General Assessment

The manuscript presents "EnsAI," a novel artificial intelligence-based emulator designed to accelerate the generation of atmospheric chemical ensembles, specifically targeting surface ammonia for emissions inversions. The premise is highly relevant to the scope of GMD, as reducing the computational bottleneck of ensemble-based data assimilation for atmospheric chemistry is a pressing need in the community. The application of a U-Net architecture to map emission perturbations to concentration perturbations, conditioned on basic meteorological variables, is a promising and computationally efficient approach.

However, several major concerns regarding the model's physical robustness, the generalization of the neural network architecture, and the validation strategy must be addressed before the manuscript can be considered for publication. The current evaluation successfully demonstrates that EnsAI can mimic the specific GEM-MACH ensemble it was trained on, but it falls short of proving that the model generalizes to unseen geographical/chemical conditions or truly improves physical inversion outcomes. Below is a detailed breakdown of the required improvements, incorporating both structural concerns and specific technical corrections.

Major Comments

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. 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.
2. The U-Net architecture is generally appropriate for this spatial task, but specific hyperparameter choices and training behaviors raise significant concerns about over-parameterization and overfitting. Expanding directly from 4 input channels to 128 channels in the first convolutional layer leads to a massive surge in network parameters. 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. 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, specify the optimizer used and any regularization techniques applied (e.g., weight decay, dropout), and explicitly state the activation function utilized at the final output layer. An study justifying the chosen channel depth and downsampling levels would also significantly strengthen this section.

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. The authors should discuss how well the posterior concentrations match the CrIS satellite retrievals post-inversion. 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.
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. 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.
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.
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.

Specific Corrections

1. 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.
2. 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.