Consolidated context for RC1 & RC2 comment' response to revise manuscript

Contents

1. Novelty: quantile forecasting rather than a median point forecast

RC2 appears to interpret the main result as essentially demonstrating that TFT produces a good "median forecast" and compares this with the concept of the median of an ensemble forecast.

This is not the intended methodological contribution.

The model is not trained simply to generate a conventional point prediction corresponding to the median. It produces a conditional predictive distribution represented by multiple quantiles. The 0.5 quantile is only one component of this distribution and was highlighted because it provides a natural central estimate for comparison with conventional deterministic/reference forecasts.

Therefore, the contribution should not be presented as "the median performs best". The important distinction is between:

- a point forecast;
- the median/central estimate of a probabilistic forecast generated by the TFT; and
- the median or ensemble statistic obtained by combining multiple model forecasts.

These are conceptually different quantities.

We agree that our terminology may have allowed these concepts to be conflated. We can revise the manuscript to consistently refer to the 0.5 predictive quantile rather than using “median forecast” without qualification and explain explicitly that the additional quantiles describe the conditional predictive uncertainty around this central estimate.

The revised manuscript will place greater emphasis on the probabilistic quantile formulation of the model. The current presentation does not consistently identify the central prediction as the 0.5 predictive quantile; for example, the caption of Fig. 1 refers only to TFT predictions. Although WIS evaluates probabilistic forecasts across quantiles, its interpretation is not sufficiently evident from the current presentation of Fig. 8. The caption also does not currently specify the quantiles included in the comparison.

Figure 5 shows individual quantiles, but only as a score – which implies that we can directly compare the two models. The revised presentation will clarify the underlying quantile-specific results rather than relying solely on the skill-score representation.

The extreme-event criticism should also be considered in this context. The model is not restricted to the central quantile. The upper and lower predictive quantiles are specifically intended to describe different parts of the conditional O3 distribution, including high-O3 conditions. We can strengthen the presentation of these results and avoid giving the impression that operational usefulness is being assessed solely from the 0.5 quantile.

2. Comparison of TFT uncertainty with CAMS uncertainty

RC1 suggests comparing the TFT uncertainty with the spread of the CAMS ensemble. The revised manuscript will clarify that CAMS serves as an external operational reference rather than as a methodologically identical forecasting system. The comparison provides quantitative context for TFT performance, while differences in spatial support and uncertainty representation limit direct equivalence between the two systems.

A direct comparison with CAMS ensemble spread is not required to establish the probabilistic formulation of TFT, because the TFT quantiles themselves define the conditional predictive distribution evaluated in this study.

The TFT quantiles estimate the conditional predictive distribution of O3 at a given forecast horizon, whereas the CAMS ensemble spread arises from the dispersion between the constituent atmospheric chemistry models. The latter therefore represents inter-model ensemble variability rather than the same conditional predictive uncertainty estimated through quantile regression.

For this reason, a direct numerical equivalence between TFT quantile width and CAMS ensemble spread would need to be interpreted very carefully.

In the current study, CAMS was consequently used primarily as an external forecast benchmark, with its ensemble product providing the closest available operational reference against which the central TFT prediction could be evaluated. We can make this distinction much more explicit and acknowledge that a rigorous comparison between calibrated probabilistic uncertainties from the two systems is outside the scope of the present study.

Such a comparison is possible in principle but addresses a different probabilistic verification question and is outside the scope of the present study.

3. Why TFT was selected

RC2 is correct that TFT is not the only Transformer/time-series architecture available. The purpose of the study was not to claim that TFT itself is novel.

The manuscript already considers alternative forecasting architectures in the literature review. TFT was retained for the experiments because the requirements of this study go beyond ordinary point forecasting: the model needs to handle multi-horizon forecasting, static station information, observed historical covariates, known-future covariates and probabilistic/quantile outputs in an integrated framework.

TFT provides these capabilities directly, together with variable-selection and interpretability mechanisms that are particularly useful for analysing the contributions of meteorological and atmospheric predictors.

This methodological justification will be clarified and strengthened in the Methods section. The revised text will not imply that TFT is the only architecture capable of quantile forecasting; rather, its selection will be justified by probabilistic, multi-horizon and heterogeneous-covariate requirements of this study.

4. Geographic transfer from Germany to South Korea

We agree with RC2 that the transfer-learning component needs to be made more prominent.

The intention of the South Korean experiment is not simply to show that the same TFT architecture can be retrained on another dataset. The important question is whether representations learned from the German station network remain useful after transfer to a geographically and meteorologically different domain.

The expanded discussion will also address probabilistic predictions across quantiles and their interpretation for high-O3 conditions.

The different fine-tuning strategies in Fig. 6 should therefore be discussed explicitly in terms of what is being transferred and what needs to adapt to the target domain. Fig. 6 will be revised to facilitate direct visual comparison of the three fine-tuning strategies, including consistent presentation of the RMSE curves and axis scales.

We can expand this section by discussing relevant differences between Germany and South Korea, including meteorological regimes, seasonality, station environments and differences in processes controlling O3. This would provide a physical context for why freezing or fine-tuning different portions of the network changes target-domain performance.

The feature-importance results should similarly be discussed as indicators of which predictors the trained model relies upon rather than simply reporting their ranking.

5. The approximately 11 ppb South Korean RMSE and the reported 5–10% degradation

Both reviewers raise an important issue concerning the interpretation of the South Korean RMSE.

RC1 appears to compare the approximately 11 ppb value in Fig. 6 directly with the approximately 2 ppb quantity shown in Fig. 3 and therefore concludes that the transfer error has increased by roughly a factor of four, contradicting the stated 5–10% increase.

If these quantities correspond to different evaluation definitions, aggregations, horizons, station subsets or normalisation/units, this needs to be stated extremely clearly. They should not currently appear to the reader as directly comparable RMSE values.

The revised manuscript will explicitly define the metric represented in each figure and the baseline to which the reported 5–10% increase refers.

The current figure presentation can reasonably lead readers to interpret the RMSE values in Figs. 3 and 6 as directly comparable. The revised manuscript will therefore make the respective evaluation definitions, aggregations, station subsets and baselines explicit.

Similarly, the statement that an RMSE of 11 ppb represents a “25% bias” should be corrected conceptually: RMSE is not bias. An RMSE of 11 ppb relative to a characteristic O3 concentration of 40 ppb indicates a substantial prediction error, but it does not imply an 11 ppb systematic bias unless the mean bias is separately shown to be approximately 11 ppb. The revised discussion will distinguish explicitly between RMSE and systematic bias.

Nevertheless, the magnitude of the South Korean error deserves discussion. The revised manuscript will report this magnitude transparently and clarify the change in performance under geographic transfer rather than implying that the transferred model reproduces German-domain performance unchanged. This discussion will be aligned with the manuscript's existing quantile-based evaluation, including the interpretation of upper predictive quantiles for high-O3 conditions.

6. Importance of O3 from approximately ten days previously

RC2 questions the physical interpretation of the apparent importance of O3 approximately ten days before the prediction.

This feature-importance result should not be interpreted as evidence that that an O3 molecule/concentration anomaly ten days earlier is directly causing the present-day concentration.

A lag identified as important by a deep temporal model can instead encode longer-timescale persistence and recurring atmospheric states: synoptic-scale meteorological regimes, seasonal evolution, persistent precursor/emission conditions and autocorrelated pollution episodes. Furthermore, importance/attention in a neural forecasting model should not automatically be interpreted as physical causality.

The revised manuscript will make this distinction explicit and cross-reference the existing time-window analysis in Appendix D.

7. Station selection and the 386/493 stations

The station subset was not selected according to model performance.

The reduction was required because of the computational/data-size constraints of the experiments. Stations were selected randomly subject to the requirement that the resulting

dataset maintain an approximately balanced representation of the relevant station/location categories (urban, suburban and rural) and adequate spatial coverage.

Importantly, the sensitivity to the amount of station data is already investigated through the ablation experiments reported in the Appendix. These experiments were included precisely to examine whether reducing/changing the available station sample materially changes model performance.

This needs to be brought forward into the main text because RC2's comment indicates that the purpose of that experiment is currently not sufficiently visible.

The relevant ablation analysis is currently difficult to identify in Appendix D because its relationship to station-sample sensitivity is not sufficiently explicit; hence Appendix D and its cross-referencing from the Methods will be revised to make this existing analysis directly identifiable. We can explicitly state the selection procedure, selection criteria and rationale in the Methods section and refer directly to the corresponding Appendix ablation results.

8. ERA5 versus station meteorology

RC2 suggests replacing ERA5 meteorological predictors with TOAR/site-specific meteorological observations because ERA5 has a substantially coarser spatial resolution than an individual monitoring station.

In the current TOAR Phase II dataset, site-specific meteorological observations are available only for a limited subset of stations and therefore cannot provide a consistent predictor set across the study network. ERA5 was consequently selected to provide spatially and temporally homogeneous meteorological information with broad coverage.

Using ERA5 provides a spatially and temporally consistent meteorological predictor set across the complete station network. Station meteorological observations may have higher local spatial representativeness, but their availability, completeness, variable definitions and measurement consistency can vary considerably between sites.

We can clarify why ERA5 was selected and explicitly acknowledge that gridded meteorology cannot resolve all station-scale meteorological variability. The revised manuscript will explicitly acknowledge that this consistency comes at the cost of unresolved station-scale meteorological variability.

9. ERA5 reanalysis and the CAMS operational comparison

RC1 raises a more substantive issue concerning ERA5.

The terminology will also be revised so that CAMS is described as an external operational reference, while the principal emphasis remains on probabilistic forecasting and geographic transfer.

If ERA5 reanalysis values at future forecast times were supplied to TFT as known-future meteorology, then we agree that the comparison with operational CAMS cannot be described as a strictly operational head-to-head comparison. The TFT would effectively have access to realised/reanalysis meteorological conditions, whereas CAMS necessarily propagates meteorological forecast uncertainty.

In that case, the comparison will be framed as a forecasting-skill/model-capability comparison under prescribed meteorological information rather than as operational equivalence, and state clearly that replacing future ERA5 with operational NWP forecasts such as IFS would be required for a fully operational comparison.

We will also replace the word “benchmark” to emphasise the paper more on the transfer learning, the quantile analysis, and the extreme values as per original intentions.

This limitation should be made prominent in the manuscript and conclusions.

If, however, future ERA5 meteorology was *not* supplied in this manner, then the current text has clearly caused a misunderstanding and should be rewritten to distinguish historical ERA5 inputs from genuinely known-future covariates.

10. CAMS spatial resolution and station-level comparison

Both reviewers correctly point out that CAMS and the TFT are not operating at identical spatial support.

The TFT incorporates local station observations, whereas CAMS represents concentrations on a numerical-model grid. Interpolating a gridded CTM prediction to a station does not turn it into a true station-scale forecast.

The CAMS comparison should therefore be described as comparison against an independent operational regional air-quality modelling benchmark, rather than implying a perfectly controlled model-versus-model comparison.

We can explicitly acknowledge the representativeness mismatch and mention that a locally bias-corrected/MOS CAMS product would constitute a more directly comparable station-scale operational benchmark where sufficiently consistent data are available.

This limitation does not invalidate the comparison, but it does constrain the conclusions that should be drawn from it.

11. Input variables, architecture and Appendix material

RC2 also found the Methods difficult to follow because important information is currently located in the Appendix. This is straightforward to address.

Detailed descriptions can remain in the Appendix because of manuscript-length constraints, while the main paper will provide a compact summary of the model inputs, their sources and their roles in the TFT in revised submission .

For example, Table A1 could be extended/restructured to provide variable abbreviation, full variable name, source and model role, while the exhaustive descriptions remain in the Appendix. Similarly, the main text can contain a concise architectural description explaining the static inputs, historical inputs, known-future inputs, variable-selection networks, temporal processing and quantile outputs.

12. Figures, terminology and language

We agree with RC2's comments regarding figure quality, ordering, terminology, font size and general language editing.

These do not require additional model experiments. We can:

- standardise “forecast horizon”, “forecast step” or another single term throughout;
- increase figure resolution and font sizes;
- ensure every figure is introduced before it appears;
- correct the ordering of Figs. 6 and 8;
- expand captions so the figures are independently interpretable and
- carefully proofread punctuation, spacing, capitalisation, italics and general English.

These revisions are all feasible within a manuscript revision.