

Response

#Revier 1

5 **Note on the organization of this response:** Reviewer comments are shown in blue, while text quoted from the manuscript is shown in red. For each comment, we provide the reviewer’s original comment, our response, the relevant original manuscript text, and the corresponding revised text. Because we have included both the original and revised manuscript passages to make all changes fully traceable, the response document is relatively long.

10 **Response for summary:**

We thank the reviewer for the positive and encouraging summary of our work. We appreciate the recognition of the motivation, dataset design, validation strategy, product intercomparisons, and CRE applications of CSRFormer. We have carefully considered all of the reviewer’s comments and have revised the manuscript accordingly to further strengthen the methodological description, validation, and discussion of the dataset.

15

Major comments 1

The 7:3 train/val split in 3.1 isn't described in enough detail — is it random at the sample level, or blocked by day/orbit? If it's a random split, footprints from the same overpass are strongly autocorrelated (similar AOD, humidity, SST across a swath), so some of that validation performance is probably leakage rather than genuine generalization. That would also explain the pattern in the results: SW R^2 drops from 0.93 (val) to 0.81 (2018 test) while LW barely moves ($0.99 \rightarrow 0.99$) — SW is more sensitive to synoptic-scale aerosol structure, which is exactly what a random split would leak across train/val. Worth clarifying how the split was actually done, and if it is sample-level random, a day-blocked validation run (even on a subset) would give a more honest read on how much the val numbers overstate true skill.

25 **Response:**

We thank the reviewer for pointing out this important issue. In the original manuscript, the training and validation samples from 2019–2020 were randomly divided at the footprint level. We agree that this strategy can lead to overly optimistic validation performance because spatially and temporally adjacent CERES footprints may share similar aerosol, thermodynamic, and surface conditions.

30

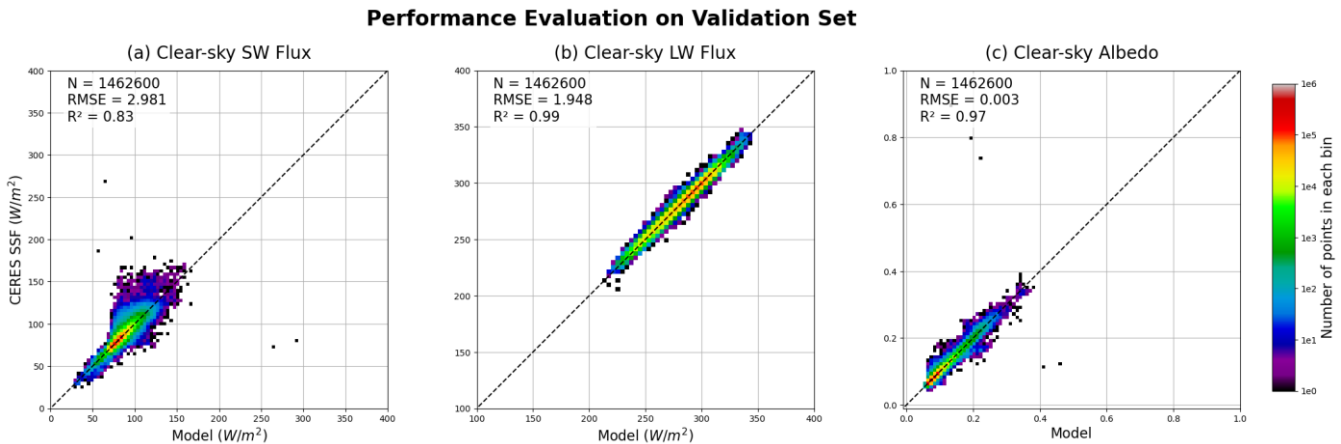
Following the reviewer’s suggestion, we have revised the data-splitting strategy and performed a new **day-blocked validation experiment**, in which all footprints from the same day are assigned exclusively to either the training or validation set. This

prevents samples from the same daily atmospheric situation from appearing in both subsets and therefore provides a more stringent assessment of generalization.

35

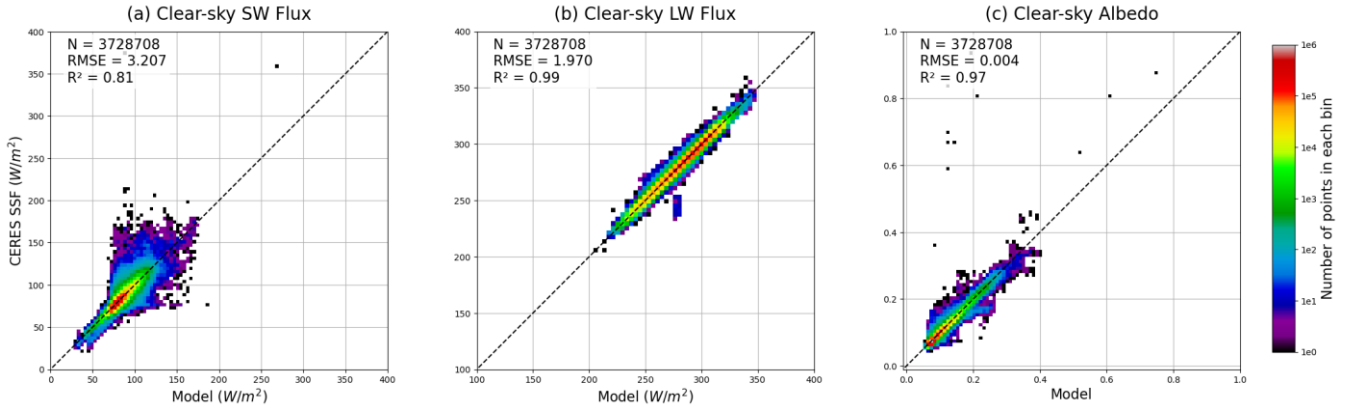
After applying the day-blocked split, the validation performance decreased relative to that obtained with the original random split, particularly for shortwave fluxes. Importantly, however, the resulting validation performance became very similar to that obtained on the fully independent 2018 test set, as shown in Figure R1 (validation set) and R2 (test set). This agreement indicates that the higher performance reported for the original random validation set was partly affected by spatiotemporal autocorrelation, whereas the revised day-blocked validation provides a more realistic estimate of out-of-sample performance. The close consistency between the revised validation results and the independent-year test results also supports the robustness of CSRFormer across temporally separated samples.

40



45 **Figure R1. Probability density distributions for the independent validation set. The x-axis shows CSRFormer predictions and the y-axis shows CERES SSF clear-sky reference values. Greater proximity to the 1:1 line indicates better agreement. (a) Clear-sky shortwave flux; (b) clear-sky longwave flux; (c) clear-sky albedo.**

Performance Evaluation on Test Set



50 **Figure R2. Probability density distributions for the independent test set. The x-axis shows CSRFormer predictions and the y-axis shows CERES SSF clear-sky reference values. Greater proximity to the 1:1 line indicates better agreement. (a) Clear-sky shortwave flux; (b) clear-sky longwave flux; (c) clear-sky albedo.**

We have therefore replaced the original validation results with the day-blocked validation results and revised the corresponding figure and text in the manuscript.

We would also like to correct an error in the original manuscript regarding the train/validation ratio. After rechecking the data-splitting code, we confirmed that an **8:2 split**, rather than the stated 7:3 split, was used in all of our experiments, including the original training and validation. We have corrected this description in the revised manuscript. Because the actual experiments were consistently performed using the 8:2 split, this correction does not affect the reported model results. Given the very large sample size in both subsets, the difference between a 7:3 and 8:2 ratio is also not expected to materially affect the training or validation performance; the more important issue is whether temporally correlated samples are separated between the two subsets.

65 **Changes in the manuscript:**

Section 2.2:

Original:

We selected input variables that describe both surface conditions and the vertical atmospheric state. For convenience, we group them into planar inputs and vertical-profile inputs. Clear-sky samples from 2019 and 2020 were divided into a training set and a validation set at a ratio of 7:3, while data from 2018 were used as the test set.

Revised:

We selected input variables that describe both surface conditions and the vertical atmospheric state. For convenience, we group them into planar inputs and vertical-profile inputs. Clear-sky samples from 2019 and 2020 were divided into training and validation sets at a ratio of 8:2 using a day-blocked strategy, whereby all CERES SSF footprints from the same UTC day were assigned exclusively to either the training or validation set. Data from 2018 were retained as a fully independent test set.

Section 3.1:

Added:

To reduce potential information leakage caused by the strong spatiotemporal autocorrelation among neighboring CERES footprints, the 2019–2020 samples were split into training and validation sets using a day-blocked strategy. Specifically, calendar days were assigned to the training and validation subsets at an approximately 8:2 ratio, and all footprints from a given UTC day were kept within the same subset. Therefore, no footprints from the same day were shared between the training and validation sets. The 2018 samples were not involved in model development and were used exclusively as an independent test set. All normalization parameters were calculated using the training subset only and subsequently applied to the validation and test data.

Section 4.1:

Original:

We evaluated CSRFormer by comparing predicted clear-sky radiative fluxes with CERES SSF clear-sky reference values at the pixel level for the validation and independent test sets. Figures 2 and 3 display the scatter density distributions and error statistics of the model on the Validation Set and Test Set, respectively. Across approximately 1.4 million samples, predictions for both SW and LW fluxes clustered closely around the 1:1 line. For clear-sky shortwave radiation in validation set (Fig. 2a), the model predictions were highly consistent with observations, with a R^2 as high as 0.93 and a RMSE of only 2.044 W/m², showing no obvious overestimation or underestimation. For clear-sky longwave radiation (Fig. 2b), the model also performed well, with R^2 reaching 0.99 and RMSE further reducing to 1.730 W/m². For the calculated clear-sky albedo (Fig. 2c), R^2 reached 0.99 with an RMSE of 0.002. This result indicates that the current model performance has exceeded that of the prediction of clear-sky albedo achieved based on an MLP model in Zheng et al. (2025) (R^2 of 0.96, RMSE of 0.004).

Revised:

We evaluated CSRFormer by comparing predicted clear-sky radiative fluxes with CERES SSF clear-sky reference values at the pixel level for the validation and independent test sets. Figures 2 and 3 display the scatter density distributions and error statistics of the model on the Validation Set and Test Set, respectively. Across approximately 1.4 million samples, predictions for both SW and LW fluxes clustered closely around the 1:1 line. For clear-sky shortwave radiation in validation set (Fig. 2a), the model predictions were highly consistent with observations, with a R^2 as high as 0.83 and a RMSE of only 2.981 W/m², showing no obvious overestimation or underestimation. For clear-sky longwave radiation (Fig. 2b), the model also performed

well, with R^2 reaching 0.99 and RMSE further reducing to 1.948 W/m². For the calculated clear-sky albedo (Fig. 2c), R^2 reached 0.97 with an RMSE of 0.003. This result indicates that the current model performance has exceeded that of the prediction of clear-sky albedo achieved based on an MLP model in Zheng et al. (2025) (R^2 of 0.96, RMSE of 0.004).

Original:

In the independent test set not involved in training (Fig. 3a), SW performance decreased modestly ($R^2 = 0.81$, RMSE = 3.207 W m⁻²), likely reflecting the higher sensitivity of shortwave fluxes to aerosol and surface conditions. Additionally, 3D cloud radiative effects—specifically, 'brightening' from lateral scattering by adjacent clouds—contribute to shortwave prediction biases. While CERES faithfully records these non-local enhancements (Loeb and Davies, 1996; Várnai and Marshak, 2002; Tijhuis et al., 2024), CSRFormer relies on single-point atmospheric column inputs lacking surrounding 3D cloud geometry. This observation-input asymmetry limits the model's ability to capture cloud adjacency effects. In contrast, the generalization performance for longwave radiation (Fig. 3b) appeared excellent and stable. R^2 on the test set remained at the high level of 0.99, and RMSE only increased marginally to 1.970 W/m², indicating stable performance across the held-out samples. Clear-sky albedo also remained in close agreement with the reference values ($R^2 = 0.97$, RMSE = 0.004).

Revised:

In the independent test set not involved in training (Fig. 3), the model maintained stable predictive performance. For clear-sky SW radiation (Fig. 3a), CSRFormer achieved an R^2 of 0.81 and an RMSE of 3.207 W m⁻², while clear-sky albedo remained in close agreement with the reference values ($R^2=0.97$, RMSE = 0.004). For clear-sky LW radiation (Fig. 3b), the model showed even stronger agreement with CERES SSF, with an R^2 of 0.99 and an RMSE of 1.970 W m⁻². The lower predictive skill for SW radiation relative to LW radiation may partly reflect its greater sensitivity to three-dimensional cloud radiative effects. In particular, lateral scattering by nearby clouds can enhance the observed SW flux in nominally clear-sky footprints (Loeb and Davies, 1996; Várnai and Marshak, 2002; Tijhuis et al., 2024). These non-local effects are recorded by CERES observations but cannot be explicitly represented by CSRFormer, which relies on single-column atmospheric inputs without surrounding cloud geometry. In contrast, LW radiation is less affected by such lateral scattering effects, which likely contributes to its more stable performance across the validation and independent test sets.

Figure 2 and 3:

We have replace original figures from Figure R1 and R2.

Major comments 2

I'd like to see more evidence that the transformer itself is doing the work, not just the physical constraint in Eq. 1. LW is pinned to σT_{skin}^4 with SST also fed in directly, and Fig. 5 shows SST as by far the dominant LW predictor — which tracks with the paper's own citations (Allan et al., 2007; Koll and Cronin, 2018) that clear-sky OLR is already close to linear in surface

temperature and column water vapor. Given that, my guess is a good chunk of the $R^2=0.99$ is just the T^4 relationship doing the heavy lifting, not the attention mechanism resolving the 37+72-level profiles — which is supposed to be the paper's main advance over the MLP in Zheng et al. (2025). A simple baseline comparison (σT_{skin}^4 minus a plain SST/TCWV regression for ΔLW , no transformer) using data already in hand would make the case for the architecture much more convincing.

145

Response:

We thank the reviewer for this valuable suggestion. To quantify how much of the longwave predictive skill can be explained by the Stefan–Boltzmann constraint and bulk thermodynamic variables alone, we constructed the suggested baseline model. As we used in our manuscript:

150

$$LW = \sigma T_{skin}^4 - \Delta LW_{pred}$$

the ΔLW_{pred} is the CSRFormer's output. But here we used a multiple linear regression with SST and TCWV as ΔLW_{pred} . So the corresponding clear-sky LW flux was then reconstructed as:

$$LW = \sigma T_{skin}^4 - \Delta LW_{regression}$$

The baseline showed very stable performance across the training, validation, and independent 2018 samples, with R^2 values of 0.911, 0.911, and 0.912 and RMSE values of 7.31, 7.37, and 7.30 $W\ m^{-2}$, respectively. These results confirm the reviewer's expectation that surface temperature and column water vapor, together with the Stefan–Boltzmann relationship, explain a substantial fraction of the large-scale variability in clear-sky LW flux.

155

However, this simple physical baseline remains substantially less accurate than CSRFormer. On the same independent test set, the baseline yields $R^2=0.912$ and $RMSE = 7.30\ W\ m^{-2}$, whereas CSRFormer achieves $R^2=0.99$ and $RMSE = 1.97\ W\ m^{-2}$ on the same sample. Thus, the RMSE is reduced by approximately 73% relative to the regression SST–TCWV baseline.

160

We interpret this comparison as showing two complementary advantages of the full CSRFormer framework. First, CSRFormer uses the vertically resolved ERA5 humidity and temperature profiles rather than reducing the atmospheric state to a single column-integrated moisture quantity such as TCWV. Second, the Transformer framework provides substantially nonlinear representational flexibility, allowing the model to learn nonlinear interactions among surface conditions and the vertically varying atmospheric state. Therefore, the high LW skill cannot be attributed solely to the imposed σT_{skin}^4 constraint or to a simple linear dependence on SST and column water vapor.

165

At the same time, we agree that this comparison does not uniquely isolate the contribution of the attention mechanism itself, because the baseline and CSRFormer differ in both input dimensionality and model complexity. We have therefore revised the manuscript to avoid attributing the full improvement solely to the Transformer architecture. Instead, we interpret the result as

170

evidence that the combination of vertically resolved atmospheric information and the nonlinear CSRFormer framework provides substantial additional predictive skill beyond the simple physically constrained SST–TCWV baseline.

175

Changes in the manuscript:

Section 3.3:

Added:

180

An additional physical baseline was constructed to examine whether the LW prediction skill primarily arises from the Stefan–Boltzmann constraint and bulk thermodynamic variability. The atmospheric attenuation term in equation (1), ΔLW_{pred} , was fitted using multiple linear regression with SST and total column water vapor (TCWV) as predictors, and the clear-sky LW flux was subsequently reconstructed with equation (1). The baseline performance was evaluated by comparing its R^2 and RMSE with those of CSRFormer.

185

Section 4.2:

Added:

190

As an additional test of whether the high longwave prediction skill is primarily determined by the Stefan–Boltzmann constraint, we compared CSRFormer with a simple SST–TCWV linear baseline for the atmospheric attenuation term. The baseline achieved R^2 values of 0.911, 0.911, and 0.912 and RMSE values of 7.31, 7.37, and 7.30 W m^{-2} for the training, validation, and independent 2018 test set, respectively. These results show that SST and column-integrated water vapor, together with the σT_{skin}^4 constraint, explain a substantial fraction of the large-scale variability in clear-sky LW flux. However, the remaining errors are substantially larger than those of CSRFormer. This indicates that the high predictive skill of CSRFormer cannot be explained by the physical constraint and bulk thermodynamic variables alone. The additional performance is associated with both the use of vertically resolved atmospheric information and the nonlinear representational capacity of the CSRFormer framework.

195

Minor comments 1

3.2: "trained with a learning rate of in a 5×10^{-5} " reads garbled — please check.

200

Response:

Thank you for pointing this out. This was a typographical error. The sentence has been corrected to:

The model was trained with a learning rate of 5×10^{-5} using the AdamW optimizer for 150 epochs and a batch size of 128.

Minor comments 2

205

Eq. 2 weights ($\alpha = 7, 3, 2, 0.5, 3$): a sentence on how these were set (tuned vs. manual) would help anyone reusing the loss design.

Response:

Thank you for the suggestion. The loss weights were not determined through an exhaustive hyperparameter search. Instead, they were manually selected based on preliminary experiments, with the aim of balancing the magnitudes of the different loss components while maintaining stable optimization of both SW and LW predictions. We have added a sentence in Section 3.2 to clarify this point.

Changes in the manuscript:

Section 3.2:

Original:

The weighting parameters α_n ($n = 1, 2, \dots, 5$) are set to 7, 3, 2, 0.5 and 3, respectively. The model was trained with a learning rate of 5×10^{-5} using the AdamW optimizer for 150 epochs and a batch size of 128, utilizing dropout (rate is 0.1) and weight decay ($1e^{-5}$) for regularization (Srivastava et al., 2014; Loshchilov and Hutter, 2017). The planar inputs are SST, AOD, wind speed, and SZA. The vertical-profile inputs are ERA5 temperature, humidity, and ozone profiles (37 levels) and MERRA-2 aerosol profiles (72 levels). The shared encoder uses 128 embedding dimensions, 4 Transformer layers, and 8 attention heads.

Added:

The weighting parameters α_n ($n = 1, 2, \dots, 5$) are set to 7, 3, 2, 0.5 and 3, respectively. **The loss weights were manually selected based on preliminary experiments to balance the relative magnitudes of the individual loss components and to maintain stable joint optimization of the SW and LW objectives.** The model was trained with a learning rate of 5×10^{-5} using the AdamW optimizer for 150 epochs and a batch size of 128, utilizing dropout (rate is 0.1) and weight decay ($1e^{-5}$) for regularization (Srivastava et al., 2014; Loshchilov and Hutter, 2017). The planar inputs are SST, AOD, wind speed, and SZA. The vertical-profile inputs are ERA5 temperature, humidity, and ozone profiles (37 levels) and MERRA-2 aerosol profiles (72 levels). The shared encoder uses 128 embedding dimensions, 4 Transformer layers, and 8 attention heads.

Minor comments 3

Fig. A6a/b: ERA5 SW comparison gives $R^2 = -2.06$ and -2.40 — a negative R^2 this large is striking enough to merit one explicit sentence confirming it's not a units/product-definition mismatch, alongside the existing physical explanation (idealized albedo, climatological aerosol).

Response:

Thank you for pointing this out. We agree that the strongly negative R^2 values in Fig. A6a–b warrant an explicit verification. We rechecked both the ERA5 variables and the processing code used in this comparison. ERA5 clear-sky reflected SW flux at the top of the atmosphere was calculated from the accumulated ERA5 fields as:

$$SW_{\text{ERA5,clr}} = (\text{TISR-TSRC})/3600$$

with units of W m^{-2} . We also tested alternative incorrect definitions, including net shortwave flux, reversed sign, and omission of the conversion from accumulated energy to flux; all produced substantially worse agreement. In addition, the strongly negative R^2 values remained stable when the comparison was repeated using random subsets of 1,000, 10,000, and 100,000 samples. We therefore confirm that the results are not caused by a unit conversion, sign convention, or net-versus-reflected-flux mismatch.

Further examination indicates that the negative R^2 mainly reflects differences in the temporal representation and dynamic range of the SW fluxes. ERA5 radiation fields represent hourly accumulated/mean quantities, whereas CERES SSF and CSRFormer represent instantaneous footprint-scale fluxes at the satellite observation time. The ERA5 clear-sky SW distribution is consequently narrower and systematically shifted relative to the CERES-based reference. Consistent with this interpretation, ERA5 clear-sky albedo still shows reasonably good agreement with CERES/CSRFormer ($R^2 \approx 0.85\text{--}0.86$), indicating that the very poor flux-level R^2 is not due to a gross mismatch in the reflected-shortwave definition. These temporal-sampling differences, together with ERA5's parameterized treatment of ocean-surface albedo and climatological aerosols, cause the ERA5 shortwave flux to reproduce the instantaneous CERES variability poorly enough that the residual error exceeds the variance of the reference data, resulting in $R^2 < 0$.

Changes in the manuscript:

Section 4.3.3:

Original:

For ERA5, the contrast between the longwave and shortwave bands is particularly clear. As shown in Fig. A6 and A7, ERA5 longwave clear-sky fluxes remain close to CSRFormer and CERES SSF, which is expected because CSRFormer uses ERA5 temperature, humidity, and ozone profiles as key inputs, so both products are anchored to a similar atmospheric thermodynamic state. In the shortwave band, however, substantial differences persist. **These are likely related to ERA5's idealized treatment of ocean-surface albedo and its use of climatological aerosol inputs (Tegen et al., 1997; Taylor et al., 2006; Hersbach et al., 2020), whereas CSRFormer uses dynamic SST, wind speed, and aerosol information to reconstruct the instantaneous shortwave response more flexibly.**

Revised:

For ERA5, the contrast between the longwave and shortwave bands is particularly clear. As shown in Fig. A6 and A7, ERA5 longwave clear-sky fluxes remain close to CSRFormer and CERES SSF, which is expected because CSRFormer uses ERA5 temperature, humidity, and ozone profiles as key inputs, so both products are anchored to a similar atmospheric thermodynamic state. In the shortwave band, however, substantial differences persist. **These differences likely arise from both the temporal representation of the radiation fields and the treatment of key shortwave radiative factors. ERA5 radiation fields represent**

275 hourly accumulated or mean quantities, whereas CERES SSF and CSRFormer correspond to instantaneous footprint-scale
fluxes at the satellite observation time. When the reflected SW flux is normalized by the corresponding incoming solar radiation,
the clear-sky albedo comparison shows substantially better agreement ($R^2=0.85$), further indicating that the very poor flux-
level R^2 does not result from a gross inconsistency in the shortwave flux definition. At the same time, ERA5 employs an
idealized treatment of ocean-surface albedo and climatological aerosol inputs (Tegen et al., 1997; Taylor et al., 2006; Hersbach
280 et al., 2020), whereas CSRFormer uses dynamic SST, wind speed, and aerosol information to reconstruct the instantaneous
shortwave response more flexibly. The purpose of this comparison is therefore not to rank the products by overall accuracy,
but to characterize the differences that arise when commonly used hourly reanalysis radiation fields are compared with
instantaneous satellite-footprint observations. The results highlight that temporal representation should be considered carefully
when such products are used for instantaneous radiative analyses.

285

Minor comments 4

The archived NetCDF (§4.5) carries flux + clear-sky fraction only. Since error clearly varies by regime (aerosol load, cloud-
edge "brightening"), a per-footprint uncertainty or QC flag would make v1.0 more usable for downstream CRE error
propagation.

290

Response:

We agree that providing information on prediction uncertainty would improve the usability of the dataset. CSRFormer is
currently formulated as a deterministic model rather than a probabilistic or ensemble model, and therefore it does not directly
provide a predictive distribution or ensemble spread from which a calibrated footprint-level uncertainty could be derived. A
rigorous propagation of input uncertainty is also difficult because the true error distributions of the driving fields are not known
295 for individual footprints, particularly under cloudy conditions. For example, aerosol loading may contain region- and regime-
dependent uncertainties, while non-local three-dimensional effects such as cloud-adjacency brightening introduce an additional
observation-model representation mismatch that cannot readily be expressed as an input perturbation of known magnitude.

300 For these reasons, we consider that assigning a formal numerical uncertainty to every CSRFormer prediction would require
assumptions that are not sufficiently constrained by the available observations. Instead, motivated also by Reviewer 2's concern
regarding the representativeness of the clear-sky training population, we use the training-domain analysis to provide a more
directly interpretable quality-control indicator for each footprint. Specifically, training support is evaluated in SST-TCWV
space for LW flux and in AOD-SZA space for SW flux. The resulting flags indicate whether the atmospheric state of a
305 prediction falls within a well-supported or low-support region of the clear-sky training population.

Therefore, motivated also by Reviewer 2's concern regarding the representativeness of the clear-sky training population, we
introduced a training-domain support QC flag based directly on the empirical sampling density of the clear-sky training

population. For LW flux, training support is evaluated in SST-TCWV space, because SST and atmospheric water vapor provide a compact description of the dominant bulk thermodynamic controls on clear-sky outgoing longwave radiation. For SW flux, support is evaluated in AOD-SZA space, because aerosol loading and solar geometry are the two dominant controls identified by the feature-importance analysis. In each two-dimensional state space, the 2019-2020 clear-sky training samples were binned and the number of training samples in each bin, N_{train} , was counted. Each predicted footprint is then assigned the support level of the bin corresponding to its instantaneous atmospheric state. This procedure is described in Sect. 3.3.

315

We use $N_{\text{train}} \geq 50$ as an operational threshold for a well-supported regime. Under this criterion, 91.41% of all CERES SSF ocean samples and 82.16% of overcast samples fall within the well-supported SST-TCWV domain. For the overcast population, 5.38% of samples occur in bins containing 1-49 clear-sky training samples and 12.46% occur in bins containing no clear-sky training samples. For SW radiation, the overlap is higher: 96.4% of all SSF samples and 93.6% of overcast samples fall within well-supported AOD-SZA bins. Sensitivity tests using alternative N_{train} thresholds produced the same qualitative conclusion that most application states are represented in the clear-sky training population, although the coverage is less complete for overcast LW conditions. These results are presented in Sect. 4.2 and Figs. 4-5, with the geographical distribution of low-support overcast conditions further shown in Fig. A6.

325

For practical use of the released dataset, we intentionally adopted a simple two-level QC flag. A flag value of 0 denotes a well-supported regime $N_{\text{train}} \geq 50$, whereas a value of 1 denotes a low-support regime $N_{\text{train}} < 50$. Thus, both sparsely sampled bins and completely unsampled bins are conservatively grouped into the same low-support category in the distributed product. The more detailed distinction between sparsely sampled and unsampled regimes is retained in the statistical analysis in Sect. 4.2, but is not encoded as separate QC levels in order to keep the footprint-level flag straightforward for users.

330

We emphasize that this QC flag is not a calibrated prediction uncertainty. Rather, it indicates the empirical support of each prediction within the training domain and can be used to identify or screen footprints for which CSRFormer is operating under relatively weak training-data support. Other sources of uncertainty, including errors in the driving datasets and three-dimensional cloud-adjacency effects, are not represented by this flag. The definition and interpretation of the QC variables included in the released dataset are now provided in Sect. 4.5 (Dataset specifications).

340

345

350

355 **Reference:**

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, 10.1002/qj.3803, 2020.

Loeb, N. G. and Davies, R.: Observational evidence of plane parallel model biases: Apparent dependence of cloud optical depth on solar zenith angle, Journal of Geophysical Research: Atmospheres, 101, 1621-1634, 10.1029/95jd03298, 1996.

Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101, 2017.

365 Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research, 15, 1929-1958, 2014.

Taylor, J. P., Edwards, J. M., Glew, M. D., Hignett, P., and Slingo, A.: Studies with a flexible new radiation code. II: Comparisons with aircraft short-wave observations, Quarterly Journal of the Royal Meteorological Society, 122, 839-861, 10.1002/qj.49712253204, 2006.

370 Tegen, I., Hollrig, P., Chin, M., Fung, I., Jacob, D., and Penner, J.: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results, Journal of Geophysical Research: Atmospheres, 102, 23895-23915, 10.1029/97jd01864, 1997.

Tijhuis, M., van Stratum, B. J. H., and van Heerwaarden, C. C.: The impact of coupled 3D shortwave radiative transfer on surface radiation and cumulus clouds over land, Atmospheric Chemistry and Physics, 24, 10567-10582, 10.5194/acp-24-10567-2024, 2024.

375 Várnai, T. and Marshak, A.: Observations of Three-Dimensional Radiative Effects that Influence MODIS Cloud Optical Thickness Retrievals, Journal of the Atmospheric Sciences, 59, 1607-1618, [https://doi.org/10.1175/1520-0469\(2002\)059<1607:OOTDRE>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<1607:OOTDRE>2.0.CO;2), 2002.

Zheng, B., Cao, Y., Huang, K. E., Liu, J., Wang, Y., Zhu, Y., Wang, M., Rosenfeld, D., Zhou, C., and Huang, Y.: Estimation of Global Ocean TOA Instantaneous Clear-Sky Albedo From CERES for Shortwave Cloud Radiative Effect Analysis Based on a Deep Learning Model, Journal of Geophysical Research: Atmospheres, 130, 10.1029/2025jd044878, 2025.

380