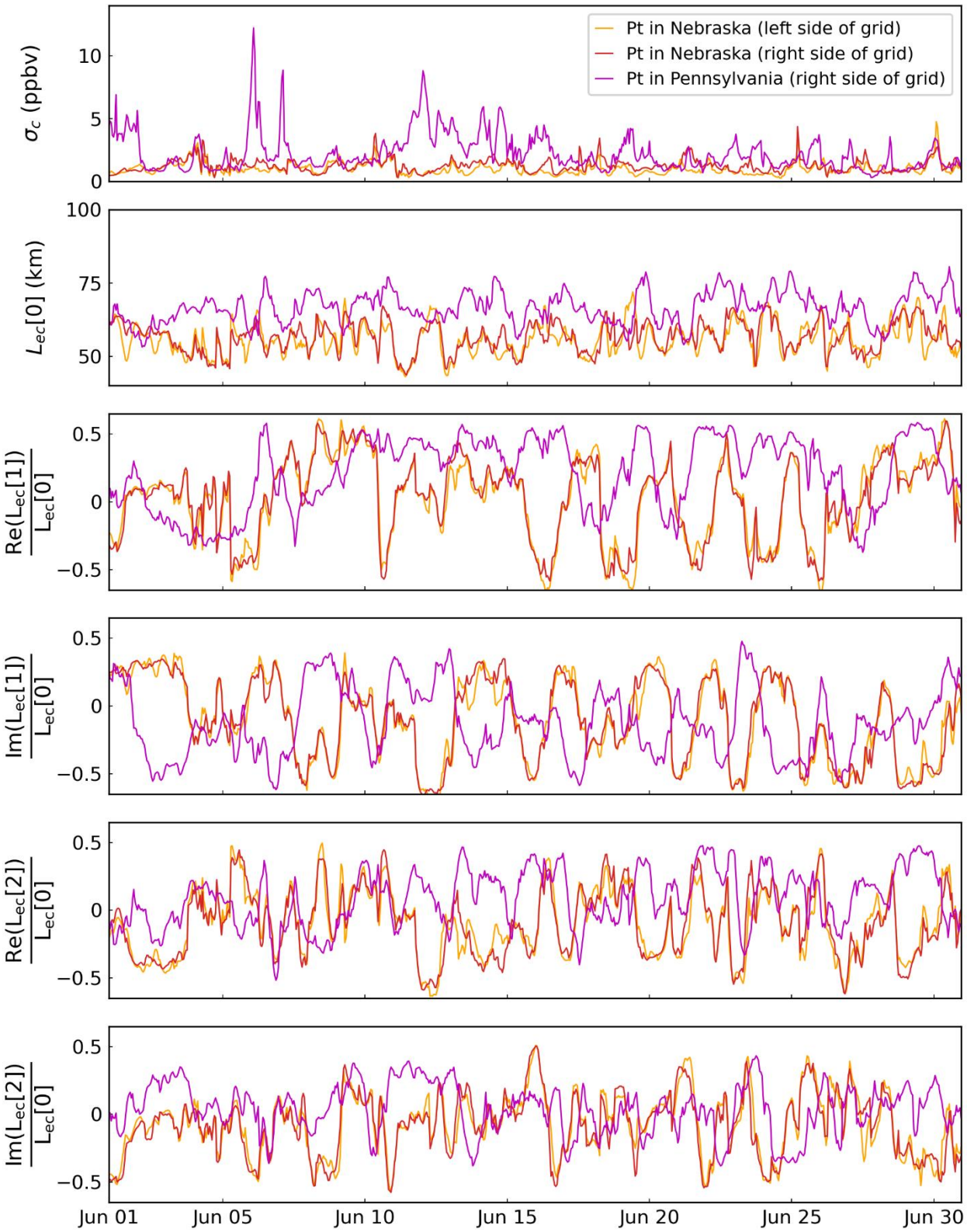
The figure below shows time series for the EnsAI results, similar to the time series in Fig. 5 to 9 of the submitted manuscript (the GEM-MACH and static covariance time series are omitted for clarity). The curve in orange labeled 'Pt in Nebraska (left side of grid)' corresponds to values in the original 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.


Pt in Nebraska (left side of grid)
Pt in Nebraska (right side of grid)
Pt in Pennsylvania (right side of grid)
σ_c (ppbv)
$L_{ec}[0]$ (km)
$\frac{\mathrm{Re}(L_{ec}[1])}{L_{ec}[0]}$
$\frac{\mathrm{Im}(L_{ec}[1])}{L_{ec}[0]}$
$\frac{\mathrm{Re}(L_{ec}[2])}{L_{ec}[0]}$
$\frac{\mathrm{Im}(L_{ec}[2])}{L_{ec}[0]}$
Jun 01
Jun 05
Jun 10
Jun 15
Jun 20
Jun 25
Jun 30

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.

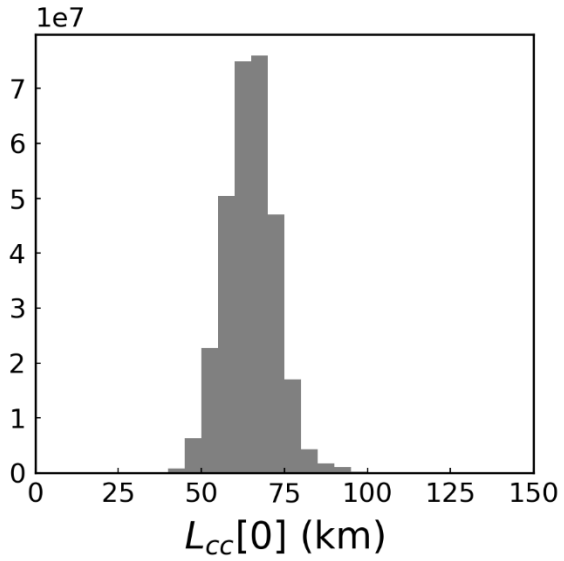
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

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 × 64 grid from a 256 × 256 grid) restricts the model's receptive field, which might limit its ability to capture larger scale regional transport phenomena given the ~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 (~hours to a day or two), which limits the scale of its advection. Consequently, its background error covariance matrix is sparse and concentrated within ~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 9th 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 occured 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 NH3 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). 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 EnsAI is this subspace, and its basis can be used to construct any number of new perturbations within the subspace. In other words, EnsAI is not restricted to perturbation patterns that it was exposed to from this set of 60 perturbations, but is restricted to the subspace that is spanned be these 60 perturbations. So the issue of generalization in this context is quite different than in other applications. The relevant question is are ‘unseen’ perturbation patterns 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 (NH3). 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."

Review #2

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.

"This work presents a valuable advance in AI-based ensemble generation for atmospheric chemistry, but its scope is currently limited to surface ammonia. The title and abstract should be revised to accurately reflect this limitation, or the manuscript should provide evidence for generalization to other chemical species."

Looking back at the manuscript, I agree that the title and some of the discussion was overly ambitious. I've tried to temper the language used in the title, abstract, and introduction. The title of

the paper has been changed to 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,

> “The aim of this work is to develop an AI-based model for efficient ensemble generation of tropospheric chemical constituents”

has been changed to

> “The aim of this work is to develop an AI-based model for efficient ensemble generation of atmospheric ammonia surface concentrations”.

The previous draft of the introduction contained the sentence

> “While generating ensembles for ammonia emissions inversions will be used as a test application for the AI-generated ensembles, another goal of this work is to develop an AI system flexible enough to be used in the future for generating ensembles for other chemical constituents relevant for air quality.”

has been moved to the ‘Conclusions’ section (and reworded). Also in the introduction,

> “an ensemble of atmospheric concentration perturbations”

has been changed to

> “an ensemble of ammonia concentration perturbations”

In the ‘Emissions Inversions’ section, the sentence

> “This study uses the application to ammonia emissions inversions as a testbed for developing the foundations of an AI-based ensemble generation system.”

has been changed to

> “This study develops an AI-based ensemble generation system to be used in an ammonia emissions inversion system”.

“The discussion of limitations and future directions should be strengthened, particularly regarding the model's applicability beyond ammonia and surface concentrations.”

The following paragraphs have been added to the conclusions section:

> “ 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.

While EnsAI was successfully implemented in an emissions inversion system for ammonia, a species with a short atmospheric lifetime, expanding EnsAI to work with longer lived species will require expanding the input capacity to the network. Most significantly, longer lived species will require EnsAI to model a prognostic relationship, which would add fields from previous time steps to the input. Input and output fields for the vertical level above the surface would also be required for long lived species."

The part about the prognostic relationship mentioned in the second paragraph will be clarified below.

1) "The manuscript should more clearly distinguish among the training period, the temporal resolution of the training samples, and the inversion window … In addition, what is the exact inversion window used for the results presented in this study: one week, several weeks, or one calendar month? The manuscript states that ammonia inversions generally use several weeks of CrIS retrievals and that each inversion estimates a time-mean emissions field over the inversion window. However, Figures S11–S13 present one emissions increment for each month, which appears to imply that monthly inversion windows were used. Please state the start and end dates, or the duration, of each inversion window explicitly."

The inversion time window is one calendar month. An unperturbed GEM-MACH model run is run for this month, which produces hourly atmospheric concentrations of ammonia. An ensemble of perturbed GEM-MACH runs were run over this month, producing hourly perturbed atmospheric ammonia fields. These perturbed concentrations fields were used as the training data for EnsAI (which is designed to emulate these perturbed atmospheric ammonia fields), so the temporal resolution of the training data is one hour. 12 months of hourly data was used for training.

The lines in the original manuscript that

"emissions inversions are generally performed using weeks of retrievals"

were just meant to convey that inversion windows need to be at least a few weeks long, but in principle could be any time period longer than that. But generally these inversions are over one calendar month, so I think this sentence was needlessly confusing and didn't serve a clear purpose. I've removed this sentence in the new draft of the manuscript.

I've changed the manuscript text to contain these lines:

"The hourly fields from an unperturbed GEM-MACH run are used as the a priori (background) model state. The ensemble used in the inversions is comprised of a set of hourly fields produced by perturbed GEM-MACH runs. The emissions inversions use a calendar's month worth of retrievals with the hourly background and ensemble model fields to produce updated emissions model parameters for a particular month. While the product of each inversion is a single field of the emissions parameters for a particular month, the

posterior ammonia emissions vary within the inversion time window as the ammonia emissions are dependent on meteorological conditions, most notably the surface temperature."

Additionally, I've added this new figure to the paper that shows a diagrammatic representation of the inversion process to further clarify the frequency of the data involved:

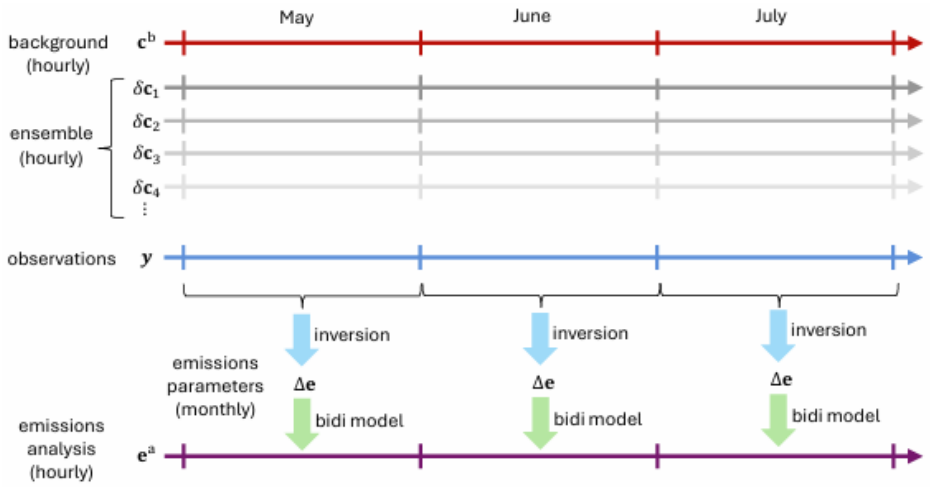


Figure 2. Diagram of the ammonia emissions inversion process. The hourly background concentrations ($\mathbf{c}^b$) and ensemble concentration perturbations ($\delta\mathbf{c}$) are used with observations ($\mathbf{y}$) within a month to produce monthly increments to the emissions parameters $\Delta\mathbf{e}$. GEM-MACH's bidirectional flux model (represented by the green arrow in the diagram) combines the monthly emissions parameters with meteorological fields to yield hourly updated emissions ($\mathbf{e}^a$).

"EnsAI is trained using one year of hourly GEM-MACH ensemble outputs. However, Equation (5) appears to represent an instantaneous mapping from the emissions perturbation and corresponding meteorological fields to the surface ammonia concentration perturbation, rather than a one-hour-ahead forecast. Please explicitly clarify the temporal relationship between the inputs and output."

The previous version of the manuscript lacked some context which is key to how EnsAI functions. In principle, Eq. (5) should be a prognostic equation, where the atmospheric concentration δc at some particular time is dependent on the concentrations at previous times. However, atmospheric ammonia has a short atmospheric lifetime, so perturbations to the concentrations are short lived, and are primarily driven by emissions. As such, an approximation was made to drop the dependency of δc on previous time steps. Note that the unperturbed ammonia concentrations are still solved via prognostic equations (as the unperturbed run comes from GEM-MACH) and is it only the concentration perturbations where this approximation is used.

The approximation of not including a dependence of δc on previous time steps was made due to the memory restrictions on the GPU, which greatly limited the number of input fields to the network. As only a limited number of input fields were feasible, only the most important fields for determining the perturbation distribution were used. The results of the manuscript show that EnsAI

was able to generate very similar covariances to the original GEM-MACH ensemble despite the fact that δc from previous time steps were not input into the network.

The results on test set show that this is a reasonable approximation for ammonia. However, neglecting δc from previous time steps would likely not work for longer lived species.

I don't think the previous version of the manuscript emphasized this point enough. I added this sentence before Eq. (5)

> "Ammonia has a relatively short atmospheric lifetime and individual perturbations to the ammonia concentrations are short-lived. We exploit these properties by neglecting concentration or meteorological fields from previous time steps to further reduce the size of the input. Hence, EnsAI models a diagnostic, rather than a prognostic, relationship."

and these sentences after Eq. (5)

> "As EnsAI describes a diagnostic relationship between variables, δu, δv, δT, and δc in the equation above are all valid at the same time (while δe is a monthly mean parameter). We emphasize that the simplification to a diagnostic relationship will only be a reasonable approximation for short-lived species."

In the conclusions, the paragraph already mentioned above was added as well:

> "While EnsAI was successfully implemented in an emissions inversion system for ammonia, a species with a short atmospheric lifetime, expanding EnsAI to work with longer lived species will require expanding the input capacity to the network. Most significantly, longer lived species will require EnsAI to model a prognostic relationship, which would add fields from previous time steps to the input."

"For example, generating seven days of hourly outputs for 60 ensemble members corresponds to 60×7×24= 10,080 member-hour inferences. These calculations may of course be performed in batches rather than as 10,080 separate program executions. Please explain whether EnsAI generates the 168 hourly outputs for each member through independent hourly evaluations, a batched multi-time evaluation, or a single trajectory-generating pass. The computational configuration and batching strategy used to obtain the reported seven-second runtime should also be provided."

As Eq (5) is a diagnostic relationship, each concentration perturbation can be generated independently of one another. In the new version of the manuscript, I've reorganized the paper so that most of the details of architecture and network training are in a Section in the appendix (so that these details are in a single place instead of being scattered around the main body of the paper, appendix, and supplement). There is a new subsection in this appendix named 'Inference' that has the lines:

> "As EnsAI describes a diagnostic relationship, inference can be done in parallel and any number of different time steps could in principle be processed simultaneously. For the 7 second run time to produce a week's worth of simulated data per ensemble member that was quoted in Section 6, inferences were performed in batches of 16."

2) “The current EnsAI formulation does not include the previous ammonia concentration, previous model state, or an explicit sequence of earlier meteorological conditions as input.”

The additional text added to the manuscript described above clarifies this issue.

“It is therefore unclear how the model represents temporal memory associated with transport, chemical transformation, deposition and the accumulated response to a persistent emissions perturbation.”

While EnsAI is not provided the previous concentration perturbation or meteorology, it can still use the input fields it does have (current meteorology and emissions perturbation) to take the processes mentioned above into account.

During training, the response of EnsAI is tuned to map the emissions perturbation to a surface concentration perturbation. Processes such as deposition and chemical transformation affect the atmospheric lifetime of ammonia, which in turn affects the spatial extent of the horizontal correlations. So these types of processing are modeled indirectly through comparing the emissions perturbation to the spatial extent of the surface concentration perturbation. Similarly, information about the transport of ammonia is inferred by examining the spatial pattern of the surface concentration perturbation.

Networks that produce prognostic fields can learn about these processes during training by seeing the temporal evolution of fields. But EnsAI learns these about these processes instead by examining the correlations of the perturbations. For example, there is an intuitive relationship between the spatial correlation length scale of the concentration perturbations and the atmospheric lifetime: In general, the longer the lifetime, the longer the correlation length scale. Not having previous time step does limit this, but this is not a dramatic limitation for ammonia (as shown by EnsAI’s ability to reconstruct the error covariance of the ammonia surface concentration). In other words, at least for ammonia, much can be inferred from the covariance structure of the ensemble at a single time step.

“I suggest further investigating this issue using a multi-hour or multi-day sequence model, or by adding lagged meteorological and concentration fields to the EnsAI inputs.”

As EnsAI models a diagnostic relationship that has shown to be successful in reproducing the flow-dependent behaviour of the surface ammonia error covariance, this is addressed in the previously mentioned additions to the manuscript regarding the diagnostic relationship. However, this only applies to short-lived species like ammonia. Extending EnsAI to longer-lived species will necessitate doing as the reviewer suggests here. The previously mentioned addition to the conclusions:

> “Most significantly, longer lived species will require EnsAI to model a prognostic relationship, which would add fields from previous time steps to the input.”

addresses this issue for future work.

“Controlled experiments could be performed using either an impulse-like emissions perturbation or a persistent emissions perturbation. The resulting EnsAI and GEM-MACH concentration responses could then be compared after 1, 6, 12, 24 and 72 hours and after one week. Relevant

diagnostics could include concentration bias and RMSE, spatial displacement of the plume, ensemble variance, emissions–concentration cross-covariance and correlation length. This analysis would determine whether the EnsAI errors remain bounded or accumulate with increasing integration time. I would refer to a persistent emissions perturbation over the inversion window, rather than an "emissions error at the initial time", because the ensemble perturbations in the present study represent time-mean emissions perturbations rather than a perturbation applied only at the initial time."

Unlike a cycling assimilation system where forecasts are launched from analyses, the emissions inversion system uses a free model run of GEM-MACH as the background and updates emissions parameters instead of initial conditions. Hopefully the addition of the figure showing the diagram of the inversion process will make this clearer.

The issue of the 'errors remain bounded' will be discussed further below in response to comment (3).

As an aside, the comment above did mention 'Relevant diagnostics could include concentration bias and RMSE'. In the revised version of the manuscript, two new sections have been added: One that examines the posterior surface concentration fields and another that compares the posterior surface concentrations to surface observations from the AMoN and NAPS networks. Although these sections do not deal with the bounding of errors, they do contain these types of diagnostics.

3) "The operational applicability of EnsAI should be evaluated using a continuous or cycling experiment covering more than one inversion window. At present, the analysis mainly demonstrates that EnsAI can reproduce GEM-MACH ensemble covariances and inversion increments for individual periods. This does not fully establish whether the system remains stable and accurate when used repeatedly. I suggest conducting parallel cycling experiments with EnsAI and GEM-MACH over several consecutive weekly or monthly inversion windows. Both systems should use the same emissions perturbations, meteorological forcing and observations. The posterior emissions obtained in one window could then be used as the prior emissions for the following window. Because EnsAI currently emulates only the perturbations of one surface chemical field, rather than the complete GEM-MACH atmospheric state, this section may be framed as an assessment of its suitability for operational ensemble generation and emissions inversion, rather than as a general operational air-quality forecasting system. The manuscript itself notes that GEM-MACH generates hundreds of chemical fields over 85 vertical levels, whereas EnsAI generates only one surface concentration perturbation field"

The earlier reviewer comments dealt with some of the language in the previous version of the manuscript being too broad. This may have additionally caused some confusion about setup of the ensemble, the inversion system, and the differences with a traditional cycling data assimilation system. I hope narrowing the language as mentioned above, in addition to the added description of inversion system, will alleviate this issue.

An operational context for ammonia emissions is very different than for NWP (or for longer lived chemical species). The inversion system always uses a free model run as the background state. The emissions inversion system uses a month-long inversion window to update monthly mean emissions parameters. Ammonia emissions vary significantly depending on the month, as they are

driven primarily by agricultural activity. So, if the emissions parameters from a previous inversion were going to be used as the background parameter values of a future inversion, it would be for the same month of a future year, not a subsequent month within the same year. In this case, a single cycle would be a year (or at minimum a single month in a particular year). So to examine if 'the system remains stable and accurate when used repeatedly' would require examining inversions of the same month over many different years. To see if the covariance diverges over time would require many years worth of inversions (likely at least a decade), which is well beyond the scope of this study.

If EnsAI is extended for use with other species in a chemical data assimilation system, in which there is a complete assimilation/forecasting cycle, these tests would be advised, but this would be for a future system, not the system being evaluated in this study.

4) "Figures S11 and S12 show the GEM-MACH- and EnsAI-based emissions increments separately, making the magnitude, sign and spatial distribution of their differences difficult to assess directly. Please add maps of the signed difference

$\Delta e^{\mathrm{EnsAI}} - \Delta e^{\mathrm{GM}}$

for each inversion period, using a consistent colour scale. Maps of the absolute or normalized difference could also be included. These plots would reveal whether EnsAI introduces systematic positive or negative biases and whether the discrepancies are concentrated in particular source regions, seasons or meteorological regimes.

Figure 16 already provides a useful comparison, but it shows the difference between two absolute deviations,

$|\Delta e^{\mathrm{GM}} - \Delta e^{\mathrm{static}}| - |\Delta e^{\mathrm{GM}} - \Delta e^{\mathrm{EnsAI}}|$,

rather than the direct signed difference between the EnsAI- and GEM-MACH-based increments. Figure 17 provides an aggregated RMSE, but it does not identify the locations or directions of the discrepancies. Figures S11 and S12 confirm that the two sets of monthly increments are currently displayed only as separate maps."

Plots of $\Delta e^{\mathrm{EnsAI}} - \Delta e^{\mathrm{GM}}$ and $\Delta e^{\mathrm{static}} - \Delta e^{\mathrm{GM}}$ have been added to the Supplement (Figs. S15 and S16). Additionally, maps of the bias and standard deviation of the errors for the emissions increments from inversions using EnsAI and the static covariances having been added to the 'Inversion Results' section of the manuscript (Fig. 17). This is accompanied by a new paragraph in this section, which reads

> "Statistics of these increments taken over the year are shown in Figure 17. The middle and right columns of the figure shows the bias and the standard deviation of errors (STDE) of the emissions increments from the EnsAI and static covariances inversions, respectively, which are defined in Section D of the Appendix. The left column shows the mean and standard deviation of the emissions analyses from the GEM-MACH ensemble inversions for comparison (for notational simplicity, we will not write the ^{a} superscript for the emission analysis is this section). From this figure, we can see that the inversions using both EnsAI and the static covariance tend to underestimate the increments in many regions, although

the extent and magnitude of this underestimation is larger for the static covariances (also see Figures S15 and S16 of the Supplement for plots of individual months). The STDE values for the emissions increment using the static covariances are also noticeably larger than that for inversions using EnsAI in many regions."

"The direct difference maps could be accompanied by monthly MAE, RMSE, spatial correlation and mean bias"

A new plot of the differences in total emissions within the model domain between the different inversions was added as well, which is now a new panel in the same figure that showed the monthly emissions RMSE (Fig. 18). The difference in total emissions is basically the bias in the emissions but just presented a bit differently. I tried plotting the correlations but the plots were not very informative. The text that accompanies this figure reads:

" The differences in the total ammonia emissions within the model domain for each month are shown in the top panel of Figure 18. These curves are expressed as percentages by dividing by the total monthly emissions (averaged over 2015) within the domain as determined by the GEM-MACH ensemble inversions. The monthly total emissions from the EnsAI inversions were within 10% of those from the GEM-MACH ensemble inversions for all months. The monthly total emissions differences for the static covariance inversions exceeded a magnitude of 10% for March to May and August to November and had larger differences than the EnsAI inversions for all months except December. The total monthly emissions from both the EnsAI and static covariance inversions were lower than those from the GEM-MACH ensemble inversions for February to November. The total annual emissions from the EnsAI inversions were 2.6% lower than from the GEM-MACH ensemble inversions, but were 6.9% lower for the static covariance inversions.

The root mean square error of the ammonia emissions increment RMSE(Δe) over the domain for each month are shown in the bottom panel of Fig. 18 (expressed as percentages relative to the mean annual emissions given by the GEM-MACH ensemble inversions). The RMSE(Δe) values for EnsAI range from 14% to 35%, and between 29% and 60% for the static covariances. The RMSE for the increments produced using EnsAI have lower values for all months as compared to the static covariances, although only slightly lower for December"

"together with regional statistics for major ammonia-emitting areas"

I've added a new section that compares the posterior surface concentrations with surface observations from the AMoN and NAPS networks. This new section includes a plot of changes in RMSE values at each station. Three new plots have been added to the Supplement that shows the biases of the posterior surface concentrations at each station, and are discussed in this new section. This is not exactly what was suggested by the comment above, but I think it close since with these new plots the reader can examine statistics of the inversion results at different stations (including those near major emissions sources as mentioned above). The new version of the manuscript contains significantly more figures than the previous version (although I've had to place a lot of them in the Supplement as to not overwhelm the main part of the manuscript), so I was hoping the new plots showing the surface observation comparison results could address this comment as well.

"January deserves particular attention because the manuscript reports poorer agreement during this period."

While looking into this comment, I found that some of the GEM-MACH ensemble files that were used for the test set contained erroneous values. These files were excluded from the test set in the new version of the manuscript (this problem was with the GEM-MACH files saved on our file system, so was not a problem with EnsAI itself). The poorer agreement was due to the files with erroneous values. The results for January now look in line with the other months of the year.

Additional minor correction

Near the end of Section 9, the manuscript states that the EnsAI-based increments are "much closer to the increments produced by the static covariances." Based on the preceding analysis and Figure 17, this presumably should read "much closer to the increments produced using the GEM-MACH ensemble."

Corrected. Thank you for catching this mistake.

Review #3

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 9^{th} 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 occured 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 $|\Delta e^{GM} - \Delta e^{static}| - |\Delta e^{EnsAI} - \Delta e^{static}|$ 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.

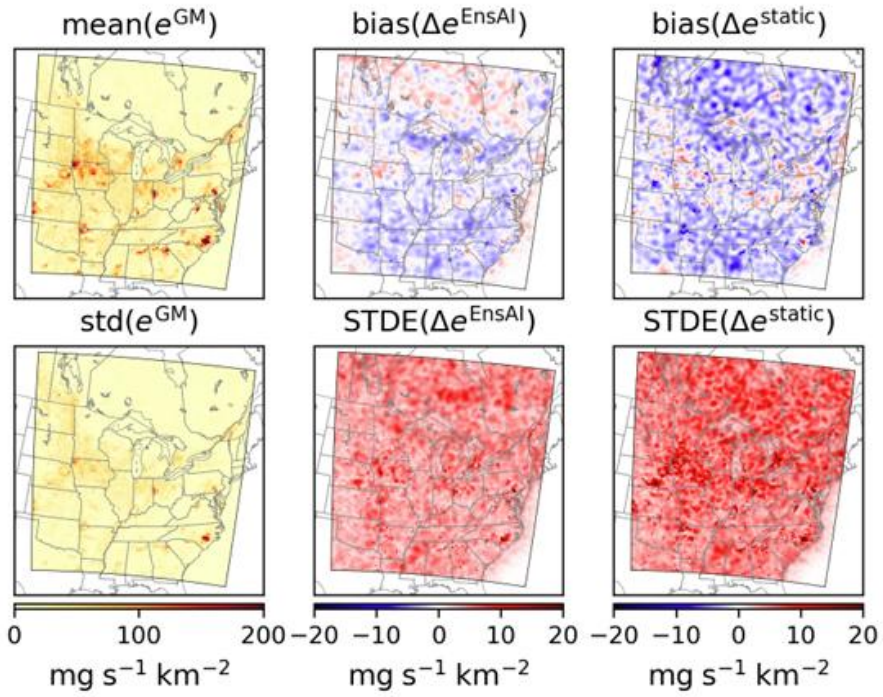


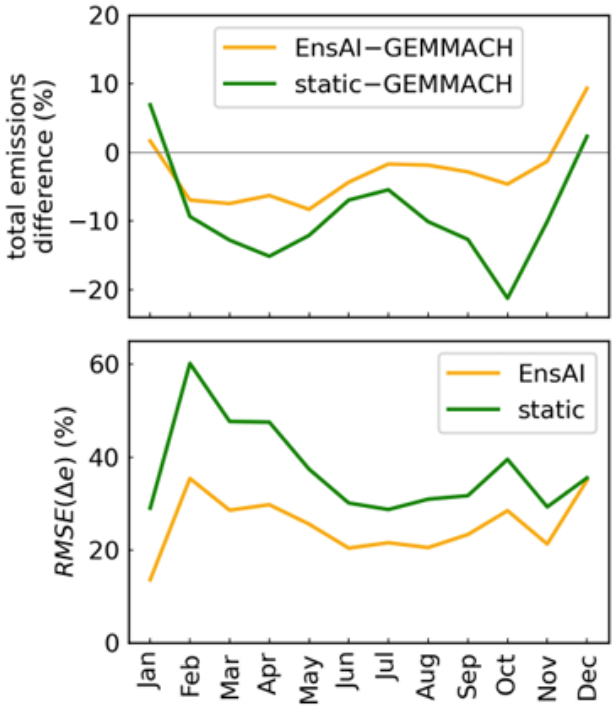
Figure 17. Statistics for the ammonia emissions inversions taken over 2015. The left column shows the mean (top panel) and standard deviation (bottom panel) of the emissions analyses from the inversions that used the GEM-MACH ensemble. The middle column shows the bias (top panel) and $STDE$ (bottom panel) 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 $\Delta e^{EnsAI} - \Delta e^{GM}$ and $\Delta e^{static} - \Delta e^{GM}$ 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 ρee, the univariate surface concentration ρcc, and…': surface concentration correlations"

Corrected.