Implementation of a machine-learned gas optics parameterization in the ECMWF Integrated Forecasting System

Peter Ukkonen[1], **Robin J. Hogan**[2]

[1]Danish Meteorological Institute

[2]European Centre for Medium-Range Weather Forecasts

Key Points:

- A machine-learned gas optics scheme, RRTMGP-NN, was implemented in a global weather model
- Compared to the original RRTMGP gas optics, RRTMGP-NN speeds up the ecRad radiation scheme by roughly 30%
- In 4 years of free-running simulations, RRTMGP and RRTMGP-NN are closer in model climate than different versions of RRTMGP are to each other

Corresponding author: Peter Ukkonen, `puk@dmi.dk`

Abstract

Radiative transfer parameterizations are physically important but computationally expensive components of weather and climate models. In previous work, it was demonstrated that the gas optics module of a radiation scheme, which traditionally rely on look-up-tables, can be replaced with neural networks (NN) to improve speed while retaining a high degree of accuracy. However, the evaluation of the NN version of the RRTMGP gas optics scheme (RRTMGP-NN) was based solely on offline radiation computations.

In this paper, we describe the implementation and prognostic evaluation of RRTMGP-NN in the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF). This was carried out by incorporating the gas optics scheme into ecRad, the modular radiation scheme used in the IFS. New NN models were trained on RRTMGP k-distributions with reduced spectral resolution. A hybrid loss function helped reduce radiative forcing errors.

Four 1-year coupled ocean-atmosphere simulations were performed for each configuration. The results show that RRTMGP-NN and RRTMGP produce very similar model climates, with the differences being smaller than those between existing gas optics schemes, and statistically insignificant for zonal means of single-level quantities such as surface temperature. The use of RRTMGP-NN speeds up the radiation scheme by roughly a third compared to RRTMGP, and is also faster than the older and less accurate RRTMG which is used in the current operational cycle of the IFS.

Plain Language Summary

The use of machine learning in weather and climate models is a growing research area, as it has the potential to reduce the computational cost of expensive simulations and improve the representation of small-scale physical processes. However, it also faces a multitude of challenges, such as making reliable predictions across a wide range of conditions (including warmer climates), being stable, and not violating conservation laws. In this study we use neural networks for a problem which does not incorporate physical equations and is more empirical, therefore avoiding these issues. Neural networks were implemented in a global weather model to predict the optical properties of the atmosphere, a component of atmospheric radiation computations that is fundamental for predicting future changes in climate due to changes in greenhouse gas concentrations. Yearly sim-

ulations show that this approach has no clear adverse impact compared to the original method it replaces, while speeding up the radiation computations by around 30%.

1 Introduction

Although atmospheric radiation is well understood and very accurate solutions are available, atmospheric models need to settle for a trade-off in the accuracy and cost of radiation computations. This trade-off can be controlled via many factors like the temporal and spatial frequency of computations (Hogan & Bozzo, 2018), simplifying assumptions (e.g. neglecting 3D effects), and spectral resolution (Hogan & Matricardi, 2020). To reduce the latter, most modern radiation schemes use the correlated-k-distribution method (e.g., Goody et al., 1989) since it allows computing broadband fluxes with high accuracy using only O(10^2 - 10^3) quadrature points, compared with O(10^6 - 10^7) of line-by-line methods which resolve individual spectral lines in the absorption spectra of atmospheric gases.

Despite this, computations remain expensive enough that many other of the aforementioned approximations need to be made, and still large-scale climate simulations can spend half of the total model runtime on radiation computations (Cotronei & Slawig, 2020). To make better use of computer resources in an era where computer hardware is becoming more heterogenous, and the gap between the theoretical peak performance and the performance of typical physics codes is probably increasing, machine learning (ML) is a promising way to simultaneously address computational challenges and potentially reduce model uncertainty by representing sub-grid processes more realistically.

Indeed, interest in the use of ML for parameterization of sub-grid processes has been growing, with a particular focus on learning convection or unified physics parameterizations from high-resolution simulations (Rasp et al., 2018; Brenowitz & Bretherton, 2018; Gentine et al., 2018; Brenowitz et al., 2020; Yuval et al., 2021). Using ML specifically for atmospheric radiation has a long history (Chevallier et al., 1998; V. M. Krasnopolsky et al., 2008; V. Krasnopolsky et al., 2010; Pal et al., 2019; Liu et al., 2020; Roh & Song, 2020; Song & Roh, 2021). These studies have attempted to replace the entire radiation scheme with a feed-forward neural network (FNN). An alternative approach, which offers better accuracy at the cost of a smaller speed-up, is to keep the radiative transfer solver intact and only replace the computation of gas optical properties with NNs.

This may be justified by considering that radiation schemes solve the radiative transfer equation (using the two-stream approximation) to obtain accurate estimates of broadband radiative fluxes. In contrast, parameterizations of other sub-grid processes, such as convection, are often more heavily based on empirical relationships and simplified theories (Wang et al., 2022). The computation of optical properties (which control the absorption, emission and scattering of radiation) within radiation schemes are likewise based on empiricism, as they rely on look-up-tables, making it a highly suitable problem for machine learning.

FNNs were consequently developed to emulate the RRTMGP gas optics scheme (Pincus et al., 2019) in two different studies, which found speed-ups of 2-6x compared to the original code (Ukkonen et al., 2020; Veerman et al., 2021). The NN gas optics was combined with a refactored radiative transfer solver to speed up the entire radiation scheme (without clouds or aerosols) by a factor of 1.8 - 3.5 in Ukkonen et al. (2020). Recently, Ukkonen (2022) compared different emulation strategies for shortwave radiation, and found that using NNs for gas optics did not sacrifice almost any accuracy, whereas replacing the entire scheme with FNNs was the fastest but also least accurate approach, with heating rates (computed from predicted broadband fluxes) having a root-mean-square-error (RMSE) of 1.35 K day^{-1}. An interesting middle-ground approach was found in using bidirectional recurrent neural networks to emulate the full radiation scheme, which produced far more accurate fluxes and heating rates (RMSE 0.16 K day^{-1}) than FNNs while offering a much smaller, but still significant speedup.

While these results confirm that gas optics emulation is more ready for operational implementation than replacing the entire radiation scheme with NNs, not least due to inherently better generalization (e.g. to arbitrary vertical grids), the evaluations were based on offline radiation computations (Ukkonen et al., 2020; Veerman et al., 2021; Ukkonen, 2022). In this study, the NN version of RRTMGP (Ukkonen et al., 2020) is integrated into the ecRad radiation scheme used in the Integrated Forecasting System (IFS), which is a global numerical weather prediction model developed at the European Centre for Medium-Range Weather Forecasts (ECMWF). New NN models are trained on reduced RRTMGP k-distributions that recently became available, which have around the same number of k-terms as the older RRTMG scheme that is used operationally in the IFS. Free-running simulations are then performed in order to test the generalization and accuracy of the NNs in a prognostic setting. In the context of wider literature on

ML-based parameterizations, our focus on gas optics represents incrementalism and aims for immediate application in numerical weather prediction (NWP) and climate modelling. On the other hand, the gas optics scheme plays a very central role in climate models as pointed out by Hogan and Matricardi (2022) and is computationally a significant part of the full radiation scheme (Ukkonen et al., 2020; Hogan & Bozzo, 2018).

The structure of the paper is as follows: Section 2 briefly describes the ecRad and RRTMGP-NN codes, and the implementation of RRTMGP-NN in ecRad. Section 3 provides an overview of the machine learning methodology, which has been refined to reproduce radiative forcings with respect to individual gases more accurately. The results are then presented in Section 4, consisting of an offline evaluation, and a prognostic evaluation where the impact of the new gas optics schemes (RRTMGP and RRTMGP-NN) on model climate is determined using year-long IFS simulations.

2 Codes

2.1 ecRad

ecRad is a radiation scheme developed at ECMWF and used operationally in the IFS since 2017 (Hogan & Bozzo, 2018). It is highly configurable with multiple options for gas optics, cloud optics, aerosol optics, and radiative transfer solvers which represent cloud heterogeneity in different ways and support various cloud overlap assumptions.

2.2 RRTMGP

RRTMGP is a recent gas optics scheme with a correlated k-distribution that is based on state-of-the-art spectroscopy. It it part of a freely available toolbox, RTE+RRTMGP, that also includes a radiative transfer solver (RTE). The radiation scheme seeks to balance accuracy, efficiency and flexibility and both the code and data continue to evolve (Pincus et al., 2019). The original RRTMGP k-distributions have a relatively large number of k-terms, also known as g-points: 224 in the shortwave (SW) and 256 in the longwave (LW), corresponding to 16 in each SW and LW band. More recently, reduced k-distributions with half the number of g-points (112/128) have been generated from the full distributions. This was done using the same approach as in the evolution from the RRTM scheme (Mlawer et al., 1997) to its reduced-resolution version designed for GCMs, RRTMG (Iacono et al., 2000), namely by iteratively combining neighboring g-points while

attempting to minimize a cost function that includes fluxes, forcing, and heating rates (R. Pincus and E. Mlawer, personal communication, May 24, 2022). The reduction in g-points was similar in both cases; RRTMG, which is used operationally in the IFS, has 112/140 g-points.

As the number of g-points gives the number of pseudo-monochromatic radiative transfer calculations, it largely determines the cost of the whole radiation scheme. The NNs developed in this paper are therefore based on the new reduced k-distributions ("Reduced-RRTMGP"), and not the k-distributions with higher spectral resolution ("Full-RRTMGP").

2.3 RRTMGP-NN and implementation in ecRad

RRTMGP-NN is a neural network version of RRTMGP described in Ukkonen et al. (2020), available with a refactored version of the RTE solver which has columns as the outermost dimension (in terms of memory location) instead of g-points outermost. The modified radiation scheme is referred to as RTE+RRTMGP-NN and like the original code, it is written in modern Fortran. The simple yet efficient NN kernel is based on BLAS routines for batched inference, exploiting the lack of vertical and horizontal dependencies - this means that gas optics computations are highly parallel. (The code can also be run on GPUs, in this case NVIDIA cuBLAS is used for the matrix multiplications and OpenACC directives are wrapped around the few remaining computations.)

RRTMGP-NN previously loaded models from ASCII files like the Neural-Fortran code (Curcic, 2019) that it is built upon. We have refined the code so that models are loaded from netCDF files, which contain not only the weights and activation functions, but also coefficients used for scaling inputs and outputs, as well as metadata about the training data. These files could in the future be expanded to replace the k-distribution files in their entirety, keeping relevant metadata and the look-up-table coefficients used to compute Planck sources from Planck fraction and temperature.

We now briefly describe the integration of RRTMGP into ecRad. The goal was to avoid larger changes in ecRad. However, since RTE+RRTMGP makes heavy use of Fortran derived types to specify e.g. gas concentrations and optical properties, use of existing RRTMGP interfaces would imply a significant amount of array copying to communicate between ecRad and RRTMGP derived types. Large changes in RRTMGP are not desirable either, because they reduce maintainability of the gas optics code.

With these conflicting goals in mind, a balance was sought with non-intrusive changes in both codes, but prioritizing minimal changes in ecRad. Firstly, the refactored radiation scheme with neural networks, RTE+RRTMGP-NN, was implemented instead of the reference gas optics code to make direct use of existing NN code. This has the advantage that RTE+RRTMGP-NN uses the same dimension order as ecRad with optical properties having g-points innermost and columns as the outermost dimension, removing the need for expensive array transposes (Ukkonen et al., 2020). While the NN fork of RTE+RRTMGP is currently only maintained by one person, the code is very similar to RTE+RRTMGP. The k-distributions are loaded from netCDF files which can be copied over as new ones are made available in the main repository.

The entirety of the RTE+RRTMGP-NN package was then added as an ecRad subdirectory (this was necessary because RRTMGP depends on RTE). The source code of RTE+RRTMGP(-NN) is kept separate: it does not use any of the ecRad modules. Instead, new interfaces were written for RTE+RRTMGP-NN for easy interoperability with ecRad while avoiding having to copy larger arrays. For instance, the new interface for the longwave (*gas_optics_int_ecRad*) replaces the derived type arguments of the original RRTMGP interface, containing optical properties and Planck sources, with explicit shape arrays that are used in ecRad. The computational kernels remain the same, as in reference RRTMGP they do not use derived types. In ecRad, another interface is then used which prepares the RRTMGP-NN gas concentrations (columns outermost) by transposing the ecRad gases (columns innermost) and calls *gas_optics_int_ecRad* (longwave) and *gas_optics_ext_ecRad* (shortwave). ecRad has similar interfaces to the RRTMG and ecCKD (Hogan & Matricardi, 2022) gas optics schemes. The overhead from transposing the gases and thermodynamic arrays is very small.

3 Machine learning

In training NNs to emulate RRTMGP, we use a similar methodology as in Ukkonen et al. (2020), where detailed offline evaluation against line-by-line computations suggested a similar level of accuracy in overall fluxes and heating rates as the original scheme, despite using fairly simple NN models with two hidden layers and 16-48 neurons in each hidden layer. The choice of outputs, loss function, model optimization, and NN complexity are changed slightly as described in the next sections.

3.1 Data

We use similar training data as in Ukkonen et al. (2020), in which a diverse and extensive data set was prepared from several sources, including atmospheric profiles used in previous radiation studies, as well as data from future climate experiments and a reanalysis. These initial data sets were synthetically supplemented, or extended, by varying greenhouse gas concentrations both manually and by using Hypercube sampling. The data in this study differs from Ukkonen et al. (2020) principally in two ways. Firstly, data provided by the Radiative Forcing Model Intercomparison Project (RFMIP, Pincus et al., 2016), comprising of 100 profiles and 18 perturbation experiments, now serves as an independent validation dataset used for early-stopping (section 3.3) instead of training. These profiles were designed to assess global mean clear-sky errors in instantaneous radiative forcing and should be well-suited as an out-of-sample test for our purposes. Secondly, a different dataset based on the CAMS reanalysis (Inness et al., 2019) is used. The new CAMS data uses the same approach as the IFS and the Correlated k-distribution Model Intercomparison Project (CKDMIP, Hogan & Matricardi, 2020), where only nine gases are considered, but the radiative forcing of many minor greenhouse gases is represented by artificially increasing the concentration of CFC-11 (?, ?). The height dependence of these gases is represented, and other RRTMGP gases are set to zero. (Neither of these generally applies to the other subsets of the training data, where all minor RRTMGP gases are included, but as scalar concentrations).

The reanalysis profiles are designed to encompass the variability in present-day atmospheric conditions, with the following steps taken to increase variance and capture extremes. Starting from an initial pool of roughly 164 000 profiles spanning global reanalysis data from 2008 and 2017 and interpolated to a 320 km resolution equal-area grid Ukkonen (2022), 1000 profiles were drawn. Of these, 17 were selected to contain the minimum and maximum of temperature, humidity and ozone at different pressure levels (a total of 9 variables) in the whole dataset, similarly to Hogan and Matricardi (2020). Another 486 profiles were selected by constructing $k = 81$ k-means clusters which are clustered in the 9 dimensions represented by the variables in the previous step. From each cluster, which the k-means algorithm ensures are as different to other clusters as possible, 6 random profiles were selected. The remaining roughly 500 profiles were randomly drawn from the entire dataset minus ones already chosen. Vertical profiles selected by the minimum-maximum, semi-random and random method are depicted in Fig. 1.

The 1000 CAMS profiles were then expanded into 42 experiments or scenarios where CH_4, N_2O, CFC11-eq and CFC12 are varied similarly to Hogan and Matricardi (2020). The $1000 \times 42 \times 60$ (layers) ≈ 2.5 million samples make up roughly 47% of the 5.42 million training samples in total. The remaining parts comprise i) end-of-century CMIP6 data corresponding to a high-emissions scenario, ii) profiles from the "mean-maximum-minimum" CKDMIP dataset, and iii) 42 profiles used for tuning RRTMGP; all of which were expanded into up to hundreds of experiments as described in Ukkonen et al. (2020).

3.2 Choice of inputs and outputs

Our RRTMGP emulator predicts layer-wise optical properties from an input vector which contains gas mixing ratios, temperature, and log-pressure. The NNs take as input all the RRTMGP gases and output all g-points, which results in better computational intensity and efficiency than computing one band at a time, and the contributions from minor gases one gas at a time, as is done in the look-up-table (LUT) kernels in RRTMGP (Ukkonen et al., 2020). In the shortwave (SW), the NN outputs are absorption and Rayleigh cross-sections, while the longwave (LW) predictands are absorption cross-section and Planck fraction. Here, cross-sections refers to optical depth divided by the number of dry air molecules in a layer N. This allows generalization to arbitrary vertical grids, since optical depths are obtained in a separate step by multiplying the cross-sections with N. Meanwhile, *Planck fraction* is the fraction of a band's total Planck function that is associated with each LW g-point, obtained by 3D interpolation in the original code. Like in RRTMGP, this is multiplied with the band-integrated Planck function at a level or layer (interpolated from a LUT using the temperature of that level/layer) to get the Planck function for each LW g-point. This retains a small LUT interpolation, but simplifies the NN model by requiring only ng outputs, instead of $3{\times}ng$ to directly predict the Planck functions used in reference RRTMGP, or $2 \times ng$ to get the Planck functions in RRTMGP-NN. (The original code has one Planck variable for each layer and two for each layer interface, the upward and downward emission, whereas RTE+RRTMGP-NN has one for each layer and layer interface. ecRad only uses one Planck function, defined at layer interfaces). Reducing the number of NN outputs can decrease model complexity and runtimes, since most of the floating point operations occur in the final NN layer given $N_{gpt} > N_{neurons} > N_{inputs}$. However, in this work a single LW model is used which predicts both absorption cross-sections and Planck fractions as one big vec-

tor. This may not be the fastest approach but has the benefit of easing the optimization procedure described in the next section.

In addition to predicting cross-sections instead of optical depths, to obtain good results with less complex models it is useful to pre-process both inputs and outputs. Specifically, square root transformations are used for all outputs and some inputs to make their distributions more uniform, and afterwards the inputs are scaled to the 0-1 range and outputs are scaled to have roughly zero mean and unit variance by using a variant of standardization that preserves correlations between different outputs (Ukkonen et al., 2020).

3.3 Should we optimize for fluxes when predicting optical properties?

Training gas optics ML models presents a tuning challenge, as the variables we ultimately care about are not optical properties but radiative fluxes and heating rates - outputs from the solver. We previously found it relatively easy to develop gas optics NNs which upon implementation in the radiation code result in low mean errors in fluxes and heating rates, but difficult to obtain accurate radiative forcings at the top-of-atmosphere or surface with respect to changes in the concentration of individual gases, especially minor ones (Ukkonen et al., 2020). The problem is likely to stem from predicting aggregated optical properties, instead of computing the contribution from minor gases separately (as is usually done in k-distributions), which is more efficient but leads to major gases dominating the loss function. Mostly accurate radiative forcings for CKDMIP gases were ultimately obtained via a time-consuming, iterative process where new models were continuously trained, evaluated, and the training data expanded. In this work we have attempted to automate the optimization with regards to fluxes, heating rates and forcings to at least some extent by adding two new techniques to the training methodology.

Firstly, errors in fluxes and heating rates were monitored during training. While these metrics can not be easily be used for optimizing the NN weights, they can be used as a criteria to know when to stop training (*early-stopping*), or to optimize NN hyperparameters. Therefore, a Python training program was written where at the end of every epoch, the NN models are saved to a file, and the Fortran radiation program is called with the new model, passing the location as a command-line-argument. The Fortran program runs RTE+RRTMGP-NN on a validation dataset, and writes some error metrics

to standard output, which are finally read by the training program. For validation we used the RFMIP dataset consisting of 100 profiles and 18 different perturbation experiments, which allowed computing radiative forcing errors with respect to CH_4, N_2O, and total forcing errors with respect to all RRTMGP gases. In addition, a benchmark line-by-line solution was available for this data, meaning that the total (parameterization) error can be determined instead of only emulation (NN) error. Our goal was to develop NNs that have a similar level of accuracy as RRTMGP; that is, emulation errors should be smaller than parameterization errors. The error metrics were normalized by the RRTMGP values, so that a value of one indicates the same level of performance as RRTMGP, and larger values indicate worse performance. An overall "radiation error" was computed by taking the RMS value of a total of 8 metrics which differ slightly for the longwave and shortwave (Table 1). This overall metric was used in the early stopping criteria and the model weights from the best epoch (a minimum in the metric) were saved.

Metric	Longwave	Shortwave
MAE Heating rate	X	X
MAE Heating rate (present-day)	X	X
MAE Heating rate (preindustrial)		X
MAE Heating rate ("future-all")		X
Bias surface downwelling flux		X
Bias TOA upwelling flux	X	
Bias TOA IRF (present-day - preindustrial)	X	
Bias TOA IRF (future - present-day)	X	
Bias TOA IRF (future - preindustrial)		X
Bias surface IRF (future - preindustrial)	X	X
Bias surface IRF CH4 (present-day - preindustrial)	X	X
Bias surface IRF N2O (present-day - preindustrial)	X	

Table 1: Metrics that comprise the overall "radiation error". TOA = top of atmosphere, IRF = instantaneous radiative forcing, "future-all" = a radiative forcing experiment with perturbed atmospheric conditions in addition to greenhouse gas concentrations.

Secondly, a custom loss function was devised to minimize the error in the difference in y associated with different perturbation experiments, in addition to mean-squared-error of y, where y are the scaled NN outputs. The new loss function indirectly measures radiative forcing errors (albeit weakly due to a non-linear dependence between spectral optical properties and broadband fluxes) and has the form:

$$loss = \alpha \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + (1-\alpha) \sum_{\substack{i=1 \\ i \text{ odd}}}^{N} \Big((y_{i+1} - y_i) - (\hat{y}_{i+1} - \hat{y}_i) \Big)^2,$$

where y and $\hat{y}$ are the target and NN output vectors, respectively, and α is a coefficient that was set to 0.6 for LW and 0.2 for SW after some testing. The second term measures the error in the difference in y between different perturbation experiments if the data is organized so that adjacent samples (of a total N training samples) correspond to different experiments but the same columns and vertical layers, which was achieved by transposing the data so that the experiment dimension is innermost. In addition, the experiments should be designed so that every odd element and its neighbour relate to the goal, which was minimizing the TOA and surface forcing errors of individual gases. Therefore, RFMIP-style experiments such as present-day versus future concentrations of all greenhouse gases, or 8X CO_2 versus preindustrial CO_2, should be avoided, as they can easily dominate the error compared to varying the concentration of minor greenhouse gases (which was the challenge to begin with). This requirement was only partially fulfilled since we wanted to make use of existing data. Though rather convoluted, and requiring bespoke data, the approach does reduce forcing errors (Figure 2).

In the end, there was still a substantial random element in results obtained, and several models were trained before settling on the final models (based on errors with respect to training data, and not the independent offline evaluation, which was only performed once). To obtain a satisfactory LW model the early-stopping criteria was loosened to 70 consecutive epochs of no improvement. (Depending on the desired level of accuracy for the radiative forcings, it may be possible to obtain adequate results without a hybrid loss function by simply training for a very large number of epochs, at the risk of overfitting if the training data is not very extensive.) In addition, increasing the number of hidden neurons compared to Ukkonen et al. (2020) seemed to improve results slightly. The final LW model has 64 neurons in two hidden layers, and the SW models have 32

neurons in two hidden layers. All models use the "softsign" activation function and were trained using the Adam optimizer, a batch size of 2048 and learning rate of 0.01.

Future studies could explore directly minimizing flux and forcing errors when training NN-based gas optics models. Doing this via gradient descent optimization would require differentiating the radiative transfer solver to obtain the derivative of fluxes with respect to changes in optical properties (and NN weights), which should be possible using automatic differentiation tools like the Python library JAX (Bradbury et al., 2018) if the radiative transfer code to be re-written in the supported language or API (JAX, for instance, has an API based on NumPy).

4 Results

In this section we evaluate the accuracy and speed of ecRad with different gas optics schemes (RRTMGP, RRTMGP-NN, and the older RRTMG scheme) in both an offline and online setting. The results were obtained using an optimized development version of ecRad which refactors the Tripleclouds (Shonk & Hogan, 2008) and SPARTACUS (Hogan et al., 2016) solvers for better efficiency and includes the new RRTMGP(-NN) gas optics. Another optimization is that reflectances and transmittances are computed in the same numerical precision as the rest of the model (in the current operational version of ecRad, these two-stream computations are always performed in double precision), which improves the single-precision performance of all solvers in ecRad. The optimizations, which are described in a forthcoming paper, have a negligible impact on fluxes and heating rates while making Tripleclouds significantly cheaper, and thus increase the share of the gas optics in the total runtime of ecRad.

The prognostic evaluation and offline timings were obtained with an ecRad configuration close to operational IFS Cycle 47r3, which is similar to the 46r1 settings given in Table 2 of Hogan and Bozzo (2018), except for replacement of the pure exponential cloud overlap assumption with "exponential-random" whereby an vertically contiguous cloud layers are partially correlated but cloud layers separated by clear sky are randomly overlapped. Furthermore, we replace the McICA solver with Tripleclouds due to the latter being noise-free and a likely candidate for operational use in a near-future cycle.

It should be noted that the RRTMGP results were not produced using the original RTE+RRTMGP package, which uses two Planck source functions for half-levels which

are then combined into one, and the LW results (Figures S1-S4 in Supporting Information S1) could be very slightly impacted by the simpler computation of Planck source in RRTMGP-NN. For simplicity the RRTMGP-NN code configured with look-up-tables and not NNs is hereafter referred to as RRTMGP.

4.1 Speed-up

The runtime of ecRad with different gas optics schemes was evaluated offline using 10,000 input profiles that were saved from a benchmark forecast run in the IFS, and a block size of 8 columns (equal to the block size "NPROMA" in the IFS). Figure 3 shows timing results obtained on a single node of the new ECMWF AMD-based supercomputer in Bologna, to which the migration of ECMWF's operational forecast is expected later in 2022.

With Reduced-RRTMGP, the runtime of ecRad is increased slightly due to the new gas optics (shown in light blue) being more expensive than the older RRTMG scheme, which is faster by a factor of 1.67 despite the similar spectral resolution. The poor efficiency of RRTMGP is explained by short inner loops in the LUT code, where inner loops are over g-points in a band (only 12-16 for the smaller k-distributions), leading to poor vectorization. However, the NN version of Reduced-RRTMGP is 2.36 times faster than the LUT-based code, and also faster than the old RRMTG scheme, leading to a total speedup of the radiation code by a factor of 1.13 compared to operationally used RRTMG.

4.2 Offline evaluation

Independent validation of the NN gas optics models was carried out by using data and tools from CKDMIP (Hogan & Matricardi, 2020). The data are from the 'Evaluation-1' dataset, which was not used for training. The accuracy of the new RRTMGP-NN SW model, relative to a line-by-line benchmark, is first shown in Figure 4 for the present-day scenario. The NN has almost identical accuracy as the Reduced-RRTMGP scheme it was trained on (Fig. 5), particularly in terms of heating rates, which in both cases have an RMSE of only 0.056-0.057 K d^{-1} below 4 hPa. Surprisingly, the upwelling flux bias and root-mean-square errors are actually smaller when using NNs. Results using the other CKDMIP concentration scenarios (Glacial Maximum, Preindustrial, and Future) are not

presented here but are very similar, with the NN gas optics resulting in better upwelling fluxes but similar heating rates.

In general, the close emulation of RRTMGP was already demonstrated in Ukkonen et al. (2020) and so the remaining results are not discussed here, but the LW results for present-day and future scenarios are provided in Supporting Information S1 (Figures S1-S4). The most significant difference to the earlier paper is that the top-of-atmosphere and surface forcings with respect to N_2O, CFC11 and CFC12 have been improved with the help of a hybrid loss function and are now excellent (Fig. 6). We note that the new Reduced-RRTMGP k-distributions with 112 (SW) and 128 (LW) g-points seem to trade only a little accuracy for a lot of speed compared to the original k-distributions with almost double the g-points (`https://confluence.ecmwf.int/display/CKDMIP`), with the exception of LW heating rates in the mesosphere which are considerably worse for Reduced-RRTMGP.

4.3 Prognostic evaluation

We now describe results from a prognostic evaluation of RRTMGP-NN and RRTMGP using 1-year free-running simulations with the IFS model. The model simulations consisted of four atmosphere-ocean coupled simulations 13 months long initialized on 1 August of the years 2000, 2001, 2002 and 2003. After a 1-month spin-up for each simulation, the remaining 12 months were averaged over each simulation. This configuration is very similar to that used in section 5 of Hogan and Bozzo (2018) to evaluate the impact of changes to the radiation scheme; the simulations are long enough to capture fast atmospheric and land-surface processes that respond to changes in the treatment of radiative transfer, but short enough that the response is not significantly affected by the longer-term changes to ocean circulation. The one-year forecast length also matches the longest operational forecast length used in ECMWF's seasonal forecasts. The model configuration was as in operational IFS model cycle 47r3 except for the use of the Tripleclouds rather than McICA solver. The horizontal resolution was $T_{Co}199$ (around 60 km) and 137 vertical levels were used. The radiation scheme was called every hour.

The impact of different gas optics schemes on annual-mean temperature from the surface to the lower mesosphere is shown in Figure 7. Because RRTMGP is not, to our knowledge, routinely tested in single precision, both single and double precision runs were

performed with both Reduced-RRTMGP and Full-RRTMGP. The NN version of Reduced-RRTMGP was only tested in single precision (internally, RRTMGP-NN always uses SP, as higher numerical precision does not benefit NNs). In general, larger differences between the runs are only seen in the mesosphere and upper stratosphere, which are very sensitive to heating-rate differences. Comparison against a reference dataset based on the Microwave Limb Sounder (MLS) instrument above 20 hPa, and ERA5 reanalysis data below this level is depicted in Fig. 7 (b) and shows a 5-K warm bias in the upper stratosphere for RRTMG, and larger in the mesosphere. The same bias was reported by Hogan et al. (2017), which they explained by the use of the older 'Kurucz' solar spectrum in the RRTMG version used in ecRad. RRTMGP uses a more recent solar spectrum with less ultraviolet radiation, resulting in closer agreement with MLS. In the mesosphere, RRTMGP-NN is clearly closer to the Reduced-RRTMGP scheme it is emulating than Reduced-RRTMGP is to Full-RRTMGP (Fig. 7 (c)). This strongly suggests that the emulation errors are small enough to be acceptable.

A height-latitude cross section of temperature likewise shows larger differences between the old RRTMG scheme and RRTMGP than between different RRTMGP configurations and the NN version (Fig. 8). A strong warm bias in the stratosphere is evident for RRTMG but less so for any version of RRTMGP, although the RRTMGP(-NN) runs do show a weaker upper-stratospheric warm bias over high latitudes and a substantial cold bias in the tropical stratosphere.

Finally, Fig. 9 compares annual/zonal means of 2-m temperature, TOA net LW and SW fluxes, and downwelling SW flux between simulations using different gas optics configurations. In general the differences relative to Full-RRTMGP are statistically insignificant as the means fall within the error bars computed from the 4-year sample. The only clear exception is RRTMG, which, for instance, at lower latitudes has significantly larger surface downwelling SW flux and smaller LW flux than Full-RRTMGP. These findings are consistent in sign and approximate magnitude with the evaluation of RRTMG against line-by-line calculations by Hogan and Matricardi (2020), although it should be stressed that these differences of 0–2 W m^{-2} are still very modest.

5 Conclusions

In this paper we have evaluated RRTMGP-NN, a neural network version of the RRTMGP gas optics scheme, integrated into ECMWF's radiation scheme ecRad, by using both offline calculations and by performing four 1-year simulations with the free-running IFS model. Emulating only the gas optics component, instead of the full radiation scheme like in previous work, results in much better accuracy at the cost of a (much) smaller speed-up. The NN models were trained on diverse datasets which cover a wide range of gas concentrations and atmospheric conditions (including pre-industrial, present-day and future conditions), and a hybrid loss function was utilized to reduce radiative forcing errors. In offline timings obtained on ECMWF's AMD-based supercomputer, using RRTMGP-NN makes ecRad roughly a third faster compared to RRTMGP.

The results from the online evaluation are highly encouraging as they show very similar model climates for RRTMGP and RRTMGP-NN. Global/annual mean temperature profiles for the reduced-spectral-resolution RRTMGP and its NN emulator are in closer agreement than different versions of RRTMGP are to each other, and all of these schemes, including RRTMGP-NN, substantially reduce stratosphere and mesosphere temperature biases relative to the older RRTMG scheme. In single-level fields, the differences between RRTMGP-NN and RRTMGP are obscured by natural variability. This, combined with virtually identical results for the two schemes in the offline evaluation, demonstrates that our RRTMGP-NN models are generally suitable for operational weather and climate models. The evaluations were based on a setup with a reduced set of LW gases (using artificially increased CFC-11 concentrations to represent further gases); for applications where the individual radiative forcings of other RRTMGP minor greenhouse gases are of importance, such metrics can be evaluated offline and new NN models trained if needed. The data and tools are freely available and should be useful for future work on radiation emulators.

6 Open Research

RTE+RRTMGP-NN is available on Github (`https://github.com/peterukk/rte-rrtmgp-nn`); the Fortran programs and Python scripts used are found in the /examples/rrtmgp-nn-training subdirectory. The data and an archived version of RTE+RRTMGP-NN have been uploaded to Zenodo (`https://doi.org/10.5281/zenodo.6576680`).

References

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., ... Zhang, Q. (2018). *JAX: composable transformations of Python+NumPy programs.* Retrieved from `http://github.com/google/jax`

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375.

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. doi: 10.1029/2018gl078510

Chevallier, F., Chéruy, F., Scott, N., & Chédin, A. (1998). A neural network approach for a fast and accurate computation of a longwave radiative budget. *Journal of applied meteorology*, *37*(11), 1385–1397.

Cotronei, A., & Slawig, T. (2020). Single-precision arithmetic in echam radiation reduces runtime and energy consumption. *Geoscientific Model Development*, *13*(6), 2783–2804.

Curcic, M. (2019). A parallel fortran framework for neural networks and deep learning. In *Acm sigplan fortran forum* (Vol. 38, pp. 4–21).

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. doi: 10.1029/2018gl078202

Goody, R., West, R., Chen, L., & Crisp, D. (1989). The correlated-k method for radiation calculations in nonhomogeneous atmospheres. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *42*(6), 539–550.

Hogan, R. J., Ahlgrimm, M., Balsamo, G., Beljaars, A., Berrisford, P., Bozzo, A., ... Wedi, N. (2017). *Radiation in numerical weather prediction* (Tech. Memo. No. 816). ECMWF.

Hogan, R. J., & Bozzo, A. (2018). A flexible and efficient radiation scheme for the ecmwf model. *Journal of Advances in Modeling Earth Systems*, *10*(8), 1990–2008. doi: https://doi.org/10.1029/2018MS001364

Hogan, R. J., & Matricardi, M. (2020). Evaluating and improving the treatment of gases in radiation schemes: the correlated k-distribution model intercomparison project (ckdmip). *Geoscientific Model Development Discussions*, *2020*,

1–29. doi: 10.5194/gmd-2020-99

Hogan, R. J., & Matricardi, M. (2022). A tool for generating fast k-distribution gas-optics models for weather and climate applications. *Earth and Space Science Open Archive*, 24. Retrieved from `https://doi.org/10.1002/essoar.10510456.1` doi: 10.1002/essoar.10510456.1

Hogan, R. J., Schäfer, S. A., Klinger, C., Chiu, J. C., & Mayer, B. (2016). Representing 3-d cloud radiation effects in two-stream schemes: 2. matrix formulation and broadband evaluation. *Journal of Geophysical Research: Atmospheres*, *121*(14), 8583–8599.

Iacono, M. J., Mlawer, E. J., Clough, S. A., & Morcrette, J.-J. (2000). Impact of an improved longwave radiation model, rrtm, on the energy budget and thermodynamic properties of the ncar community climate model, ccm3. *Journal of Geophysical Research: Atmospheres*, *105*(D11), 14873–14890.

Inness, A., Ades, M., Agusti-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., ... others (2019). The cams reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, *19*(6), 3515–3556.

Krasnopolsky, V., Fox-Rabinovitz, M., Hou, Y., Lord, S., & Belochitski, A. (2010). Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: climate simulations and seasonal predictions. *Monthly Weather Review*, *138*(5), 1822–1842.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2008). Decadal climate simulations using accurate and fast neural network emulation of full, longwave and shortwave, radiation. *Monthly Weather Review*, *136*(10), 3683–3695.

Liu, Y., Caballero, R., & Monteiro, J. M. (2020). Radnet 1.0: Exploring deep learning architectures for longwave radiative transfer. *Geoscientific Model Development*, *13*(9), 4399–4412.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997). Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, *102*(D14), 16663–16682.

Pal, A., Mahajan, S., & Norman, M. R. (2019). Using deep neural networks as cost-effective surrogate models for super-parameterized e3sm radia-

tive transfer. *Geophysical Research Letters*, *46*(11), 6069–6079. doi: https://doi.org/10.1029/2018GL081646

Pincus, R., Forster, P. M., & Stevens, B. (2016). The radiative forcing model intercomparison project (rfmip): experimental protocol for cmip6. *Geoscientific Model Development*, *9*(9), 3447–3460. doi: 10.5194/gmd-9-3447-2016

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3074–3089. doi: https://doi.org/10.1029/2019MS001621

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. doi: 10.1073/pnas.1810286115

Roh, S., & Song, H.-J. (2020). Evaluation of neural network emulations for radiation parameterization in cloud resolving model. *Geophysical Research Letters*, *47*(21), e2020GL089444. doi: https://doi.org/10.1029/2020GL089444

Shonk, J. K., & Hogan, R. J. (2008). Tripleclouds: An efficient method for representing horizontal cloud inhomogeneity in 1d radiation schemes by using three regions at each height. *Journal of Climate*, *21*(11), 2352–2370.

Song, H.-J., & Roh, S. (2021). Improved weather forecasting using neural network emulation for radiation parameterization. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002609. doi: https://doi.org/10.1029/2021MS002609

Ukkonen, P. (2022). Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, e2021MS002875. doi: 10.1029/2021MS002875

Ukkonen, P., Pincus, R., Hogan, R. J., Nielsen, K. P., & Kaas, E. (2020). Accelerating radiation computations for dynamical models with targeted machine learning and code optimization. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002226. doi: https://doi.org/10.1029/2020MS002226

Veerman, M. A., Pincus, R., Stoffer, R., Van Leeuwen, C. M., Podareanu, D., & Van Heerwaarden, C. C. (2021). Predicting atmospheric optical properties for radiative transfer computations using neural networks. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200095.

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940.

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. doi: https://doi.org/10.1029/2020GL091363

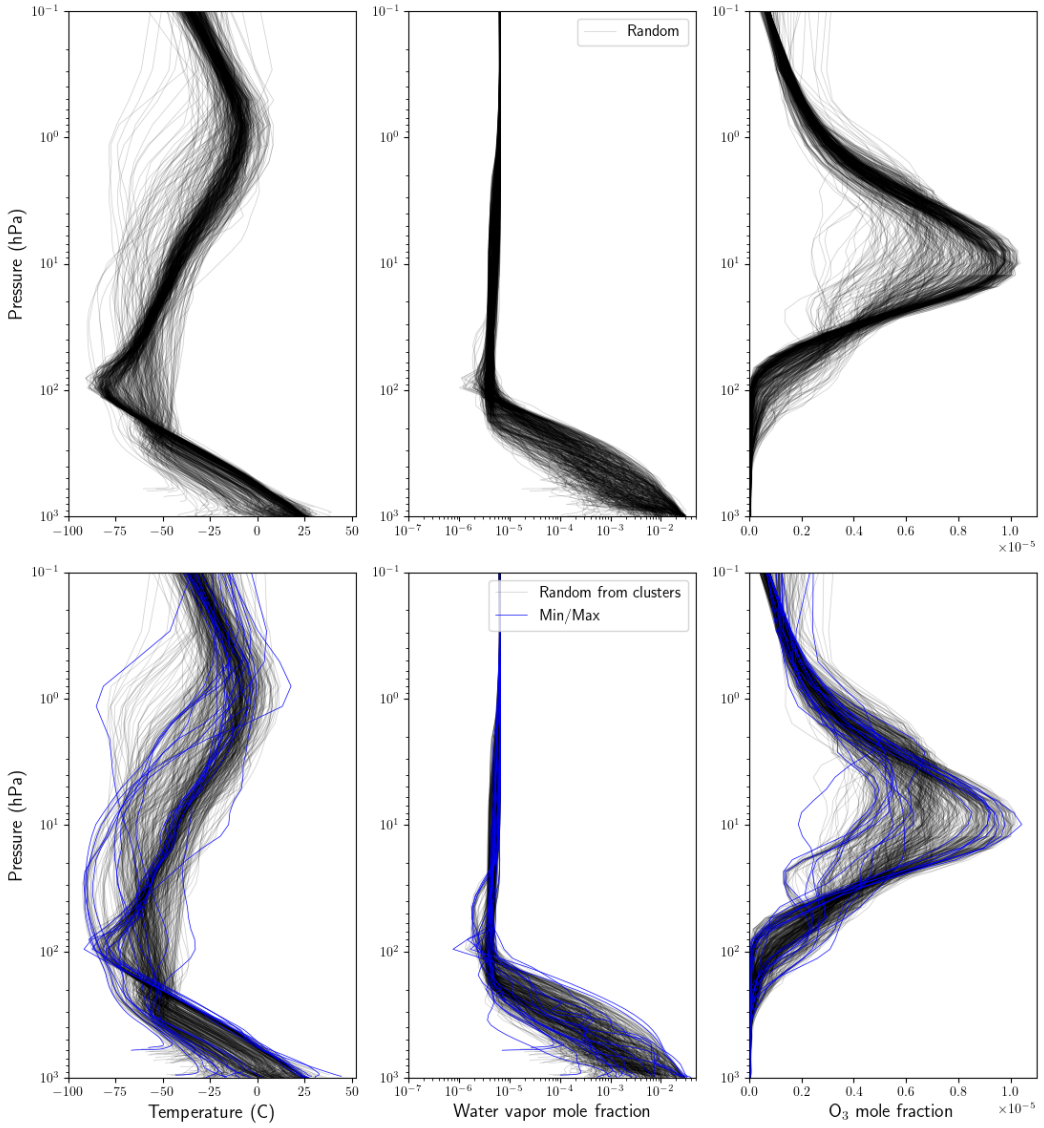

Figure 1: Vertical profiles of temperature, water vapor and ozone selected from the CAMS data as described in Sect. 3.1. The top panel shows 486 random profiles (black), and the bottom panel shows 486 profiles drawn from k-means clusters (black) and 17 that were selected to sample minimum and maximum values (blue).

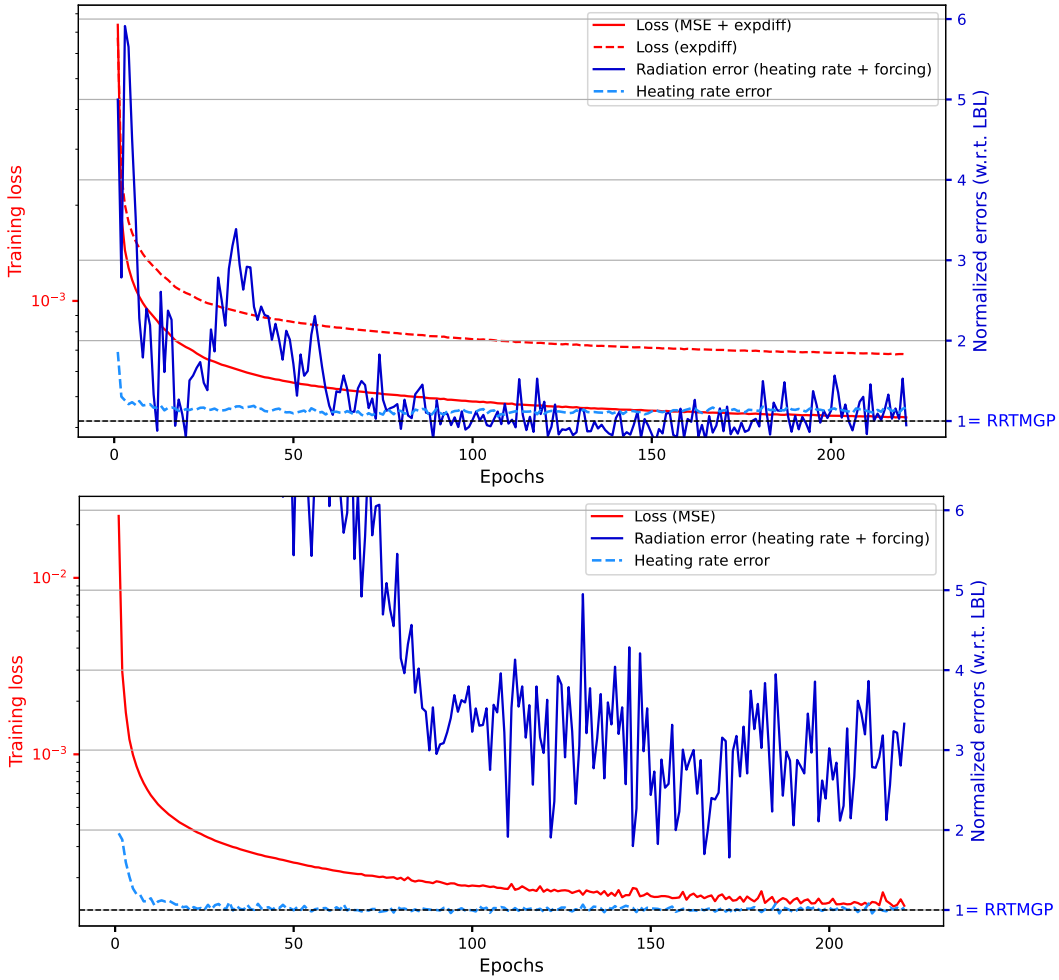

Figure 2: Monitoring of heating rate error (dashed cyan line, given by the mean of the heating rate metrics in Table 1), and the total radiation error (solid blue line, given by the RMS of the metrics listed in Table 1) when training the final LW gas optics model using a hybrid loss function and early stopping (top), and training for the same number of epochs with a regular loss function (bottom). The monitored errors (right axis) are computed using the RFMIP data with respect to line-by-line results and normalized by the RRTMGP value. Also shown are the training losses (solid and dashed red lines). The larger radiation error when not using the hybrid loss function (b) was largely due to a single metric, the surface radiative forcing of N2O (not shown).

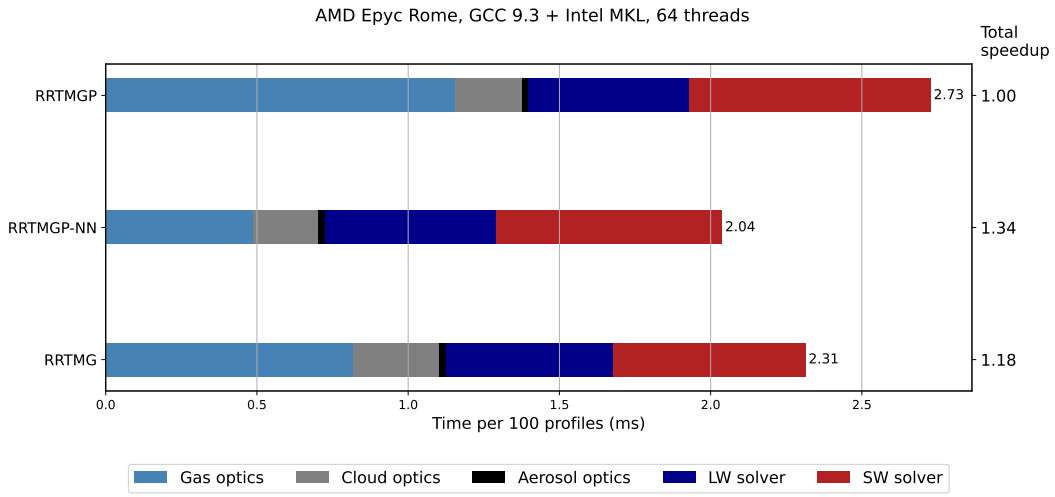

Figure 3: Runtime of ecRad in single precision per 100 atmospheric profiles, broken down by component. Three gas optics schemes are compared: RRTMGP scheme with reduced spectral resolution (112 SW and 128 LW g-points), its neural network version (RRTMGP-NN), and the older RRTMG scheme with 112 (SW) and 140 (LW) g-points. The runs were repeated 10 times and the values for each component were computed by taking the average of the per-thread values reported by the General Purpose Timing Library. CPU: 64-core AMD Epyc Rome. Software platform: GNU Fortran compiler version 9.3 and Intel MKL library 19.0.5 (used for general matrix-matrix multiplication (GEMM) in RRTMGP-NN)

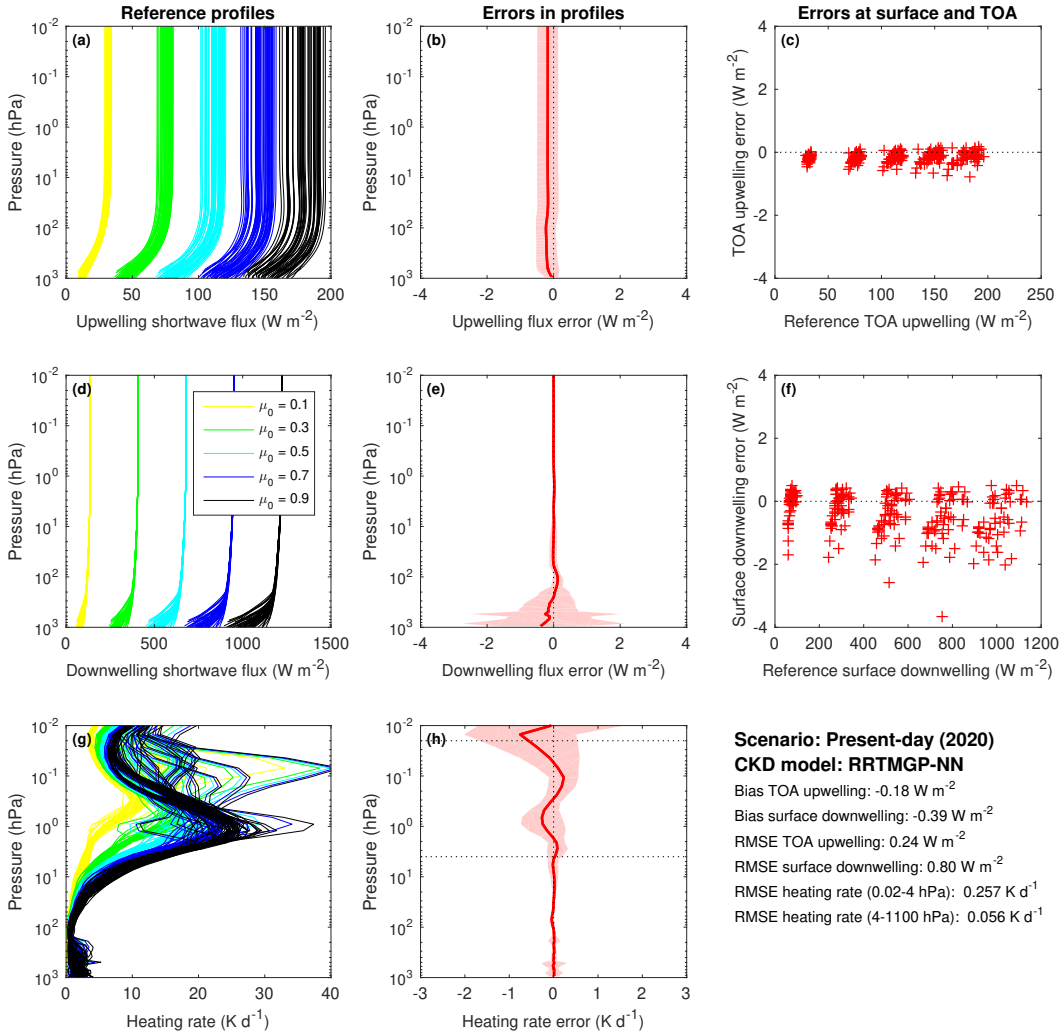

Figure 4: Evaluation of Reduced-RRTMGP-NN shortwave fluxes and heating rates using the 50 independent profiles of the CKDMIP Evaluation-1 dataset with present-day concentrations of greenhouse gases. The left column (a, d, g) shows the reference profiles of upwelling flux, downwelling flux and heating rate from LBL calculations with five different values of the cosine of the solar zenith angle, μ_0 (0.1, 0.3, 0.5, 0.7 and 0.9). The middle column (b, e, h) shows the corresponding biases (solid lines) and 95th percentile of errors (shaded area) using all 250 data points. The right column (c, f) depicts instantaneous errors in upwelling TOA and downwelling surface fluxes with the clusters corresponding to the different solar zenith angles.

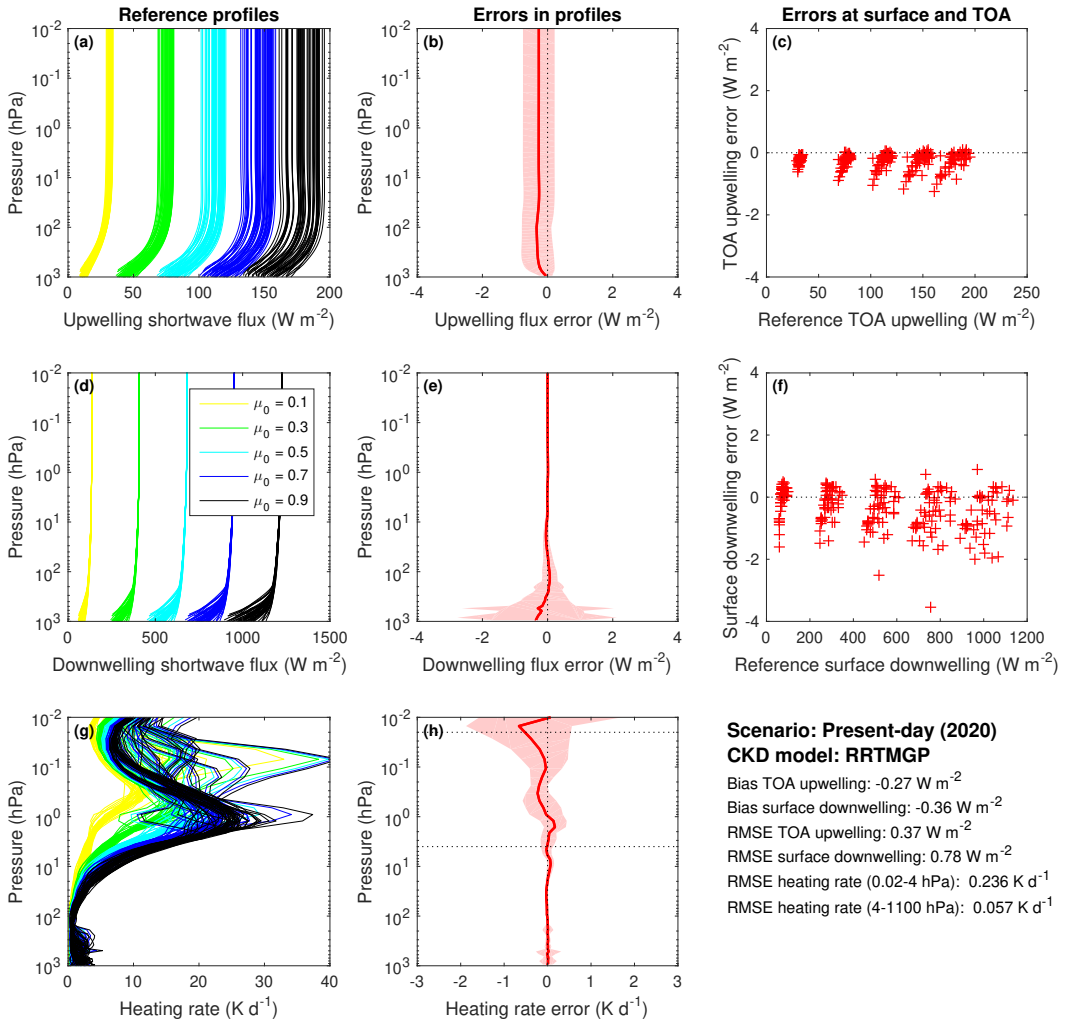

Figure 5: As in Fig. 4 but for Reduced-RRTMGP.

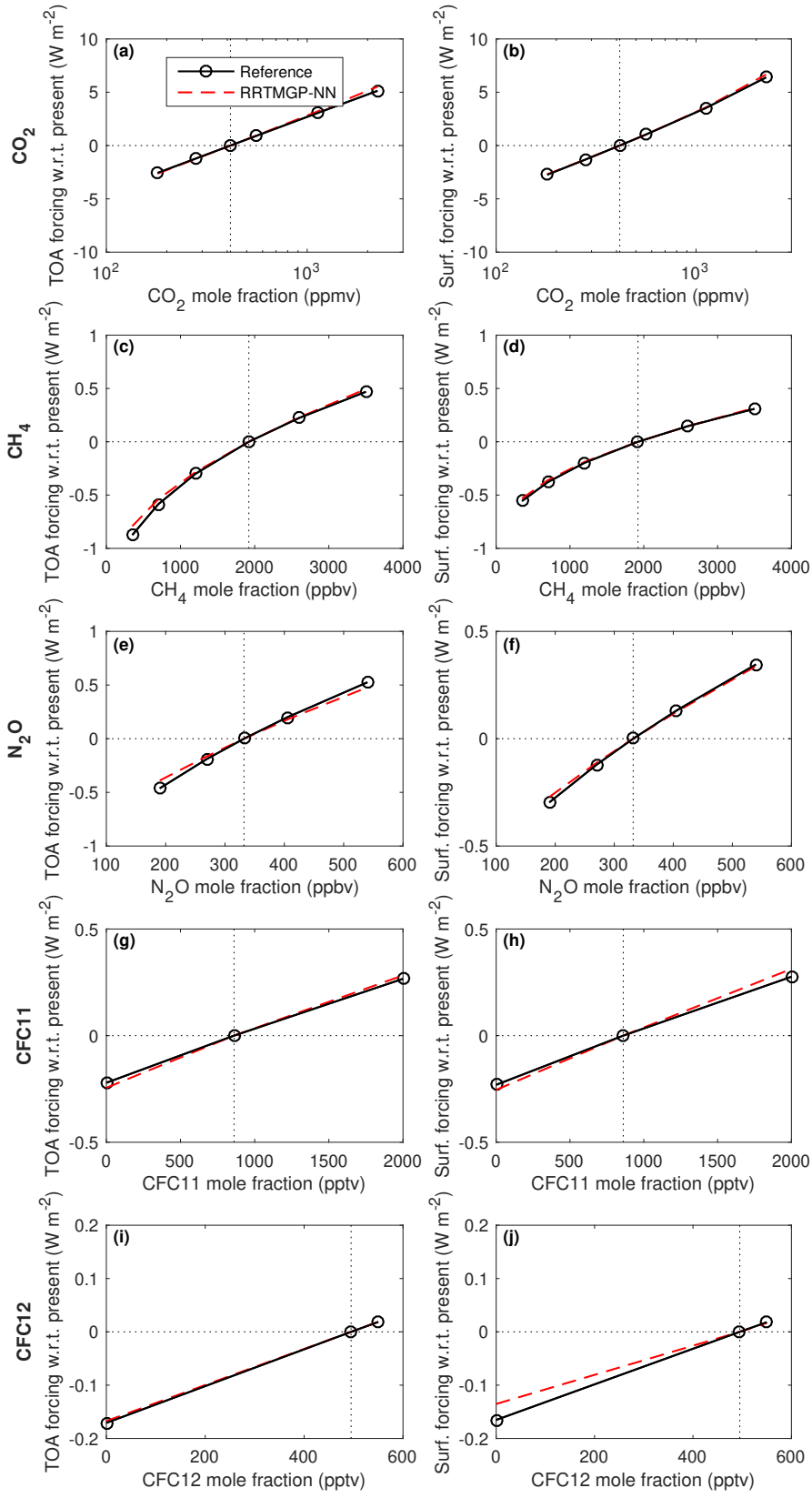

Figure 6: Comparison of RRTMGP-NN and reference LBL calculations of instantaneous longwave clear-sky radiative forcing at top of atmosphere (left column) and surface (right column) when perturbing different greenhouse gases (rows), averaged over the 50 profiles in the CKDMIP Evaluation 1 dataset.

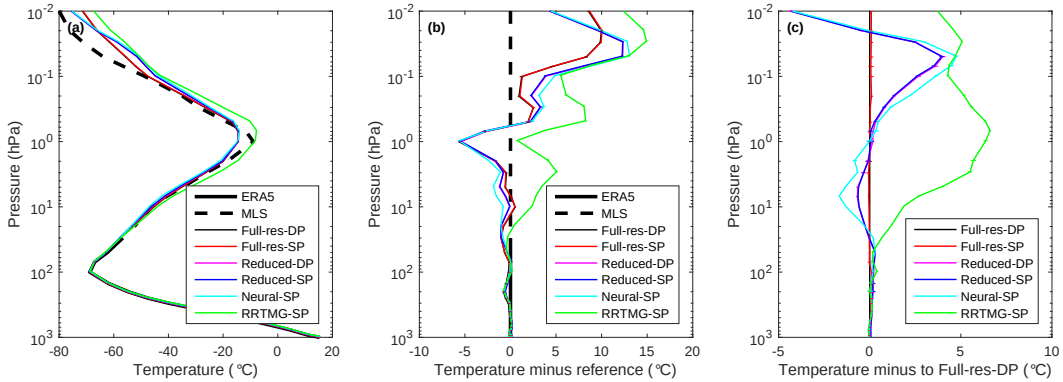

Figure 7: Evaluation of global/annual-mean temperature profiles from free-running simulations by the IFS. (a) Mean temperature, (b) difference against a reference dataset consisting of the MLS climatology above the 20-hPa height level and ERA5 below this level, and (c) difference against a simulation using full-resolution RRTMGP in double precision. The small horizontal bars give the 95% confidence interval as computed from differences between different years. As indicated in the legend, the simulations were performed in double or single precision (DP or SP) using the full- or reduced-resolution RRTMGP, the NN emulation of (reduced-)RRTMGP, or the older RRTMG gas optics scheme.

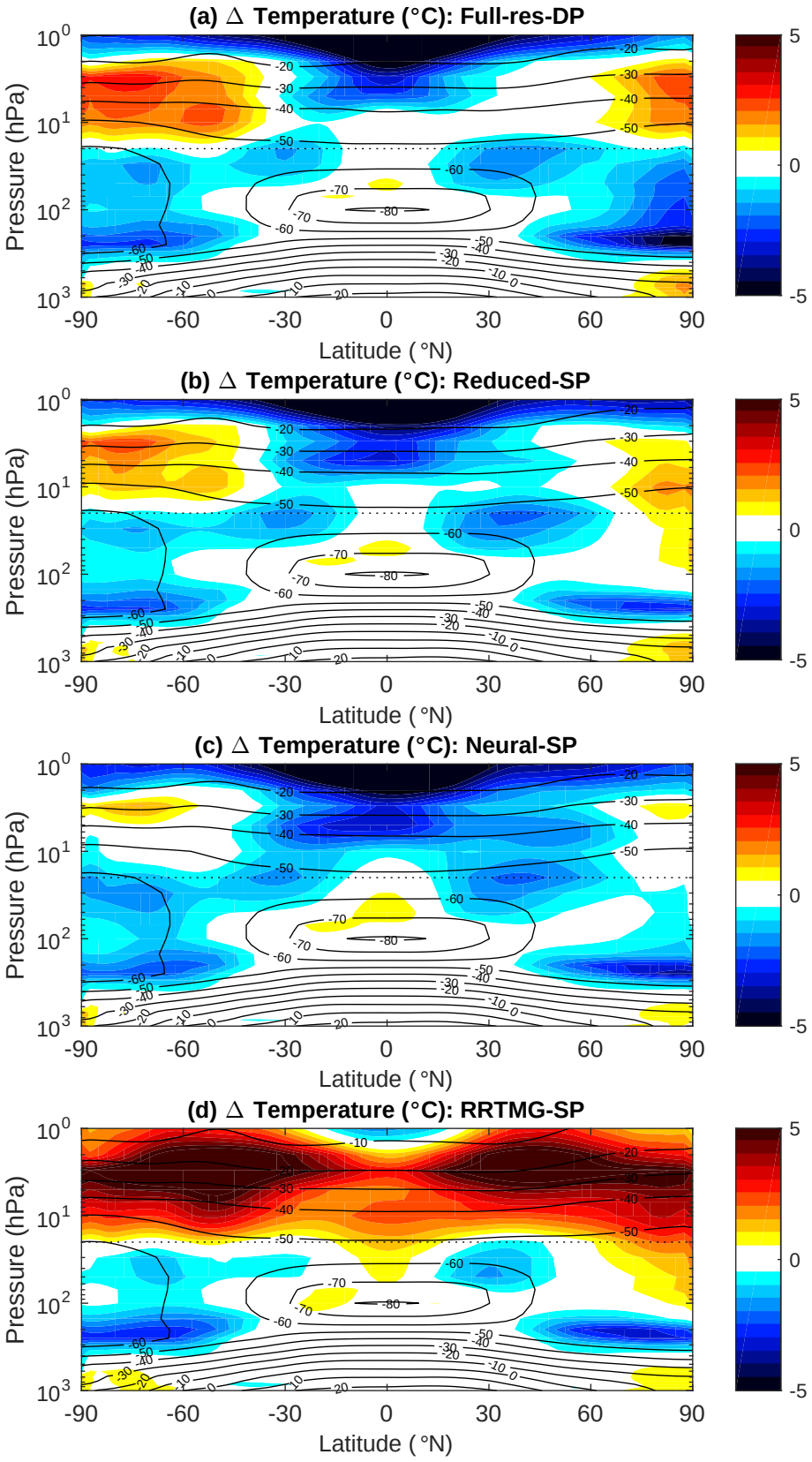

Figure 8: Similar to Fig. 7 but showing the the height-latitude cross section of mean temperature (black contours) and temperature difference (colors) against the reference datasets, and only until 1 hPa.

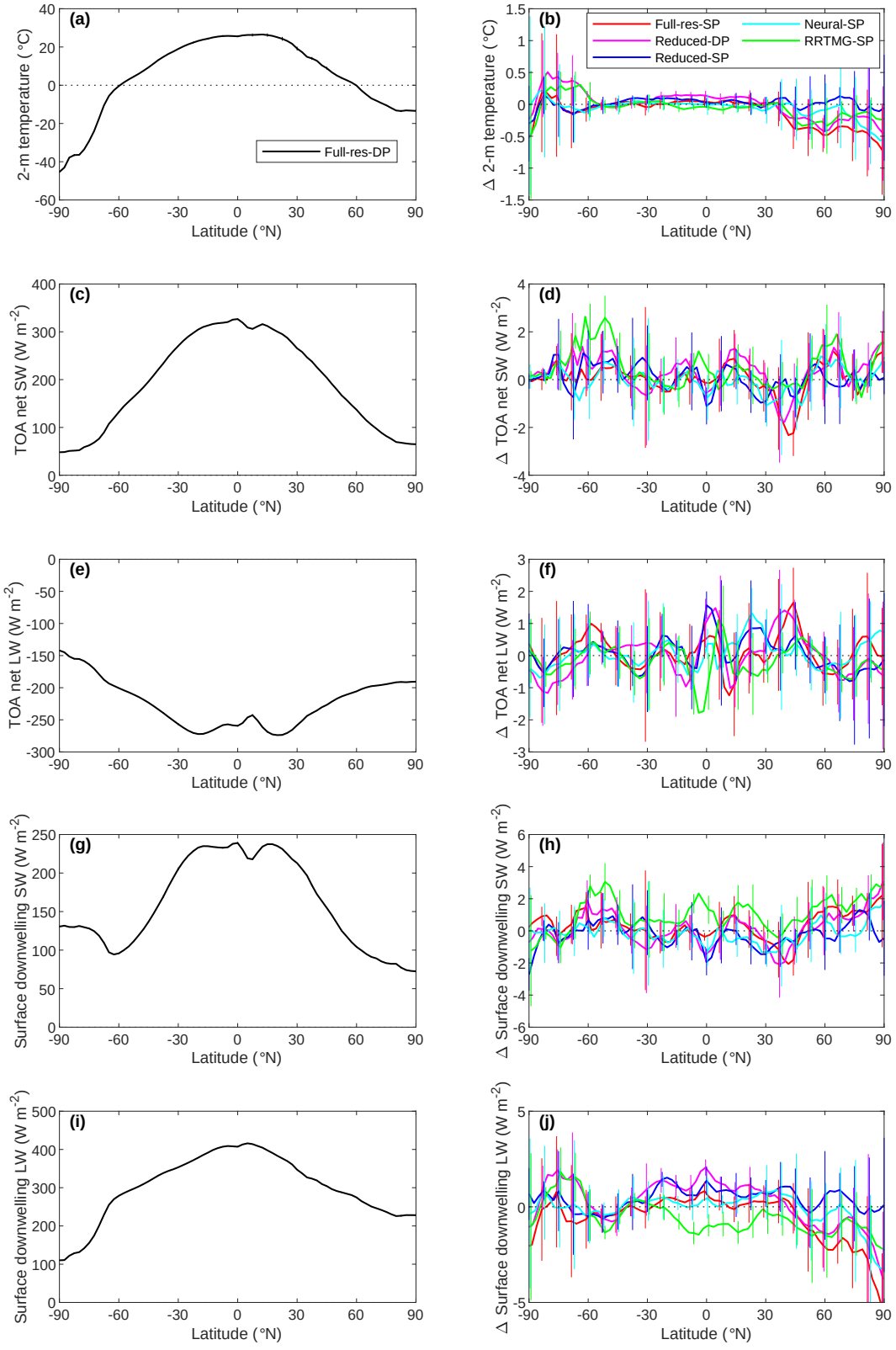

Figure 9: Zonal mean of different single-level quantities using the reference Full-RRTMGP run in double precision (a, c, e, g, i) and differences relative to this run (b, d, f, h, j) with the vertical lines indicating the 95% confidence interval: (a, b) 2-m temperature, (c, d) TOA net SW flux (where net is defined as downwelling minus upwelling), (e, f) TOA net LW flux, (g, h) surface downwelling SW flux, and (i, j) surface downwelling LW flux.