Bachelor Thesis

Improving Physical Consistency in ML-driven Radiative Transfer Parameterization for Climate Model Emulation

Charlotte Lange
1st Supervisor: Dr. Christopher Irrgang
2nd Supervisor: Prof. Gordon Pipa

April 8, 2023

Abstract

In this work, we study the application of physically motivated constraint machine learning (ML) methods to parameterize radiative transfer (RT) in Earth System Models (ESMs) to address the challenge of climate change mitigation and adaptation. We highlight the limitations of current RT parameterizations due to computational limitations and physical bias, and the challenges faced by ML surrogates. We discuss the potential of using theory-guided ML methods, integrating scientific principles of climate physics to emulate RT parameterization and especially the use of constraint methods, which have shown promise in improving the accuracy, efficiency, and generalization capacity of parameterizations in the context of climate models. We highlight the lack of research on ML-driven RT parameterizations using these methods and propose a combination of physically motivated hard and soft constraining methods and Transfer Learning, to improve RT parameterization conducted by numerical and data-driven benchmarks. Our analysis focuses on physical validity and generalization, showing improvement in overall performance in our resulting models compared to existing baselines.

Contents

1 Introduction

Climate change has visible effects on the Earth's natural systems and human societies and requires mitigation and adaptation within thirty years to avoid catastrophic consequences [1].

To avert serious consequences such as extreme weather events, sea level rise, and the loss of biodiversity, urgent action is required. Mitigation efforts, such as reducing greenhouse gas emissions, are essential to slow the rate and magnitude of climate change. At the same time, adaptation measures, such as building resilient infrastructure and enhancing natural systems, must be implemented to cope with the changes that are already happening. The window for effective action is rapidly closing. According to a report by the Intergovernmental Panel on Climate Change (IPCC), global temperatures are projected to rise by 1.5°C above pre-industrial levels as early as 2030, which could have severe consequences for food and water security, as well as human health and safety [1], suggesting that we have only thirty years to significantly reduce emissions and adapt to the changing climate.

Climate models, in their most general form known as Earth System Models (ESMs), are valuable tools for addressing climate change as they provide a better understanding of the Earth's climate system to scientists and policymakers and are being used to evaluate the effectiveness of different strategies for mitigation and adaptation to climate change [2]. However, ESMs remain impractical and inflexible due to computational limitations [3]. Parameterizations, which are approximations of processes introducing free parameters, are employed to simplify climate models and make them more computationally feasible, but they introduce uncertainty, bias, and further computational bottlenecks [4]. One particular challenge within ESMs is the parameterization of Radiative Transfer (RT), describing the passage of radiation through the atmosphere, which is a key driver in the earth's climate. Current RT parameterizations are imperfect and suffer from coarse resolution, vast computational resources and uncertainties introduced by cloud dynamics. To overcome the limitations of parameterizations in climate models, Machine learning (ML) has emerged as a promising tool. However, ML surrogates face challenges, such as requiring an abundance of training data, showing bad generalization behaviour to out-of-distribution (OOD) conditions, and a risk of producing unphysical results. In climate model research, physical knowledge is plentiful and theory-guided machine learning, which integrates existing scientific principles of climate physics into ML, emerges as a promising field to improve the accuracy, efficiency and generalization capacity of climate model parameterizations [5] [6]. Physical motivated constraint ML methods can potentially solve parts of the problems of parameterizing RT in climate models, but previous work is limited. However, given the imperfect training data, coarse nominal and temporal resolution, need for trustworthy predictions and generalization to unseen conditions in the context of ESMs, exploring such methods could unlock great potential.

The resulting challenge is to emulate RT parameterization, optimizing for the tradeoff between design complexity, flexibility, physical consistency, OOD generalization, inference time and amount of training data needed.

Prior work trying to tackle this challenge includes data-driven ML methods that were shown to mimic existing numerical RT parameterizations while optimizing inference time [7], however, the physical consistency of models and data was not addressed and neither was OOD performance. Work exists on including physical knowledge encoded in differential equations into an ML driven RT parameterization [8], however, experiments were only conducted in idealized small-scale simulations and are hard to transfer to the context of ESMs. Studies using constraint methods, which are more general, flexible ways to promote physical consistent models, to emulate various parameterization schemes have shown promising results [9] [10] [11], but none were conducted in the context of RT.

To address this gap, we combine different machine learning methodologies, including Transfer Learning and a combination of physically-motivated hard and soft constraining methods [12] and evaluate resulting models comparing to purely data-driven and numerical benchmarks [7].

Our key contributions include:

- Analysis of Data and ML Models of an RT parameterization benchmark dataset (ClimART [7]) with respect to various physical constraints and general physical validity.
- Improvement of benchmark models by combining physically motivated bias, design and training methodologies into ML models.
- Analysis of computational resources, OOD generalization and physical validity of resulting models.

In the remainder of this work, we will first revisit literature on climate models, radiative transfer, parameterization and incorporating physical knowledge into ML models (see Section 2), followed by an introduction to the benchmark dataset our work is based on (see Section 3), and the methodologies we use to build on top of it (see Section 4. Our analyses of the data and results of our methods are summarized in Section 5, followed by a discussion and an outline of the limitations of our work (see Section 6).

2 Background

2.1 Climate Models

Climate models are useful tools for understanding and adapting to climate change. Climate models can be used to make projections about future climate change and its impacts on natural systems and human societies. These projections can help policymakers and communities plan for and adapt to the impacts of climate change. Also, these models can help scientists better understand the fundamental processes that drive climate variability and change, such as the interactions between the atmosphere, oceans, and land surface, and the effects of changes in solar radiation and greenhouse gas concentrations. Thus, climate models can be used to test different scenarios for reducing greenhouse gas emissions and mitigating climate change, and to evaluate the potential effectiveness of different strategies [2],[13],[14] [3].

A climate model is a numerical representation of the Earth's climate system, monitoring processes such as the atmosphere, oceans, land, and ice cover [4]. It uses Differential or Partly Differential Equations (PDEs) from physics like the conservation of energy, the Stefan-Boltzmann (describing the relationship of heat and radiation), the Clausius-Clapeyron equation (describing the vaporization of liquids), and the Navier-Stokes equation (describing conservation of mass and momentum balance of liquids) to calculate the state of the climate system. There are many types of climate models, which differ in their complexity and the subparts of the whole climate system that they model, including Energy Balance Models, Radiative Convection Models, Regional Climate Models and Global Climate Models (GCMs), the latter of which represent atmospheric and oceanic processes globally [4]. The most conventionally used most thorough models are so-called Earth system models (ESM), extending their GCM predecessors by including chemical and biological processes alongside the purely physical ones.

To numerically represent the state of the climate system, climate models need to discretize both space and time, meaning they divide space and time into a finite number of discrete points or intervals. The discretization of space involves dividing the Earth's surface into a grid of cells or elements, with each cell representing a small portion of the surface. The size of these cells can vary, depending on the resolution of the model and the computational resources available. Typically, cells range in size from hundreds of kilometres to a few hundred meters. The most common method for discretizing space in climate models is through the use of a latitude-longitude grid. This involves dividing the Earth's surface into a grid of equally spaced lines of latitude and longitude. The spacing between these lines determines the resolution of the model, with higher resolutions requiring more closely spaced lines. If the atmosphere is included in the model's representation, discretization is happening in the vertical dimension as well, introducing atmospheric layers of different altitudes. All atmospheric levels at a specific longitude-latitude grid cell are referred to as an atmospheric column. The discretization of time involves dividing time into a series of discrete intervals or time steps. The length of these time steps can vary, depending on the dynamics of the system being modelled and the computational resources available. In climate models, time steps can be in the order of minutes to hours, allowing the models to simulate processes that occur on short timescales, such as the formation and dissipation of clouds. However, simulating longer timescales, such as seasonal or decadal variability, requires models to run for many years or decades and usually involves chunking time into steps ranging over months or even years [3].

Climate models take in a wide range of inputs, including initial conditions, boundary conditions, and external forcings. Initial conditions refer to the starting conditions of the model, such as the state of the atmosphere and oceans at a particular point in time. Boundary conditions refer to the conditions at the edges of the model domain, such as the temperature and composition of the atmosphere at the top of the model atmosphere. External forcings refer to factors that influence the Earth's climate system, such as solar radiation, volcanic eruptions, aerosols and human-induced greenhouse gas emissions. [3] Forcings can be best estimates of past conditions or part of a scenario projecting future pathways of greenhouse gas emissions and socio-economic developments. The climate model experiments include

historical runs, future warming scenarios, control runs, abrupt studies, and other specialized runs. The Coupled Model Intercomparison Project (CMIP) [15] is a collaboration of many climate modelling groups to compare models by aligning them and using the same inputs to make sure that the difference in projections is due to the models and not because of the setup [4]. Important future projections run on climate models and used in policy-making include the Representative Concentration Pathways (RCPs), describing the concentration of greenhouse gases over time, and Shared Socioeconomic Pathways (SSPs) which describe possible trajectories of socioeconomic development based on assumptions about factors like population-growth, international cooperation, economic growth and others [16] [17]. The output of a climate model provides an exhaustive picture of Earth's climate, including thousands of different variables, such as temperature, precipitation, atmospheric pressure, wind speed, and sea level. The output of climate models is typically presented in the form of maps, graphs, or tables that show the simulated values of these variables over time and across different regions [3].

Challenges in climate models include their validation, which is done by comparing them against historic observations, but only little trustworthy historic data is available to do so[4]. Another limitation is the issue of scale, as many processes such as modelling clouds, can only be accurately computed on higher resolutions, however, due to the already complex nature of climate models, such resolutions are often omitted in exchange for computational performance. The time it takes to run SSP scenario predictions on climate models can vary widely depending on a variety of factors, including the specific model being used, the computational resources available, and the complexity of the scenario being simulated. For example, a study [2] by the Intergovernmental Panel on Climate Change (IPCC) used 27 different climate models to simulate the effects of various SSP scenarios on global temperatures and other climate variables, with simulations taking anywhere from a few weeks to several months to complete, whereat they operated with temporal resolutions in the realm of months only [18]. This timeframe does not yet include the actual development process and the 'spin-up' time these models need to calibrate. It is worth noting that the computational resources required to run these simulations are often substantial, with most studies using high-performance computing clusters or supercomputers to speed up the process [19]. According to [20] one scenario runs on average 390.5 days on top performance clusters, with computation steps ranging in the realm of quadrillions.

Because of these limitations in scale, validation and computational complexity, climate models remain impractical and inflexible to use to explore the full space of possibilities in climate research [21] and are often replaced by one-dimensional impulse-response models or even simple pattern scaling approaches to explore the physical system responses to a given scenario [22], which simply lack in accuracy.

2.2 Parameterizations

To simplify climate models and make them computationally more feasible, climate modelling groups make use of parameterizations. Parameterizations refer to the simplified representation of complex physical processes that are not fully resolved by the model's grid size or computational power. These parameterizations are essential in climate modelling because they allow the model to simulate climate processes that occur at scales smaller than the model's resolution. These processes may be unobserved or their physics unknown for the resolution in question [4]. A popular example is accounting for clouds, which involves particle physics that cannot be computed on a grid scale of several hundreds of kilometres [23]. The parameterizations are usually developed through guidance from theoretical principles, observations and empirical relationships, or a combination of these [24].

However, parameterizations also have several problems and limitations, as described in [3]. One major problem is that parameterizations rely on assumptions and simplifications, which can introduce errors and uncertainties into the model's output. These errors can accumulate over time and lead to significant discrepancies between model simulations and actual climate observations. Additionally, parameterizations can be difficult to validate, since the small-scale processes they represent are often

poorly observed and understood. Finally, different parameterizations can produce different results, leading to uncertainty and confusion among policymakers and stakeholders [3].

Hence, parameterizations are a great tool to speed up and generalize climate modelling, but they remain a source of uncertainty, introduce bias and often remain computational bottlenecks as is the case with the parameterization of Radiative Transfer in all ESMs.

2.3 Radiative Transfer

Radiative transfer (RT) is the process by which energy is transferred from the sun to the earth and then from the earth back again out into space. As such, RT controls the balance of energy in the atmosphere, which makes it a key process in the climate system. In the context of radiative transfer, fluxes refer to the amount of energy per unit area, time, and wavelength that is transported through a medium, such as the atmosphere. The flux is defined as the product of the intensity and the cosine of the angle between the direction of the intensity and the normal to the surface [25]. Intensity is a measure of the radiation propagating in a particular direction and is typically expressed as a function of wavelength and direction [25]. Fluxes are important because they determine the energy balance of the atmosphere and the temperature distribution of each layer, which results in the heating or cooling of these layers over time, a process that is quantified by heating rates. The radiative transfer equation is used to calculate the fluxes for each layer of the atmosphere [25] [26]. The key quantities in radiative transfer calculations are the optical properties, which include the optical depth, scattering albedo, and extinction coefficient [25] [26]. The optical depth describes the passage of radiation through a medium and is dependent on the density and absorbent properties of molecules in that medium. The scattering albedo measures the likelihood that a photon will scatter rather than be absorbed, and the extinction coefficient represents the sum of the absorption and scattering properties of the medium. These quantities are wavelength-dependent. Generally, radiation travels through the atmosphere in many directions and photons are absorbed and re-emitted or scattered multiple times in interaction with chemicals that are present in the atmosphere before they exit it, which may even alter the photons' wavelengths [25]. See 1 for a schematic overview of this process.

Therefore, fully-fledged radiative transfer calculations of fluxes and temperature at each layer of the atmosphere are computationally intractable, as they require integration over all possible wavelengths, angles and directions [25].

2.4 Radiative Transfer Parameterizations

Radiative Transfer is conventionally parameterized in all current Earth System Models. Although details of the parameterization, like specifically chosen parameters, differ from model to model, most share underlying principles. In the following, we will explain how the RT parameterization is done in the Canadian Earth System Model (CanESM5), which we take as a representative, as it is part of the CMIP6 project [21] and will also serve as the basis of the data we will deal with in the remainder of this work. The CanESM5's RT parameterization consists of three major part that deals with one of the significant challenges of calculating RT, namely having to integrate over all possible directions, having to integrate over all possible wavelengths and having to deal with all possible interactions of photons [27].

Instead of considering all directions in which fluxes can travel, a common approach is to only model fluxes travelling in four distinct directions, meaning along the vertical and the horizontal axis. In most climate models, like CanESM5, things are further simplified by neglecting the horizontal propagation of radiation as well - this approach is referred to as the "two-stream approximation". Although a very coarse approximation, neglecting the horizontal flow of radiation can be reasoned for based on some physical principles and practicalities. In theory, the earth is said to have a "shallow" atmosphere, meaning that it extends much less than the earth's radius and shallow atmospheres are considered to have layers that are homogenous in the horizontal in planetary physics [25]. In practice,

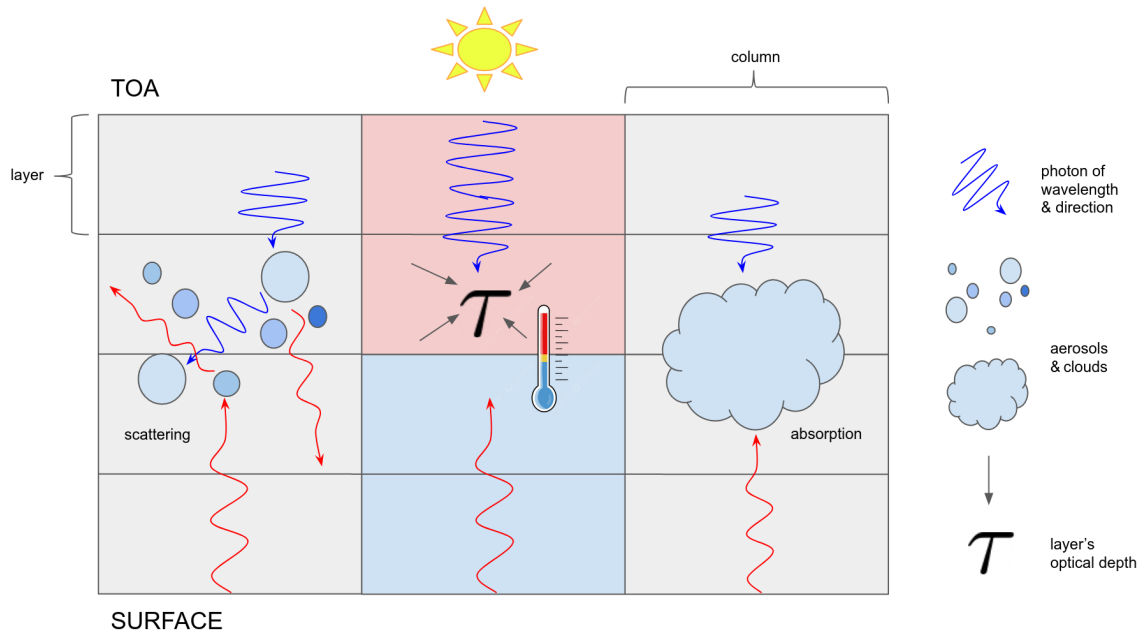


Figure 1: **Schematic Overview of Atmospheric Radiative Transfer:** The cooling and heating of atmospheric layers is dependent on various processes happening within that layer. Key quantities to consider are the amount of incoming energy by solar or thermal radiation. Photons making up the radiation travel at a specific wavelength (shortwave or longwave) and direction. Solar radiation is made up of shortwave photons arriving at the top of the atmosphere (TOA) whereas longwave radiation is coming from the surface. These influence whether the photons are absorbed or scattered by molecules present in the atmosphere, such as for example aerosols, clouds. Their density and absorbent properties are quantified by the optical depth , which describes the total passage of radiation through an atmospheric layer.

the rectification of the two-stream approximation is based on the spatial resolution the model assumes. Given columns of comparably small width, in contrast to the atmosphere's height, the vertical flux by far overwhelms the horizontal flux, in which case the horizontal flux may be dropped [27]. In two-stream approximations, the problem is then reduced to only account for up- and downwelling fluxes, meaning the amount of radiation travelling up and down between adjacent atmospheric layers, and the layer's heating rate, which can be computed based on the fluxes coming in and out of that layer.

To deal with the abundance of different wavelengths, CanESM5 makes use of a correlated k-distribution model. In a k-distribution model, instead of considering single wavelengths to account for the parameterization, the spectral gets chunked into bins, the number and size of which are calculated by probability distributions considering the optical properties of the atmosphere [28] [27]. As a result, the calculation does not have to occur integrating over a next to infinite amount of wavelengths, but only over k spectral chunks. In using this approximation, any potential altering of wavelengths caused by interactions of photons and other chemicals are neglected. Even after these simplifications, retrieving a flux computation remains computationally exhaustive, which is why CanESM5 uses the Monte-Carlo Independent Column Approximation (MCICA) to accelerate the process further [27]. This approximation is, as the name suggests, based on the assumption that atmospheric columns are independent of each other. As a consequence, it is valid to just subsample a random set of sub-columns from all possible columns at each integration point and only include these in the RT calculation, using averages to fill in the gaps for columns that have not been considered [29].

Nevertheless, this way of parameterizing radiative transfer in the earth's atmosphere still poses

problems: In the end, there remain some free coefficients whose values need to be determined, and selecting specific values for these is referred to as choosing a closure. Closures can have different implications, some even can produce unphysical results, and in general lead to bias in different climate models [25]. Another major problem is accounting for clouds, as their existence queries all the mentioned approximations [30]. A cloud may cause the up- or downwelling flow of energy to bend direction, dependent on the cloud's opacity, thus violating the assumption the two-stream approximation is based-upon. Further, given that columns are sufficiently small in width, which again is a prerequisite for the two-stream approximation, the flow of energy of a column filled with a cloud may then most certainly transgress to other columns. In that case, columns cannot be seen to be independent of each other, violating the Monte-Carlo Independent Column Approximation. The use of MCICA can also not be justified in the case of layers only being partly filled by clouds.

In the end, even though significantly sped up, calculating fluxes based on optical properties using the approximations from above is still substantially slow and thus, only done every few iterations in running a climate model, using interpolations for in-between and therewith introducing even more errors [27] [7].

In summary, conventional Radiative Transfer Parameterizations remain imperfect and are impacted by resolution and consideration of clouds, and on top of that still form computational bottlenecks in climate modelling.

2.5 Machine Learning for Climate Modelling

Machine learning (ML) has emerged as a promising tool to emulate climate models and parameterizations to overcome their limitations in speed and accuracy. Machine learning (ML) is becoming increasingly useful in climate prediction as data from climate modelling projects becomes more and more available [5]. This development comes hand in hand with the availability of more computational resources and the accessibility and distribution of AI tools [6]. ML can be used to create surrogate models emulating either the full climate model or they can be used to replace only parts of it, which in theory is called 'hybrid-modelling' [6]. [31] discusses a hybrid approach by using an ML-based image-filling technique to reconstruct missing climate information, which was used to correct previous global temperature records. Hybrid approaches are especially useful for approximations, and often ML models are used to completely replace sub-grid scale parameterizations that cannot be explicitly simulated. For example, Neural Network surrogates have been shown to speed up parameterizations of wave drag, meaning the propagation of large-scale waves through the atmosphere, in weather forecasting systems [32] and of cloud processes [33] in a General Circulation Model. Various forms of Neural Networks have also been applied to approximate turbulent processes in ocean models [34] and to model sub-grid processes, especially clouds, in the Community Atmosphere Models (CAMs) [35]. Although bringing substantial benefits in speed and accuracy in most cases, all the above studies mention drawbacks in the generalization behaviour of the Neural Networks (NNs) to out-of-distribution settings and the large quantity of training data needed [34] [35]. Out-of-distribution, in the context of climate change, refers to the phenomenon of the earth shifting its climate state in response to gradual changes in natural or anthropogenic forcings. This can be due to very rare but high-risk extreme events or long-time responses, makes it hard to test any prediction of an ESM or ML surrogate. Due to the explicit encoding of physical principles, ESMs stand a chance to predict these events accurately, whereas purely data-driven trained NNs that learned to uncover and categorize patterns to solve a specific task only, being prone to overfitting, can lack the physical knowledge needed to generalize to unseen climate states [6] [36]. However, a climate model emulator can only be useful for policymakers and climate scientists, if it can faithfully and accurately reproduce the response of ESMs.

To meet the demands ML surrogates face, meaning mostly the challenging generalization conditions and the risk of producing unphysical - and as such untrustworthy - results, designing such models should

be carefully integrated into and guided by existing scientific principles of climate physics [5] [6].

2.6 Theory-Guided Machine Learning

In climate modelling, theory-guided machine learning shows promise for improving the accuracy, efficiency and generalization capacity of climate models, especially in the context of parameterizations, while maintaining trustworthiness concerning physicality. Merging theoretical knowledge with data-driven models is a relatively new field within the machine-learning community. [36] summarizes different pathways of integrating scientific knowledge in data science models and therewith coins the term "Theory-guided Data Science" (TDSC), whereas other papers refer to similar concepts as "Process-guided" [10], "Scientific" [37] [38] or "Physics-informed" [11] [39] Machine Learning. TGDS aims to learn dependencies that have sufficient grounding in physical principles and thus have a better chance to represent causal relationships, yielding better generalizability, accuracy, simplicity and consistency [36]. There are several major options to ground a model in physical knowledge: 1. Restricting the search space of models to exclusively physically consistent ones, 2. Restricting the model's output space to an exclusively physically consistent one and 3. fine-tuning search and/or output space to increase the likelihood of physical consistency.

The first option, limiting the search space of models, can be done by using domain knowledge to initialize and constrict the points of the search space, i.e. using theory-guided priors and relationships to initialize, constrict and regularize model parameters. One concrete example of this is Physics Informed Neural Networks (PINNs) [39] [37]. PINNs work by having a network predict physical quantities of a known PDE. Residuals of the PDE, like terms defining initial and boundary conditions, are then computed based on the gradients of the network and then added as extra terms to the overall loss function of the network. Thus, PINNs can produce an estimated solution for a given point in the integration domain of the PDE. A related idea is the one of Neural Differential Equations (NDE) [40] which are deep neural networks predicting a hidden state which is then passed through a differential equation solver. The network is then trained by backpropagating through the solver. There have been efforts to define a universal framework to bring together deep learning methods and knowledge encoded in differential equations [38], however, all of these methods require more or less exact knowledge of the structure of underlying mathematical processes, which is not always available.

Thee second option, constricting the output rather than constricting the search space, only requires knowledge about conditions resulting outputs must meet to be physically valid, but not about the structure of the underlying process. Strict constraining of the output space can be achieved by, for example, using a meaningful architecture design, parameter initializations and probabilistic relationships within the model that are grounded in domain knowledge. In the literature, these approaches are referred to as Hard Constraints on Neural Networks, whereas so far only linear constraints have been sufficiently studied[12]. Hard (linear) Constraints are implemented by adding additional parameters to the network that can be interpreted as defining a linear equation system operating upon the immediate network's output as inputs [12]. Reparametrization by such layers enforces that outputs obey the constraints encoded in the linear equation system. Theory-motivated constraint methods can also be much simpler, such as rescaling and ensuring positivity (through for example the application of a softmax layer in a neural network) to meet conservation laws [9]. In the context of climate modelling, enforcing linear constraints have been shown to improve emulating of various subprocesses within ESMs, such as ensuring conservation laws in convective processes [41], for emulating aerosol microphysics [42] and also have been shown to improve downscaling of ESMs (meaning to increase their nominal resolution thus the resolution of the grid on which the data is presented on = extrapolate to smaller grid sizes) [9]. Hard Constraints show promising results with respect to producing physically accurate results and generalizing to out-of-distribution scenarios, however, Hard Constraints can also limit the flexibility of the model, making it difficult to fit the available data [12].

Asa third option, rather than using firm constrictions of search and output spaces, explorations of data-driven models can also be guided more softly by, for example, introducing physics-guided loss

terms [43]. This fine-tuning of the output space is also referred to as "Soft Constraints" in the literature [12]. In the context of climate modelling, studies have been conducted using physics-based loss penalties to ensure relationships between the density of water at different depths in lake water temperature predictions [11]. This work has been extended to include more complex physical relationships that happen on a temporal scale, like energy conservation in the thermal gain of the lake in response to in- and outcoming thermodynamic fluxes [10]. In these studies, the fine-tuning of the output space happened using observational data, subsequent to pretraining the model parameters to predictions of a process-based model. While Soft Constraints are reported to increase OOD generalization capabilities, are flexible and allow a better fit to data than Hard Constraints, resulting models may not always be physically meaningful.

While all these methods use knowledge from theory to introduce meaningful bias in data-driven models, TGDS can also be thought the other way around, augmenting theoretical models with data-driven feature extraction that may operate on observational data [6] [5]. The aforementioned study using an image-filling technique to reconstruct missing climate information can also be interpreted as an example of that [31].

In conclusion, TGDS seems like a promising candidate to solve the problem of parameterizing RT in climate models, being able to tap into the resources of already allocated physical knowledge and combining them with the predictive power of data driven methods.

2.7 Theory-Guided Machine Learning for Radiative Transfer Parameterization

There is only little previous work on ML-driven RT parameterizations available, mostly due to a lack of awareness, data and benchmarks outside of the climate-modelling community. Accessibility of climate modelling problems to ML research is a general problem, seeding the slow progress. Recent efforts trying to bring change include benchmark datasets built to present the data in a way to get ML developers started without needing domain knowledge and to establish benchmarks, against which new methods can compete. Up to the start of this work, only two published benchmark datasets exist, one concerned with general ESM emulation [22] and one especially concerned with RT parameterization, named ClimART [7].

ClimART provides input and target data extracted from the Canadian Earth System Model (CanESM5) and it's RT parameterization in different conditions: With incorporating clouds and aerosols, only including aerosols or nothing of the two. A ClimART ML model's goal is to predict up- and downwelling fluxes based on a bunch of features of the atmosphere, like for example chemical abundances of different gases, pressures and geographical information. The flux prediction happens on a scale of a single atmospheric column and heating rates are numerically computed based on these, as this step in the numerical RT parameterization can be done accurately and fast [7]. ClimART includes data from pre-industrial, near-present and future simulations drawn from the actual climate model, additionally providing an evaluation pipeline for OOD testing. The work includes released statistics on training very simple NN-based models to fitting to the target RT parameterization, which shows promising results in the realm of accuracy and especially interference speed-up, however, these baseline models especially lack in their generalization capabilities [7]. Except for introducing some architectural bias by using Graph Neural Network (GNN) to exploit the nature of the structure of the input data (i.e. translating layers of an atmospheric column in a GNN where nodes represent layers and only adjacent layers/nodes are connected) no experiments with incorporating or exploiting physical knowledge were conducted, nor did they question the physical meaning of model predictions and target data, although the latter is known to be physically imperfect resulting from the approximations employed in CanESM5.

Little research exists for theory-guided ML techniques in the context of RT: [8] elaborates on using PINNs to solve the RT equation solving for radiative intensity based on a given point in space,

time, angles and frequencies [8]. PINNs are trained to emulate numerical approximations of RT in specialized and more general simulations, showing good generalization capacities at a fraction of the cost of original numerical methods [8]. However, their experiments are handcrafted high-resolution physical simulations and are hardly transferable to a more general application like ESMs, as it is the challenge for example in ClimART where information like angles and frequencies are not available, due to the many approximations done by CanESM5, like the two-stream approximation (only vertical flow, no angles) or the spectral-binning (no individual frequencies, just chunks of the spectrum), confounding underlying PDEs to invalidity.

A much more promising approach to tackle RT parameterization in ESMs in practice seem to be theory-motivated constraint methods as they trade-off flexibility, accuracy and OOD generalization performance. In the discussion of whether Hard Constraints or Soft Constraints constraints yield better results, opinions are mixed. In the aforementioned study [12], arguments are in favour of hard constraining methods, as they claim that the joint optimization that soft constraining requires is inefficient and leads to convergence issues. Opposing that, another study argues that Hard Constraints are computationally feasible but theoretical benefits and results do not outweigh existing soft constraining methods [44]. Thus, imposing Hard Constraints must be carefully done using handcrafted theoretical knowledge. In summary, both approaches have advantages and disadvantages depending on the specific problem being addressed, whereas Soft Constraints can be more flexible and allow for some violations of physical principles, which may be necessary for accurate predictions, and Hard Constraints ensure exact satisfaction of physical principles and can provide stronger guarantees on the validity of the model outputs, but are less flexible and require more accurate knowledge in their design. However, both types can be beneficial, especially for OOD generalization, which is one of the main issues reported for the ML baseline models in ClimART [7].

Given the setting of imperfect training data, coarse nominal and temporal resolution, a need for trustworthy predictions, a need for generalization to unseen conditions, availability of plentiful physical knowledge and no prior research conducted on the matter, we see the potential of exploring physical motivated constraint ML methods for emulating RT in the context of ESMs.

3 Data

3.1 The ClimART dataset

ClimART is a Benchmark Dataset for Emulating Atmospheric Radiative Transfer in Climate Model parameterizations[7]. It is based on a single climate model, namely the CanESM5 [27], and contains global snapshots of its atmospheric state covering historic and future runs with a frequency of 205h, meaning 43 snapshots per year for training data.

The spatial resolution offered is 128 in longitude and 64 in latitude, resulting in a 3-degree resolution. Due to the spherical shape of the earth, grid cells will have varying sizes, ranging from around $300km^2$in the polar region up to $900km^2$ near the equator. Each grid cell also consists of multiple layers in altitude which altogether forms an atmospheric column, resulting in 8192 columns in total. The dataset considers two conditions: "pristine sky", meaning no aerosols and no clouds and "clear sky", meaning just no clouds, which differ in the number and nature of variables included in the input and impacts the nature of the targets taken from the CanESM5 model.

3.2 Inputs

Every single item in the dataset represents one of the atmospheric columns. The input variables are defined on different scales: There are layer variables such as pressure, water vapour etc. which have one value for each layer in the column. Then there are level variables, with levels being the interface between adjacent layers, including pressures, height and temperature information. These also have one value per level. Additionally, there are global variables which are independent of discretization such as optical properties, geographical information, boundary conditions etc.

Dimensions Determining the dimensionality of the dataset includes considering a set of years y_s, the number of examples pear year s_y (meaning global snapshots at a specific point in time during that year, in short referred to as snaps) and the number of atmospheric columns these snapshots entail (set to 8192 in ClimART). The number of snaps differs between training years (43 or 42 for leap years) and test years (15). Thus, the total number of items in the dataset N with each entry being one column, is computed as $N = y_s \times s_y \times 8192$. Each column has $S_{lay} = 49$, $S_{lev} = 50$ number of layers and levels (number of layers plus one as these represent interfaces between layers) per column with differing number of features associated with each layer ($D_{lay} = 14$ (pristine sky condition) or 45 (clear sky) and level ($D_{lev} = 4$) plus additional global features ($D_{glob} = 82$).

Inputs are represented by longitudinal and lateral indices or the grid that get translated to 3D coordinates to ease usage by ML models. This is done by mapping the latitude and longitude information on a unit sphere to retrieve the coordinates, as can be seen in Figure 2:

$$x - cord = cos(lat) * cos(lan)$$
$$y - cord = cos(lat) * sin(lon)|$$
$$z - cor = sin(lat)$$

These coordinates are simply included as part of the global features that come with each column.

The full dimensionality of the 3-parted dataset is thus as follows:

- layers (N, $S_{lay} = 49$, $D_{lay} = 14 or = 45$)
- levels (N, $S_{lev} = 50$, $D_{lev} = 4$)
- globals (N, $D_{glob} = 82$)

A schematic of input variables can be seen in Figure 3.

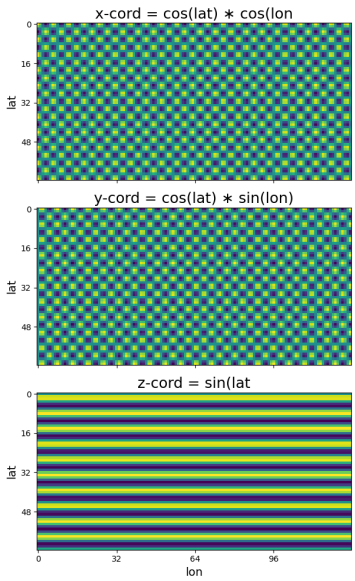


Figure 2: **Visualization of grid representation:** The 2D grid information (longitude, latitude) is mapped onto a unit sphere, retrieving a 3D representation to ease use by an ML model.

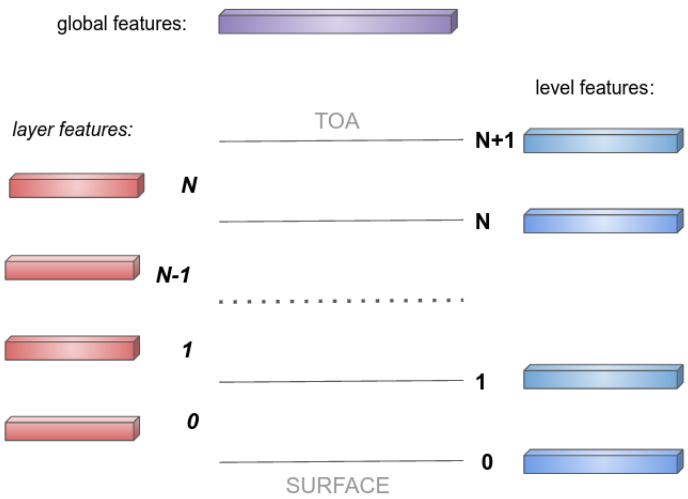


Figure 3: **Schematic overview of input variables:** Inputs are provided columnwise with N atmospheric layers, each of which represented by a feature vector of layer variables (red bar) and $N + 1$ interfaces between these layers (levels) each of which represented with a feature vector of level variables (blue bar). Some variables are not spatially tied to levels nor layers and are represented by a global feature vector (purple bar).

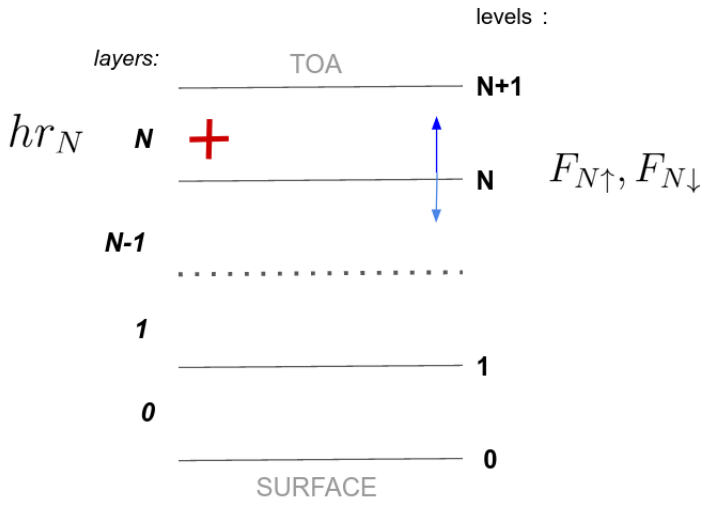


Figure 4: **ClimART Target Data Structure:** For each level (=interfaces between atmospheric layers) n an upwelling flux $F_{n\uparrow}$ and a downwelling flux $F_{n\downarrow}$ are predicted (depicted with blue arrows). From the fluxes of adjacent levels, the rise or fall in temperature in a corresponding atmospheric layer is computed, captured in the layer's heating rate hr (depicted as red plus sign).

3.3 Targets

The targets delivered by ClimART come from the CanESM5s RT physics model. Like the inputs, these come in two conditions: pristine and clear sky. For each condition, both shortwave and longwave calculations are available, each giving up- and downwelling fluxes and heating rates of the column in question. The fluxes are level variables, whereas the heating rates are layer variables. A schematic overview of target variables can be seen in Figure 4

In total, up to 6 target variables can thus be retrieved for each condition offering the opportunity to freely choose the desired target variables (e.g. only considering shortwave radiation, only longwave, both...). An example for downwelling shortwave flux targets for different snaps of a test year are shown in fig 5.

3.4 Dataset Splits

The recommended test split includes data covering years 1979 to 2004, excluding the years 1991-1993 due to the prevalence of an extraordinary volcano eruption. For validation, the years 2004 and 2005 are used. Next to the main test set, ranging from 2007 to 2014 with 15 snaps per year, thus targeting longer-term predictions of more present day conditions. Additionally, to this historic data, the dataset offers test sets with distributional shifts in order to evaluate the robustness of a potential climate model emulator.

Out-of-distribution (OOD) test set: This set covers the year of the Mount Pinatubo eruption 1991, thus probing how well the ML model can cope with a sudden and temporal shift. To avoid information leakage by remnants of increased stratospheric aerosols, the subsequent two years are excluded from training, as the impact of the eruption on the global climate was still measurable over this time [45].

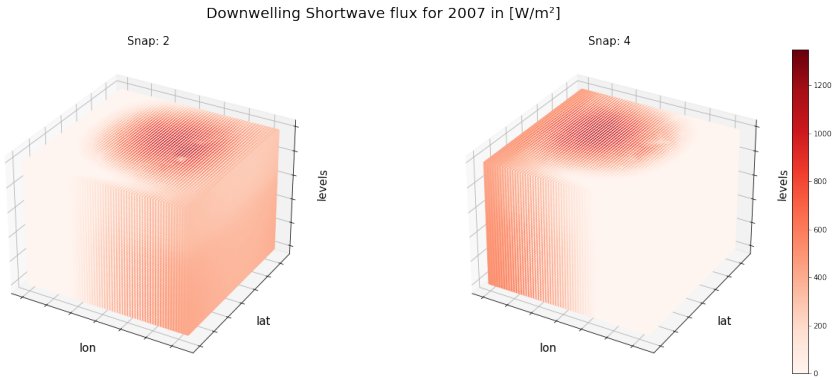


Figure 5: **Target Downwelling Shortwave Radiation Example:** Example of Downwelling Shortwave Radiation (rsdc) for different snaps of the test year 2007 taken from the Target RT Parameterization. For each atmospheric column located at one grid location over longitude and latitude, we get one positive flux value (W/m^2) for each level.

Historic test set: Pre-industrial years covering 1850-1851 can probe the model's response to unseen surface-atmosphere conditions and assess physical correctness when greenhouse gas concentrations are reduced as the target parameterization tends to be fine-tuned to real observational data.

Future test set: Far-future predictions shows how well the ML model can extrapolate conditions differing from the current state of the climate in response to the forcing following one of the CMIP6 scenarios (SSP585 - a pathway representing an upper boundary with green house gases emissions). For this split, the far-in-the-future years covering 2097-2099 from the 100-year coverage of the scenario projection were taken (we note, that this required some sort of interpolation by the authors of [7] as raw CMIP6 scenario data is only available in 10-years intervals).

3.5 Normalization

As the input variables, regardless of belonging to layer, level or global variable type, cover distinctly different units and ranges, these should be normalized such that all values end up being in the operation range of the respective activation function used by the network architecture of choice.
In [7] this is efficiently done by precomputing statistics on all variable types of all training years provided in the dataset, meaning computing a mean and standard deviation for each variable type over all years. Using these precomputed statistics, a normalization method such as z normalization (resulting in zero mean and unit standard deviation) or a choice of min-max with or without log normalization can be applied.
The same procedure is applied to the targets.

Years that are not part of the training set, such as test years or OOD test years, are normalized using the statistics obtained from the training years and not the test years. This is necessary to avoid any information leak from the independent test set. However, this procedure should be taken with a grain of salt: Distributional shifts in the test years may result in a distribution consisting of many far outliers in the new normalized space.

3.6 Heating Rate Computation

As mentioned, heating rates are dynamically computed from the flux predictions, as this part of the RT parameterization is accurate and fast [7]. The heating rate hr_l of any given layer $l \in 1, .., S_{lay}$ is calculated based on the up- and down-welling fluxes of the two adjacent levels as follows:

$$hr_l = \ c \times \frac{(F^{\uparrow}_{l+1} - F_{t+1}) - (F^{\uparrow}_{l} - F_{t})}{p^{lev}_{l+1} - p^{lev}_{l}} \tag{1}$$

where $c \sim 9.76e{-}3$. $F^{\uparrow}_l$, F_t, p^{lev}_l are the respective up- and downwelling fluxes and level pressures of a level $l \in 1, .., S_{lay} + 1$ [7].

4 Methods

4.1 RT Constraints

After analysing the data and pretrained models, we selected two physical requirements any RT calculation should meet, from which we derive soft and hard constraint methods for training. We selected those on the basis of the magnitude of violation by the target data and the baseline models, and their feasibility in implementing into the training pipeline. For a more detailed discussion of the selection process of these requirements, please see Appendix section 8.2.

4.1.1 Energy Conservation Constraint

Considering that Radiative Transfer describes the transfer of energy, Energy Conservation is a a natural choice to formulate a constraint.

Following the assumption that atmospheric columns are independent of each other, as it is made in [7], the conservation of energy must be fulfilled by a single column. This means that the incoming energy at TOA must equal the energy absorbed by the atmosphere (A_{atm}) plus the energy absorbed by the surface (A_{srfc}) plus the energy leaving the atmosphere again at TOA as flux ($F_0^{\uparrow}$).

One would expect to have to handle the energy conservation calculation differently for the two wavelength forms, shortwave and longwave, as there is no incoming longwave radiation at TOA.

However, this is dependent on how the cut-off line separating shortwave and longwave wavelengths is defined. For example, if one takes a wavelength of roughly 4 microns as the dividing line between shortwave and longwave radiation, there is about $12W/m^2$ of incoming solar radiation at wavelengths greater than 4 microns, resulting thus in an "incoming logwave flux". The treatment of incoming radiation at the TOA for longwave is particular to the radiative transfer model. Most radiative transfer models put this incoming radiation into their shortwave calculations, albeit in different manners, e.g., adding the additional incoming to the wavelength interval close to 4 microns. In the RT code underlying the parameterization data from ClimART however, the incoming solar radiation was explicitly accounted for in the longwave radiative transfer calculation [27]. This is very specific to the exact climate model (CanESM5) and allows us to treat the two types in the same way in our analysis.

Thus, considering a single column, N layers and $N+1$ levels with index 0 describing the top of the atmosphere (TOA) and N being the surface, up- $\vec{F}_{\uparrow}$ and downwelling $\vec{F}_{\downarrow}$ fluxes defined over levels, and the two wavelength-types given in the data $w \in SWR, LWR$, we can formulate this error mathematically as follows:

$$ECC = \sum_{w \in \{SWR, LWR\}} F_0^{w,\downarrow} - A_{atm}^w - A_s^w - F_{c_0}^{w,\uparrow} \tag{2}$$

$$A_{atm}^w = ((F_{c_0}^{w,\downarrow} - F_{c_0}^{w,\uparrow}) - (F_{c_{N+1}}^{w,\downarrow} - F_{c_{N+1}}^{w,\uparrow})), \tag{3}$$

$$A_{srfc}^w = (F_{c_{N+1}}^{w,\downarrow} - F_{c_{N+1}}^{w,\uparrow}) \tag{4}$$

We compare the energy conservation error of both the data and predictions of a simple MLP baseline, concluding that absolute errors for both target RT and model predictions remain rather small over all test sets. The target RT seems to account for energy conservation in its design, and the data-driven models seem to be able to learn and encode that property. For a detailed analysis, refer to the Appendix section 8.2.1.

All in all, the energy conservation constraint is already implicitly encoded in training but must be respected in the further design of any deep learning methods such that it still withholds.

4.1.2 Integral Balance Constraint

After an analysis of significance using pre-trained models, we selected a column-wise constraint that both model and target parameterization do not account for. This constraint explores the relationship between fluxes, heating rate and level of pressures and will be referred to as the "balance constraint" in the remainder of this work.

The integral balance constraint as described in [46] (Appendix section A) states that the two outputs from the radiative transfer parameterization, fluxes and heating rates, are not completely independent of each other, as there exists an integral relationship between them. This relation is described per atmospheric column by equation 12. Here $k = 1..N$ stands for the vertical levels, with 0 being TOA and N being the surface level. $\alpha_k = (p_k - p_{k-1})G^{-1}$ describes the pressure difference between levels k and $k-1$ with the constant $G = (86.400s * 9.81ms^{-2}/1005.5Jkg^{-1}K^{-1})$ set to seconds per day times the gravitational constant divided by the specific heat constant. G is necessary to make the units consistent with each other. h_k stands for the heating rate at a given level k. Φ 13 is usually computed differently for the two cases, short wave and long wave radiation, as there is no natural incident longwave radiation at the top of the atmosphere. However, the CanESM5 uses a trick to account for solar radiation with a wavelength slightly below the dividing line between shortwave and longwave radiation, simply factoring it in as incoming radiation in the longwave radiative transfer calculations (as described in Section 4.1.1 [27]). Hence, in our case, Φ is equal to the SWR case for both types of radiation. The equations below describe the balance constraint with up- and downwelling fluxes noted as $\vec{F_\uparrow}$ and $\vec{F_\downarrow}$ respectively and hr being the heating rates.

$$BC(F^\uparrow, F^\downarrow, hr) = \frac{\sum_{k=1}^{N} \alpha_k h_k}{\sum_{k=1}^{N} \alpha_k} + \frac{\Phi}{\sum_{k=1}^{N} \alpha_k} = 0 \tag{5}$$

$$\Phi = \begin{cases} F_0^\uparrow - F_N^\uparrow + F_N^\downarrow, LWR \\ F_0^\uparrow - F_0^\downarrow - F_N^\uparrow + F_N^\downarrow, SWR \end{cases} \tag{6}$$

We compare the integral balance error of both the data and predictions of a prtrained simple MLP baseline model provided by the authors of ClimART, concluding that, although rather small for the main test years, absolute errors for both target RT and model predictions grow bigger for OOD test sets alongside a greater divergence in behaviour between target RT and models. For a detailed analysis, refer to Appendix section 8.2.5.

All in all, the integral balance constraint seems like a promising candidate to incorporate in the training design of deep learning methods to help improve physical consistency and generalization ability.

4.2 Soft Constraints: Loss Functions

4.2.1 Base loss

For the base loss, we consider the loss as it is computed in [7], meaning a simple mean squared error computation weighted and summed over the different parts of the radiative transfer parameterization. Spelt out, a weighted sum over each wavelength type SWR, LWR is computed, for each of which another weighted sum over the similarity of the upwards and downward fluxes $\vec{F^\uparrow}$, $\vec{F^\downarrow}$ as well as the heating rates $\vec{hr}$ between the model's predictions y and the target parameterization t is computed. The total base loss can thus be written as follows:

$$\mathcal{B}(\vec{y}|\vec{x}) = \sum_{k \in k_{SWR}, k_{LWR}} k \times (k_{up} \times MSE(\vec{F_y^\uparrow}, \vec{F_t^\uparrow}) + w_{down} \times MSE\ (\vec{F_y^\downarrow}, \vec{F_t^\downarrow}) + w_{hr} \times MSE(\vec{hr_y}, \vec{hr_t})) \tag{7}$$

with k being a placeholder for the free parameters, indicating the amount of contribution for each subpart. This loss will push the models to imitate the behaviour of the target parameterization, already indirectly learning basic dynamics of RT, like conservation of energy.

4.2.2 Theory-guided: Constraint loss

The single value return calculated from the violation of the Integral Balance Constraint (see Section 4.1.2 can be interpreted to be a deviation from the ground truth, which in this case is always 0. We thus use a squared version of the balance error based on a model's flux prediction $BC(\vec{F}^{\uparrow}_{y,w}, \vec{F}^{\downarrow}_{y,w}, \vec{hr}_{y,w})$ for $w \in SWR, LWR$ (see equation 5) to serve as our constraint-based loss. As we treat the two wavelengths independently, we can formulate an overall loss can as a weighted combination of the balance errors per wavelength type with free weight parameters k:

$$\mathcal{C}(\vec{y}|\vec{x}) = \sum_{w \in SWR, LWR} k_w \times BC(\vec{F}^{\uparrow}_{y,w}, \vec{F}^{\downarrow}_{y,w}, \vec{hr}_{y,w})^2 \tag{8}$$

Models trained with this constraint loss will only learn to satisfy the constraint, but not learn about any other properties of flux predictions. However, it is possible to minimize the constraint loss without changing absolute amounts of fluxes by mere redistributing of fluxes across the vertical. This is due to the weighted integration of heating rates that are directly computed from the fluxes. This loss could thus be used to refine the output sapce of models.

4.2.3 Theory-guided: Combined loss

As described in [41], [9][12][10], terms expressing how well physical constraints are satisfied by a model's output can be added to the model's original loss term during training. By doing so, the model learns to minimize the error connected to a physical property that the output should maintain. Formally, this is known as soft constraining. Hereat, $\mathcal{L}$ is the total loss of some prediction y given an input x and the target t using some base loss function $\mathcal{B}$, often implemented as a similarity function, and an instance of a constraint error function $\mathcal{C}(y)$. The amount of contribution of the original loss and the constraint loss to the total loss can be controlled by introducing additional free weight parameters (k_b, k_c). The overall loss in our case, considering the ClimART base loss and the integral balance error, can thus be expressed as follows:

$$\mathcal{L}(\vec{y}|\vec{x}) = \; k_b \times \mathcal{B}(\vec{y}, \vec{t}) + k_c \times \mathcal{C}(\vec{y}) \tag{9}$$

This combined loss allows models to learn to imitate the target parameterization, thus acquiring basic dynamics of RT, while at the same time trying to minimize the external constraint which would mean deviating slightly from the targets. Exploitation of imitating the targets and exploration by moving away from them to satisfy the external constraint can be controlled, even dynamically, by the weighting parameters.

4.3 Bias and Hard Constraints: Architectures

4.3.1 ClimART Base Models

For the first stage of our training workflow, meaning the training of base models, we choose two different baseline architectures from the previous work on ClimART [7]. The authors train an ensemble of models ranging from a simple Multilayer Perceptron (MLP) [47] and a one-dimensional Convolutional Neural Network (CNN) [48] to more advanced architectures such as a Graph Convolutional Neural Network (GCN) [49] and a Graph Neural Network (GNN) [50].

While the MLP is reported to perform the worst in terms of evaluation metrics, it is reported to have the best inference time. We thus choose this model as a baseline for comparison making it possible to assess if an increase in performance can be singled out to be due to the altered workflow and not only due to a more elaborate model design. As a second base model, we choose the GNN architecture, as it is reported to have similar performance in terms of flux predictions to its CNN competitors, but a significantly reduced inference time given the availability of a GPU. Also, a GNN can be quite flexibly extended to exploit the physical properties of the input data [50] even more, making it a good starting point for future work based on this project.

ClimART MLP The MLP presented in ClimART is a shallow, fully connected 3-layer neural network. The input data gets transformed into a single flattened unstructured feature vector, mapping to a single output vector. For the predictions, up- and downwards fluxes for each wavelength type considered are stacked upon each other. By slicing the flat output vector into pieces of size N (number of levels), we obtain the specific predictions for each component.

ClimART GNN Graph neural networks (GNNs) [50] deep learning based that have been on the rise in recent years, due to their convincing performance exploiting structured data for a multitude of classification and regression tasks [51]. GNNs operate on a graph domain, meaning that input data is presented in the form of a structured graph and the learnable parameters of such a network are also defined structurally. This allows for the introduction of several learning biases, specifying information flow and relationships present in the data. The specific design of the graph network architectures can be adjusted to the task at hand and should be carefully chosen, to provide useful and, at best, physically consistent biases and not confound or allow unphysical processes. Once a graph has been built, several tasks can be learned based on it, e.g. graph classification or node regression, thus either working on the full graph or only on certain aspects of it.

A Graph Neural Network can entail several blocks, each of which can be formally defined as a 3-tuple $G = (u, V, E)$ of nodes, edges defining relationships between nodes and global information [50]. For the set of nodes $V = v_{ii=1:N^v}$, each v_i is a node's attribute (can be represented as a vector of size N^v or even another graph). Nodes are connected by edges, in the set of which $E = (e_k, r_k, s_k)_{k=x:N^e}$ each e_k is an edge's attribute, connecting a receiver node with index r_k with a sender node with index s_k. Edges thus define the structure and information flow of the graph. The global attribute u, often also referred to as another node, can be a vector set or even another graph and is typically connected to all other nodes in the graph, making certain information globally available.

In addition to these network parameters, a Graph Network Block consists of several update and aggregation functions which are applied in a specific order to subsequently distribute information within the block and update its parameters. Typically, edges are updated by incorporating information from adjacent nodes first. Next, for each node in the network, an aggregation function mapping each updated edge adjacent to the node in question to a single update is applied, which is then used to update the node. Next, the updated edge and node features are aggregated to map to the update function of the global features. The per-edge and per-node aggregation and update functions can be reused on all edges and nodes respectively and thus induce a combinatorial generalization bias, meaning the final Graph Network Block can operate on a graph of different sizes and shapes.

Several of these graph network blocks can be stacked upon each other, although it has been shown that deeper Graph Neural Networks typically deteriorate in performance in comparison to their unstructured counterparts [52]. Once the final graph is obtained, one can either build a classification task by classifying the full graph or parts of it. Dependent on the task in question, taking updated node or edge features as the final output can also make sense, which is referred to as performing node or edge regression respectively.

In the ClimART training pipelines, atmospheric columns are treated as identically and independently distributed samples, justifying that each column can be interpreted as a graph of the same

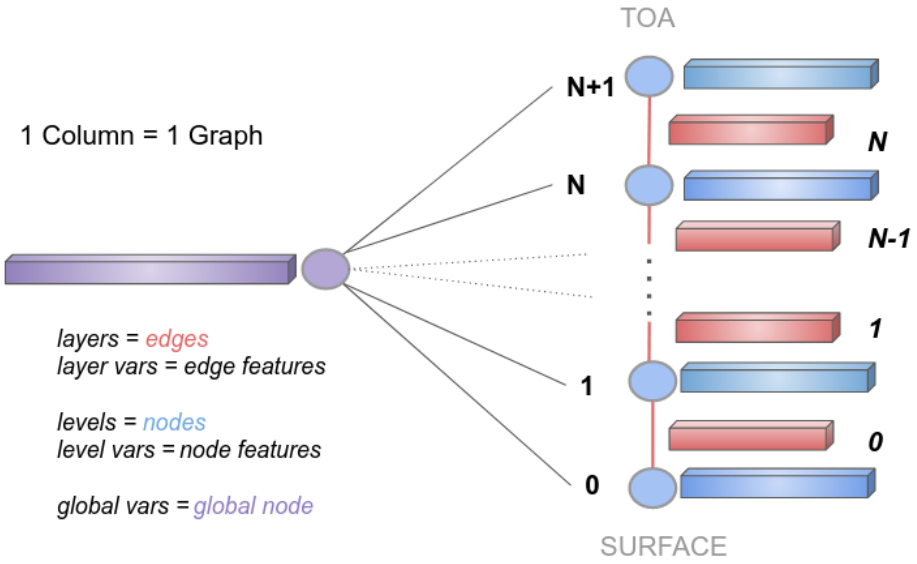


Figure 6: **Forming a ClimART GNN from the input data structure:** levels variables become nodes (blue circles) with associated node features (blue bar), layer variables become edges (red lines) connecting adjacent levels / nodes with associated edge features (red bars) and the global variables (non-spatial) become the global node (purple circle) with an associated feature vector (purple bar).

structure. Translating the column features to a graph is rather intuitive: Level features are defined to be the graph's nodes, layer features the edges, and global features the global node. For the adjacency matrix, defining the structure of the graph, only nodes that correspond to levels being right above or below each other are connected, resulting in a line graph. The translation from the ClimART data structure into the GNN is shown in Figure 6.

The GNN baseline in ClimART is a three-layer GNN, meaning three sequential GNN blocks that do not share weights. The three update functions of each block (global, node and edge updates) are modelled by single-layer MLPS and aggregation functions are modelled by mean-pooling.

In principle, one could interpret the updated node features of the last layer's block to already being the flux representations, turning this into a node regression task. However, in ClimART the authors use an additional readout layer, mapping the final mean-pooled node representations to a flattened output vector just like in its simple MLP counterpart.

In the remainder of this work, we will adapt the ClimART GNN design, which accounts for basic physical relationships present within the input data.

4.3.2 Theory-guided: 2 head models

In the simple transfer paradigm, pre-existing weights from a model, which was usually trained on a slightly different task or data coming from a similar domain, are taken as initialization weights for a new model for either all layers, or just the early ones [53]. This new model will then be trained on the new task. There are different approaches in respect to what weights should be allowed to change at what rate during training on the new task. In the following, we will define a physically motivated 'head' architecture. This head architecture introduces additional parameters to a flux-predicting pretrained model, optimally one trained on ClimART and its base loss. These additional parameters can then be learned to allow fine-tuning of flux predictions.

Two-Head MLP We define a Two-Head model to consist of two parts: A pre-trained base model that has been trained on the original task and a second head, taking the output of the pre-trained model as input. While fine-tuning in a subsequent learning stage, all weights of the base model remain unchanged and only the weights of the new readout layer - or second 'head' - will be adjusted by gradient descent.

The task of the second head will be to rescale the flux predictions according to some loss while not changing certain physical properties of the base model's flux prediction:

- All flux predictions remain positive.
- The total energy contained in the up- and downwelling fluxes respectively may not change.

From this, we derive Hard Constraints that we encode in the Two-Head architecture, namely a positivity constraint and an energy conservation constraint:

First, we obtain the base model's flux predictions, from which we determine the total energy conserved in the two flux flows respectively by taking the sum over levels for each aspect (up- and downwelling). Next, the base models' flux predictions per aspect are passed through a readout layer to obtain rescaling logits. These logits are passed through a softmax, again per aspect, ensuring both positivity and a unit sum. The resulting values are then rescaled using the total sum of flux energy per aspect computed from the original base model prediction to obtain the final rescaled flux prediction.

Thus, this architecture is easy to implement on top of existing ClimART baseline models, allowing deviations while maintaining conservation of energy. This mixture of transfer learning and constraint-motivated architecture choice should introduce a bias in the model space, making training more efficient. A simple overview of the Two-Head design is shown in Figure 7.

4.3.3 Flux rescaling

The ClimART paper has an issue in its training pipeline design: The models are allowed to learn to predict an external factor (incoming flux at the top of the atmosphere = solar radiation coming into the atmosphere) that should not be able to be influenced by the model. That happens because this feature is only present in the targets, as TOA downwelling flux value, but not in the model's input data.

Luckily, this value stays approximately constant over time and space, with an average incoming annual solar radiation at the top of the Earth's atmosphere of around $1361 W/m^2$, which models may indirectly learn. Nevertheless, models should still not be able to alter this value to, for example, minimize any external constraints by just minimizing the absolute amount of new energy in the system and thus minimizing absolute errors.

However, as in column-wise flux predictions, the relationship between resulting heating rates and net TOA flux is near linear [54], it is possible to rescale flux predictions by a constant and still retrieve meaningful predictions. Thus, as a work-around, we rescale the RT target fluxes by dividing with the incoming solar flux at TOA and use this exact same value to scale up the model flux predictions for each training sample. Additionally, the TOA fluxes of the models predictions are manually set to match the incoming flux as present in the targets.

We can understand this rescaling approach as another theory-guided hard constraint we employ in our training design.

4.4 Training Procedure

Following the work of [10] [9], we devise a deep learning framework based on the principles of introducing hard and Soft Constraints in combination with transfer learning to achieve a tradeoff optimization

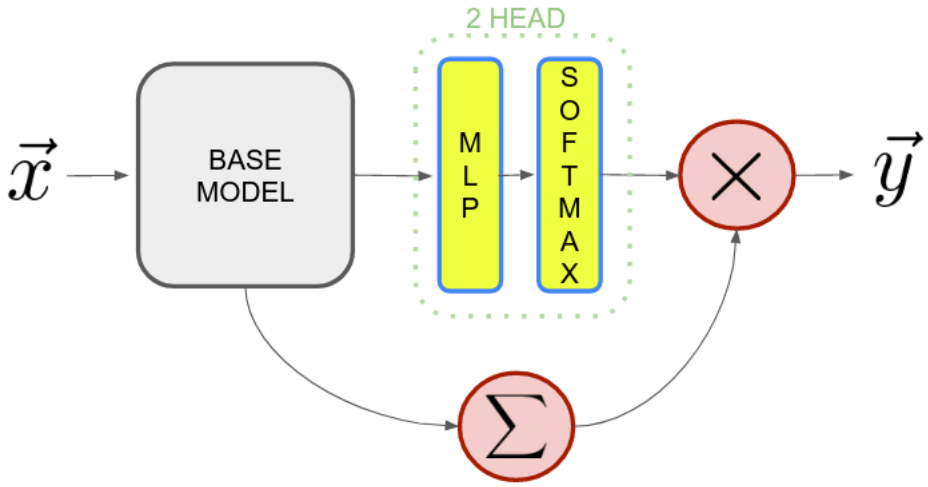


Figure 7: **Simplified view of the two head model:** Considering one wavelength and one aspect (up- or downwelling flux), the input $\vec{x}$ is turned into an original flux prediction by passing to a base model. From there, we have two pathways. In the first, the base model's flux predictions are passed through the additional head, which is another layer followed by softmax ensuring positivity and unit sum. The second pathway simply computes the total sum of the base model's prediction and multiplies this value with the output of the first pathway to obtain the final prediction $\vec{y}$.

for both: similarity to the expensive radiative transfer parameterization present in climate models and accounting for external physical constraints the said parametrization itself does not satisfy. Given the nature of the ClimART training pipeline, where we are looking at single columns at a time, we will explore loss functions and constraint formulated on the level of columns. For the sake of simplicity and reducing computational resources needed, we devise our framework such that pretrained models on the original loss, if present, can easily be plugged in by incorporating transfer learning. The principal workflow consists of two training stages.

First, an ensemble of baseline models are trained optimizing for similarity to the ClimaART target data in terms of predicting up- and downward radiative fluxes. In contrast to the previous work done in [7], we already make use of our first hard constraint in this first training stage, namely rescaling the predictions according to the incoming TOA solar flux and making sure this value matches the targets. After sufficient convergence, the best model of the group will be evaluated and selected for further fine-tuning.

In the second stage, we switch out the loss with an error function for a given constraint completely, or alternatively, with a combination of the original loss and the constraint loss (see the Section 4.1.2 and 4.2.3). Thus, we employ a mixture of transfer learning and soft constraining. In general, this second stage should be much shorter. Additionally, the models' architecture gets modified such that the predictions may change, but energy conservation and positivity of flux predictions are necessarily given by hand constraints (see Section 4.3.2).

See Figure 8 for a structured overview of the training stages, and the ensemble of biases and constraints applied per stage.

After training, the final models are evaluated on the tradeoff between their two optimization goals

1st Stage: **Pretraining**

- base loss
- *hard constraint* → flux rescaling

2nd Stage: **Transfer Learning**

- *physical bias* → transferring of weights (implicit energy conservation, similarity to RT target parameterization)
- *soft constraint* → constrained based loss (integral balance constraint loss / combination of base and integral balance loss)
- *hard constraint* → flux rescaling + positivity constraint and energy conservation constraint in Two-Head architecture

Figure 8: Overview of bias and constraints applied per training stage.

and their physical plausibility based on data that has not been present in either of the training stages.

4.5 Evaluation Metrics

4.5.1 Flux Target Similarity

Quantitative: Following [7], we report the root-mean-squared-error (RMSE) and the mean-bias-error (MBE) of the difference between the models' flux predictions and the target parameterization. We average these scores over levels, columns, snaps and years per training test split and report them for up- and downwelling fluxes and wavelength type separately.

Qualitative: Additional to mean statistics over snaps, columns and levels, we create height error plots, to show where models deviate the most from the target prediction in relation to height (level / layer) in the atmosphere.

4.5.2 Balance Constraint

Quantitative: As the ground truth balance error should be as close to zero as possible, we report a root-squared value for the balance error averaged over columns, snaps and years per test set, respectively. The balance error summarizes up- and downwelling fluxes of all layers, but treats wavelengths separately.

Qualitative: As the balance error can be determined-column wise, we assess a models' ability to fullfill the balance constraint by looking at heat maps of a model's column-wise prediction for a single snapshot. Due to the uneven distribution of area per column located at each grid cell, which in turn is due to the spherical shape of the earth in combination of the gridding method used in the data [7], we report an area weighted column-wise balance error per column to assess its qualitative significance. The weighting is retrieved by multiplying each error per column by the columns' respective covered area, dividing by the total land area present. A more thorough explanation on weighting by column contribution can be found in Appendix section 8.1.

4.5.3 Energy Conservation

As an additional check-up, as the energy conservation constraint is not explicit in the training procedure, we create several heating maps showing the violation of energy conservation by the models' flux predictions similar to the balance constraint heating maps.

4.6 Experimental Setup

As the availability of pre-trained baseline models from the authors of the ClimART paper was limited to running a few tests, retraining was needed. However, training fully fledged models that yield compatible results to those reported in [7] can take up to several days of training time and needs several hundred gigabytes of data. As the goal of this work is not to develop full-performance models but to test the effect of physically motivated training on both performance and training costs, we use a data-bagging concept [55] to reduce computational efforts.

In data-bagging, we instantiate an ensemble of the same models that are only trained on a random subset of the full dataset. We choose an ensemble size of 3 for each model design, including each base model and each transfer version based on either of the base models, trained with either only the constraint loss or a combination, thus resulting in $3 \times 2 + \ 3 \times 2 \times 2 = 18$ models in total. For the training split, each ensemble member is trained on a random 60% split of the main training data set, while validation and test split remain unchanged. For the base MLP used in this work, we stick to the ClimART baseline, meaning a simple three-layer MLP with hidden dimensions $< 512, 256, 256 >$ with layer normalization. For the Graph Net base model, we stick to the ClimART implementation of a three-block network, modelling the update function of each component, node, edge and global features with a one-layer MLP of size 128 plus a readout MLP after the blocks that takes the mean-pooled node information of the last block as input. For the Two-Head models, we use a full-connected randomly initialized layer with a zero-bias initialization as read-out layer.

All input data is normalized using a z-normalization based on pre-computed statistics (see Section 3.5). All models are trained on predicting shortwave fluxes only, whereas losses based on up- and downwelling flux predictions contribute equally to the overall loss, meaning a contribution weight of 0.5, whereas there is no heating rate loss contribution. For the combined loss approach, flux and constraint loss contribute equally to the overall loss.

In contrast to [7], where all models train for 100 epochs, we shorten the training time to a maximum of 65 epochs for the base models as we work with a reduced size in training data. For Two-Head models trained on the constraint loss only, we shorten this further to 25 epochs, while we extend it to 35 epochs for combined loss Two-Head models as they face more difficult convergence requirements. For all other training hyperparameters, we stick to the previous work: We use early-stopping after 12 epochs and clip the l2 gradient norm at 1. For the optimizer, we choose Adam with a learning rate of 2e-4 and an exponential decay learning scheduler with $\gamma = 0.98$ and a minimum learning rate of 1e-6. These values were set throughout all experiments. For training, the batch size is set to 128, for testing, a batch stretches over all columns of a snap, thus is set to 8192. All models are trained on an NVIDIA Quadro RTX 8000.

The codebase is a framework based on PyTorch Lightning and Weights&Biases and adapted from ClimART, which is available on GitHub.

Model	test/flux score/rmse/mean	test/flux score/mbe/mean
Base MLP	1.984 ± 0.331	0.042 ± 0.325
ClimART MLP	0.701 ± 0.04	-0.16 ± 0.07
Base GNN	0.764 ± 0.038	0.063 ± 0.087
ClimART GNN	0.648 ± 0.04	-0.142 ± 0.03

Table 1: **Average Test Flux Score:** Scores are averaged over all ensemble members, test years, up- and downwelling fluxes, vertical levels and respective errors (rmse and mbe). We show this computed mean score and its standard deviation. Note that the scores for the base models are computed on shortwave only, whereas the ClimART scores are averaged over both wavelengths.

5 Results

We report training convergence and generalization ability for flux error and balance constraint-based performance statistics for all models, as well as a qualitative assessment of the flux predictions.

5.1 Quantitative

5.1.1 Base Model Performance

We can report that the base models trained with data bagging, although trained with fewer epochs and only parts of the training data available, perform similarly to the reported baselines in ClimART in terms of flux prediction statistics, with the GNN outperforming the MLP (see Table 1). All models performed worse as the test years moved further away from the training years, which is also consistent with the results reported in ClimART. It shows that the GNN architecture outperforms the MLP architecture in terms of faster convergence and more stable performance. Both base models show good convergence in terms of training loss, with the GNN showing faster and more stable convergence than the MLP. For a detailed performance report, see Appendix section 8.3.1 for an analysis of training convergence and Section 8.3.1 a detailed report on all testing statistics of the base models.

5.1.2 Balance - Flux Tradeoff

Looking at training convergence, results show that all Two-Head models plateau quickly in terms of the evaluation metric. While models trained on a combined loss show difficulty in improving the constraint-based evaluation metric, purely constraint-based Two-Head models terminate their training before the maximum number of epochs allowed. See Appendix section 8.3.1 for a detailed report on training convergence of Two-Head and transfer models.

In Figure 9 we visualize the tradeoff between balance constraint satisfaction and diverging from the target parameterization for the flux predictions.

In terms of flux statistics, the purely constraint-based Two-Head models tend to perform worse than their base model but show less variation over the different test sets. The combined Two-Head models perform equally or even slightly better for far-future scenarios. In terms of balance constraint satisfaction, all Two-Head models outperform their respective base models, whereas purely constraint-based models perform the best. However, combined Two-Head models still significantly reduce the balance error as compared to the base models, which is especially evident for the combined Two-Head GNN.

Overall, the results show that the Two-Head models have the best tradeoff performance in terms of balancing the flux and balance error RMSE. Whether purely constraint-based or combined Two-Head models are deemed to be the best is subject to the question of weighing the two errors against each other. We suggest that the combined models deliver the best performance in terms of staying close

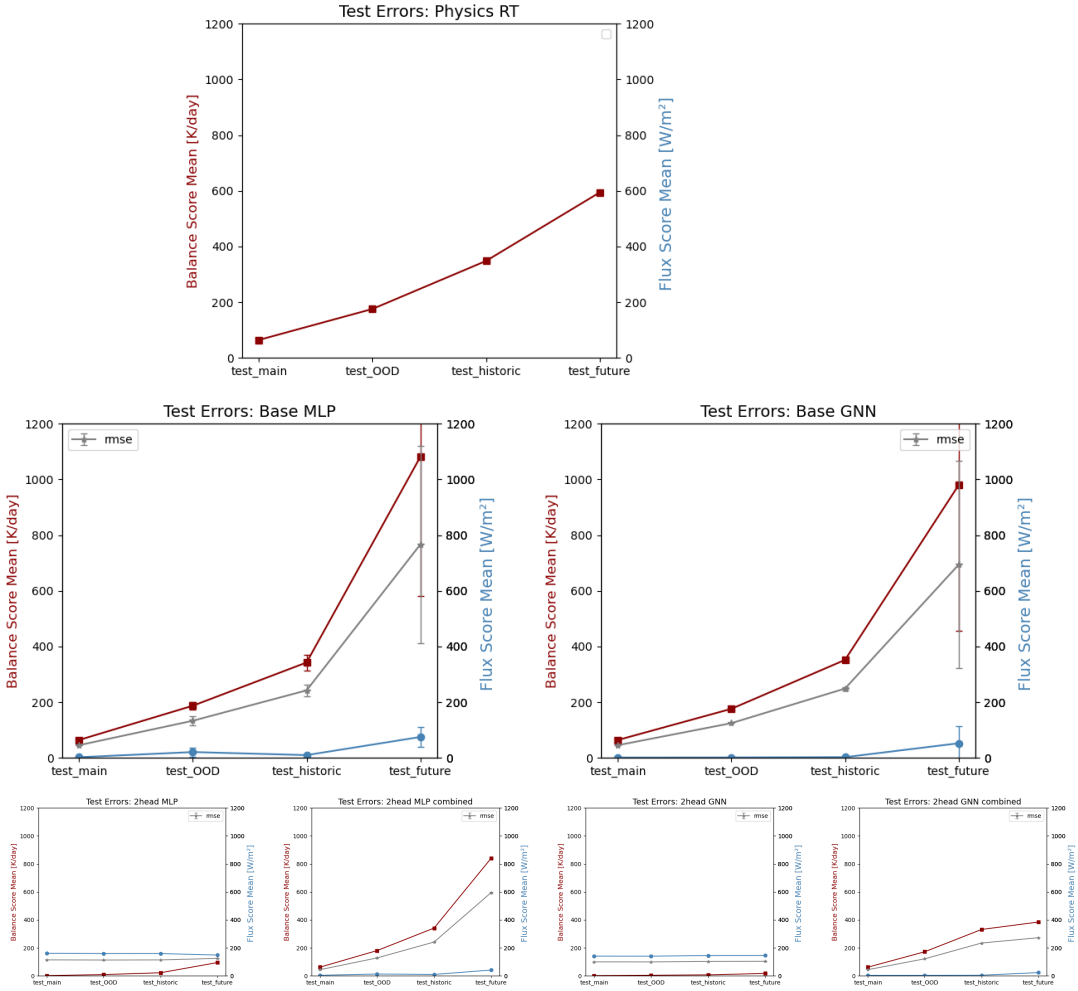


Figure 9: **Balance-Flux Error Tradeoff:** Each plot shows the tradeoff between the balance error RMSE (red, y-axis left) and the flux error RMSE (blue, y-axis right, averaged over the error for up- and downwelling shortwave fluxes) averaged over each test-set split (x-axis). The root-mean-squared average between the two errors is shown in gray. First row: Errors in the Target RT Parameterization (note that flux errors are not considered, as this is our 'ground truth' for flux predictions). Second row: MLP and GNN Base Models. Third row: Two Head Models, one per base model and transfer-loss (pure balance loss vs combined loss.

to the target RT while still reducing the external constraint violation and being the most stable to generalize to unseen and out-of-distribution conditions.

For a more detailed report on all test metrics results and scores, see Appendix section 8.3.1 and 8.3.1.

5.2 Qualitative

5.2.1 Balance Constraint Violation Heat Maps

To assess the qualitative results of the models, we focus on balance error heat maps. The heat maps show the violation of the balance constraint by the target parameterization and models, weighted by

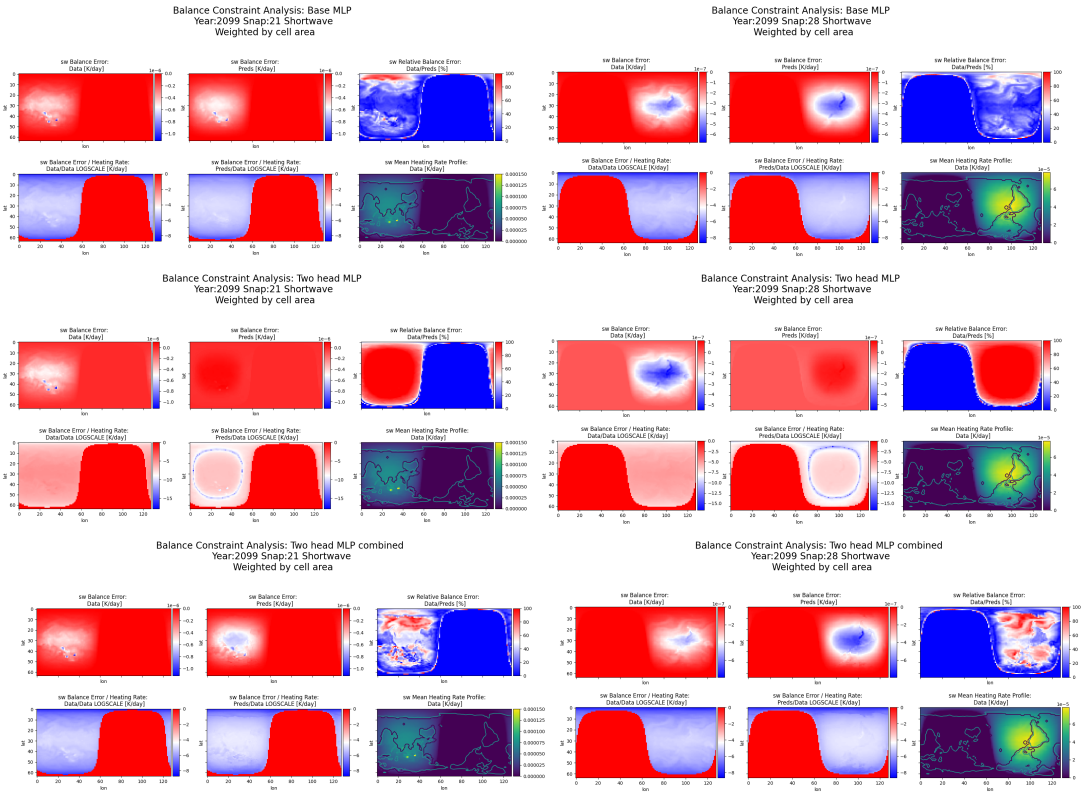


Figure 10: **Selection of MLP-based Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents the same two snaps of a year belonging to the far-future test set for a different model. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by the cell area covered by that column.

the cell area coverage of the respective column. The balance error is also plotted in relation to the target vertical average heating rates, to explore the significance of the error and where local adjustments are done by fine-tuned models.

Exemplary, we show heating maps for a year taken from the far-in-the-future projection test set, as patterns become the most observable there, for a selection of MLP-based models (Figure 10) and GNN-based models (Figure 11). For exemplary plots for each test set and model, please see Appendix section 8.3.2.1.

It shows, that base models learn to imitate the constraint violation patterns of the target parameterization very well for years closer to the training years. Moving away from training years in time, exhibited patterns start to diverge from the targets but still show similar absolute violation of the constraint.

All Two-Head models make local adjustments when asked to learn how to satisfy the balance constraint. Fore Two-Head models trained with a purely constraint-based loss, the error vanishes almost

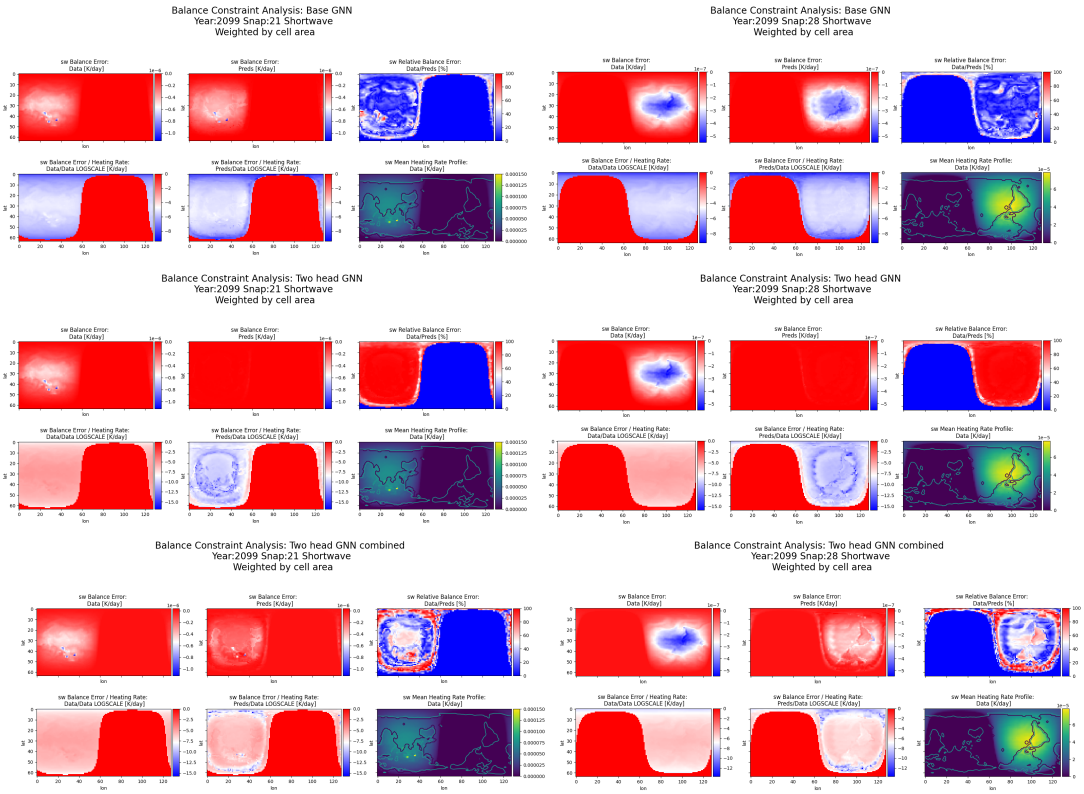


Figure 11: **Selection of GNN-based Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents the same two snaps of a year belonging to the far-future test set for a different model. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by the cell area covered by that column.

completely and, if still present, tend to cluster around the very middle and the very edge of the sun's incoming flux imprint. The Two-Head models trained with a combined loss show mixed adjustments dependent on the choice of baseline architecture. For MLP-based models, the violation of the error is either not altered or even greater in value and broader in patterns. GNN-based combined Two-Head models, however, show definite improvement of the error, even though not as much as their purely constraint-based counterparts. With them, the locality of where violation still remains becomes even clearer, again clustering in the centre and the edges of incoming solar radiation.

All in all, this qualitative assessment suggests that tuning the Two-Head models to diverge from the targets in order to satisfy the balance constraint was successful.

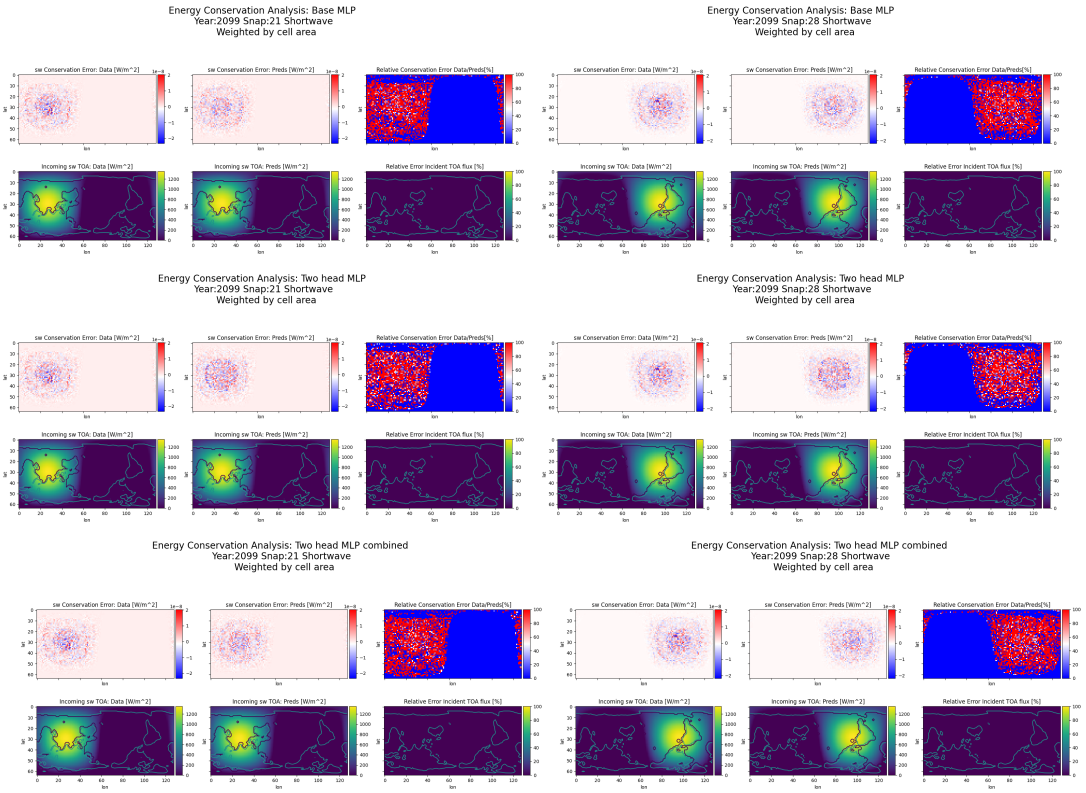


Figure 12: **Selection of MLP-based Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of conservation of energy as exhibited in the target parameterization versus a trained model. Each row represents the same two snaps of a year belonging to the far-future test set for a different model. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

5.2.2 Energy Conservation Violation Heat Maps

To assess the physicality of resulting flux predictions and the success of hard constraining the Two-Head, we show heat maps to test whether Two-Head models were able to maintain their energy-conserving properties. The heat maps show the violation of the energy conservation constraint by the target parameterization and models, weighted by the cell area coverage of the respective column. We also plot the incoming solar flux at TOA for both, targets and models, to assess if the flux-rescaling issue is solved. For both aspects, we also show relative errors between models and targets.

We show the heating maps for the same selection of snaps and years of the far-in-the-future projection test set that we used for the balance constraint heatmaps, for a selection of MLP-based models (Figure 12) and GNN-based models (Figure 13). For more exemplary plots for each test set and model, please see Appendix section 8.3.2.1.

Base models imitate the target parameterization behaviour, whereas absolute error values are low. Given that these years are the test years furthest away from training years in distribution, this suggests that our base models were able to abstract conservation of energy as a basic property from looking at

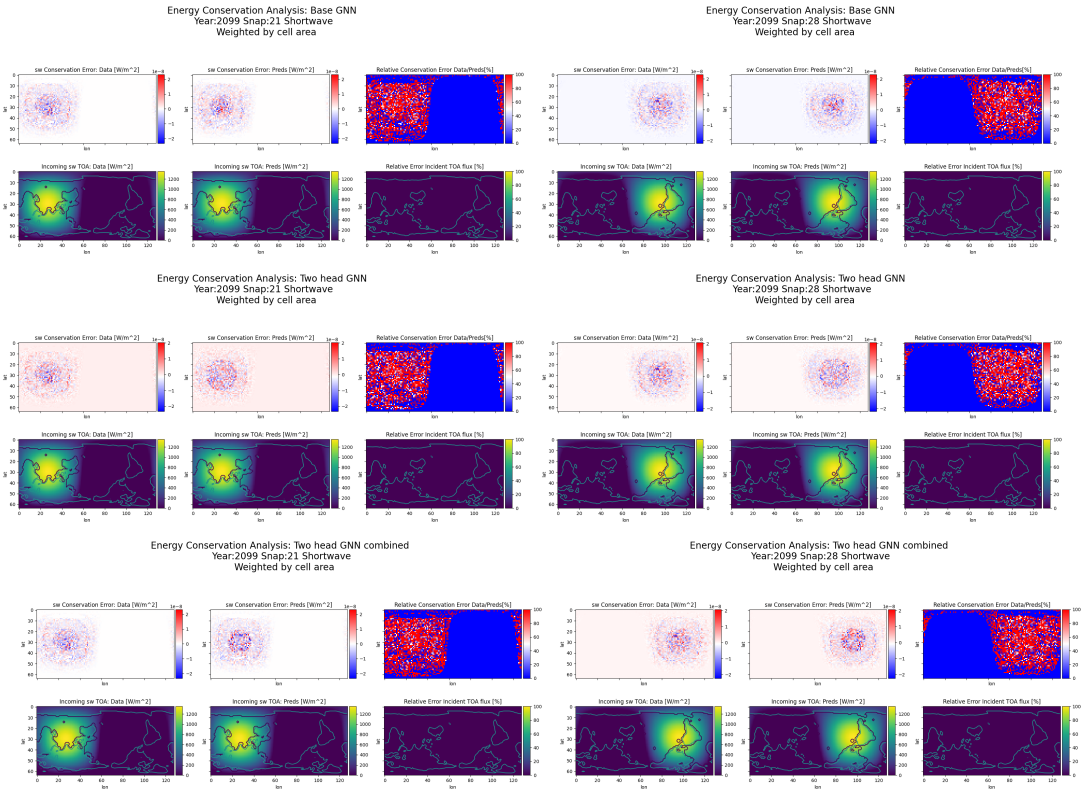


Figure 13: **Selection of GNN-based Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of conservation of energy as exhibited in the target parameterization versus a trained model. Each row represents the same two snaps of a year belonging to the far-future test set for a different model. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by the cell area covered by that column.

the target parameterization, which has conservation energy encoded in its process.

All 2head models, independent of the choice of loss function, show slight alterations in patterns, but no significant increase in absolute values of energy conservation errors. Thus, Two-Head models equally satisfy energy conservation in comparison to their base models and the target parameterization.

All models show no deviation in incoming shortwave flux at TOA.

All in all, this qualitative assessment suggests that the energy-conserving and fax-rescaling hard constraint introduced into the Two-Head models was successful in constricting the output space.

5.2.3 Height Errors

To show where Two-Head models make adjustments to the flux predictions, we show mean deviations from the targets, averaged over columns, years and snaps by height in the atmosphere.

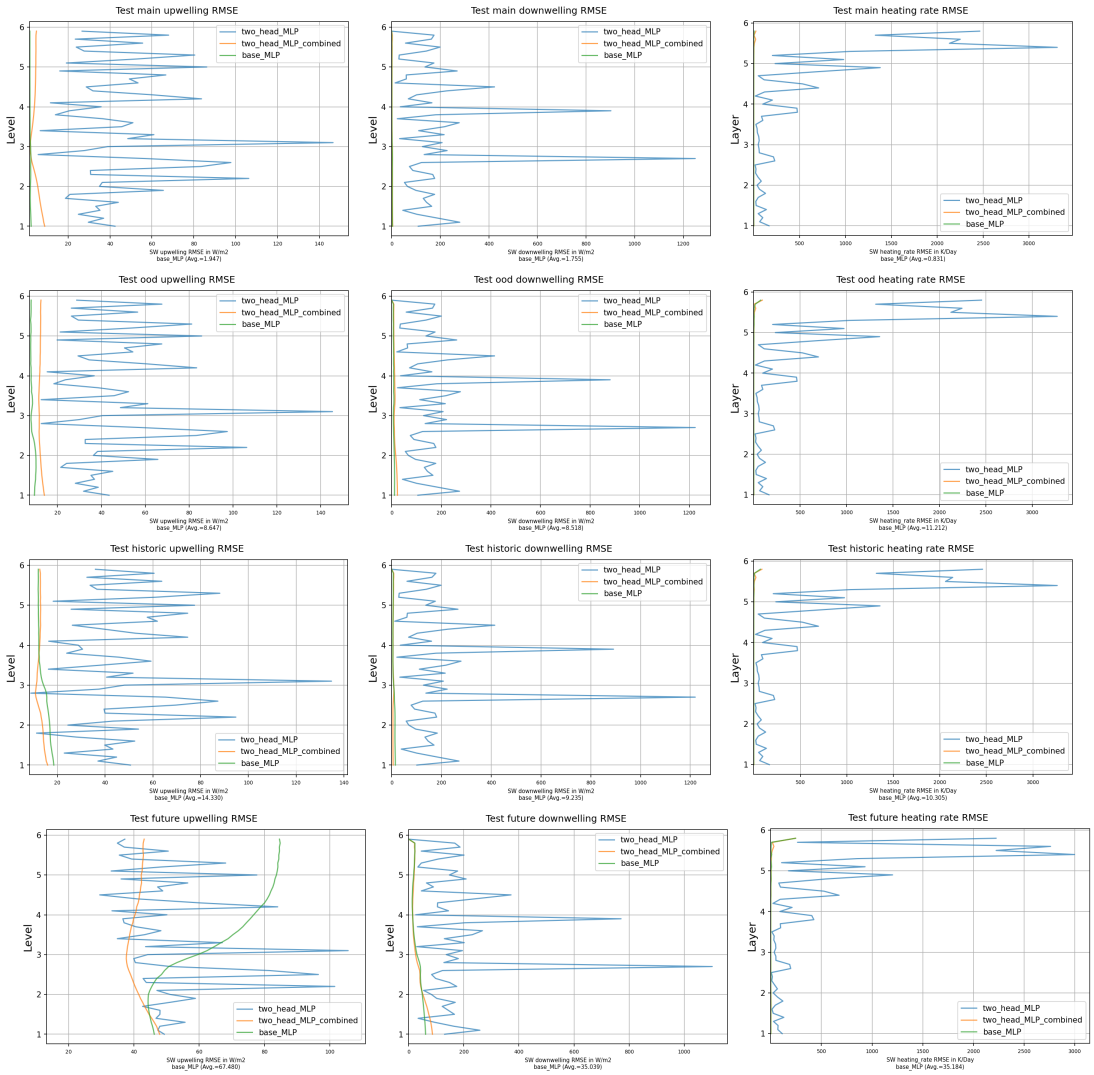


Figure 14: **Height Errors MLP-based models:** This plot shows the difference between model flux predictions and target RT (RMSE) over height in the atmosphere. The lines show errors averaged over years belonging to the main test set, the OOD test set, the historic and the future test set respectively. The first column shows errors of the upwelling flux predictions (y-axis equals levels), the second errors for downwelling flux predictions (y-axis equals levels) and the third shows the difference in resulting heating rates (y-axis equals atmospheric layers).

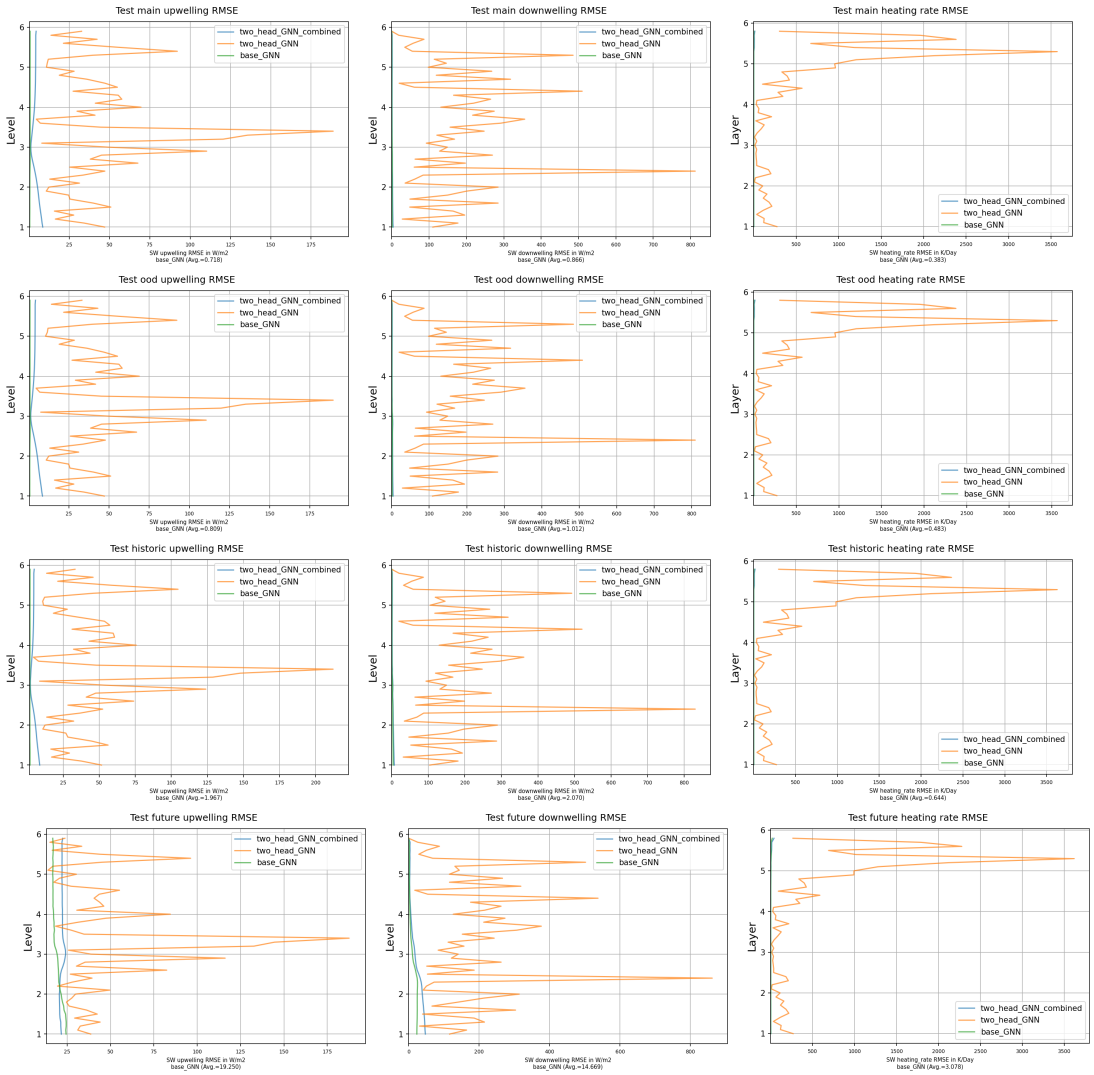


Figure 15: **Height Errors GNN-based models:** This plot shows the difference between model flux predictions and target RT (RMSE) over height in the atmosphere. The lines show errors averaged over years belonging to the main test set, the OOD test set, and the historic and the future test set respectively. The first column shows errors for the upwelling flux predictions (y-axis equals levels), the second errors for downwelling flux predictions (y-axis equals levels) and the third shows the difference in resulting heating rates (y-axis equals atmospheric layers).

On the resulting height error plots (Figures 45 (MLP-based models), and 46 (GNN-based models) it is visible that adjustments are relatively consistent among every test set (main, historic, OOD and future) and are the most distinct for Two-Head models trained on the purely constraint-based loss. Up- and downwelling fluxes show the largest deviation around the mid-layer (level) area, mostly in the form of two major "spikes". The largest deviations in heating rates occur near the top of the atmosphere.

As these deviations are consistent, independent of the base model architecture (MLP vs. GNN), this suggests that they are more than just mere statistical artefacts.

6 Discussion

6.1 Results

Our findings demonstrate the efficacy of the external physical constraints we employed in enhancing the performance of our Two-Head models. These constraints led to improved accuracy when compared to target RT and base models, and resulted in systematic adjustments of flux predictions.

We observed that the Two-Head models trained with combined losses exhibited a remarkable ability to either match or surpass the baselines reported in ClimART [7], regarding both optimization goals: similarity to the target RT and minimizing external constraints. This suggests their enhanced effectiveness. Notably, our Two-Head models required significantly less training data than the ClimART baselines due to our data-bagging paradigm, and yet their total training time was approximately the same (100 epochs). This suggests that the introduction of biases and constraints improves efficiency of training.

Our Two-Head models demonstrated superior generalizability to out-of-distribution conditions, meaning more stability in their predictions. We suspect that this is due to the hard constraints restricting the output space of the models, simply allowing less variation.

Overall, GNN baselines and Two-Head models based on a GNN architecture show slightly better performance, indicating that the physical bias introduced through the architecture, exploiting the relationships between layers, leverages training efficiency.

Purely constraint-based Two-Head models show distinct and consistent alterations of the base models fluxes, showing great ability to satisfy the external constraint while maintaining energy conservation column-wise. Also looking at the vertical, they exhibit distinct emerging patterns, in the forms of spikes in deviations from the Base Models in always the same height proximity. However, the physical reasoning of the appearance of these patterns in response to the integral balance constraint is to this point not clear to us.

Despite all of that, it is important to note that these findings all rely on the available evaluation metrics at hand, which may be imperfect, due to the nature of the available data, and incomplete, due to the limited selection of explored constraints. Further investigation and comparison to observation are needed to validate the Two-Head model's superiority.

6.2 Limitations

Our experimental design has several limitations.

First, we did not conduct an exhaustive hyperparameter search despite having numerous free parameters, such as the loss contribution weights. This includes weighing the flux-based and balance-based terms in our combined loss, as well as up- and downwelling flux, and the contribution of heating rates to the flux-based loss. Furthermore, we did not vary the number of model parameters, the optimizers, learning rate schedules, or early stopping conditions. Exploring the model space through hyperparameter tuning could potentially lead to even better performance.

Second, we did not explore alternatives to using precomputed statistics to normalize the input data. For example, one could compute normalization statistics dynamically on each mini-batch or employ spatial normalization on both the full dataset and mini-batches. Spatial normalization in this case means computing the statistics separately for layers and levels. This separation is already available within the statistics data provided by the authors of ClimART. All these alternatives may leverage the expressiveness of the input space for models to learn.

Also, we did not study the effect of using data bagging on Base Model versus Two-Head model performance. Our baselines are trained on a one-time selection of 60% of the training data, whereas Two-Head models experience a switch in between, very likely leading to having seen more training

years in total than the baselines. Experiments should be conducted, allowing more random switches of training data, which should very likely increase performance while still keeping computational costs low.

Furthermore, we also did not explore any alternatives to the GNN design rather than using node-pooling and a read-out layer to make predictions, It could be interesting to exploit the graph structure directly, like through a direct node-regression technique, which may enhance performance but above all, would offer a more physically interpretable resulting model.

Additionally, our computational resources were limited, which hindered our ability to train our base models and Two-Head models on all available data. Although convergence trends suggest that more training time on a subset of the data could lead to improved performance, we acknowledge that this is a limitation of our study.

Lastly, we did not model both longwave and shortwave, which may impact the accuracy of our models.

External validation: The most significant limitation of our approach is the absence of true ground truth data to compare to, as the dataset targets are only an RT approximation themselves. Consequently, we have no means to assess whether the adjustments made by our Two-Head models were any closer to reality than the RT parameterization or our base models. However, we believe that including observational data could improve our model predictions. For instance, we could plot actual heating rate profiles over atmospheric layers of our models and compare them to idealized theoretical profiles or observational ones. It is important to note that the height errors shown in Section 5.2.3 do not depict heating rate profiles, but rather deviations in heating rates between the models and the target RT. There exists a study [56] that used theoretical and observational heating rate profile data to evaluate the performance of a range of RT approximations, with a focus on the effect of clouds. However, our current models are unable to model the effect of clouds, thereby forming another limitation of our approach.

6.3 Outlook and Further Research

One potential solution to the limitation of not modelling clouds is to extend our approach to account for the fact that columns need to be regarded as interdependent data points, thus having to be able to model inter-column interactions. This basically translates to moving away from a 2-stream approximation to at least a 4-stream (allowing the horizontal flow of fluxes) or even to a more general multi-stream approximation (allowing flow in all directions). Why this is the case becomes evident in Figure 16: Clouds in the atmosphere can block the vertical flow of radiation and may cause accumulation of heat, eventually leading fluxes to trespass columns. This will definitely be the case for columns of small width, but can happen even for columns that are wide in range or that are only partly filled by clouds and lead to potential domino effects in the immediate neighbourhood, or in extreme events may even evoke teleconnections, meaning linking phenomena in locations that are separated by several thousand of kilometres. All in all, modelling column-interdependency is crucial for incorporating clouds as well as for extending to different resolutions.

To achieve this, we propose interventions in the ClimART training pipeline and architectures, operating on full snaps rather than single columns. We suggest extending the ClimART GNN to allow edges between neighbouring levels in all directions, resulting in a star-shaped graph structure outgoing from each node. However, this full-scale graph would be computationally intensive and require large amounts of GPU memory to train. Alternatively, sub-graphs from immediate neighbourhoods could be sampled to train a scalable inception network, as shown in [57]. This approximator would be interpretable as a 3D radiative transfer model, trained with 3D constraints, making it more physically consistent and more flexible. Potentially, it would be able to a) infer flux predictions for specific regions from fluxes given for other regions and in general better explore the impact of locality and b) could be able to include observational data in training, even if these are only partially available.

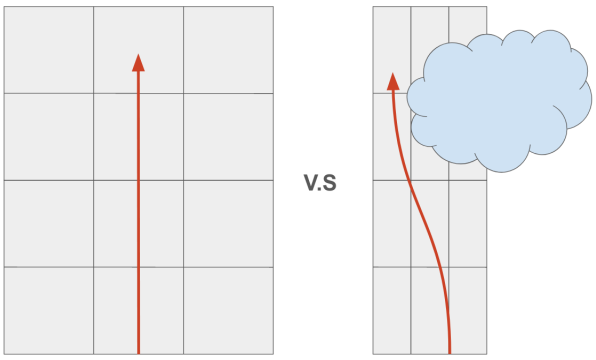


Figure 16: **The effect of clouds:** Clouds in the atmosphere can block the vertical flow of radiation and may cause accumulation of heat, eventually leading fluxes to trespass columns in the horizontal. In this case, one would need to move away from a 2-stream approximation to a 4-stream approximation to reasonably model RT.

The creation of such a 3D GNN poses several challenges, as the layer's heights in the ClimART dataset are of uneven distribution across columns (see Figure 24). Therefore, interpolation or extending to different gridding methods from the underlying climate model would be required to link fluxes for correct heights. Nonetheless, this approach offers more generalizability and could be leveraged to problem statements where the 2-stream approximation runs into problems, such as modelling clouds in RT and increasing model resolution. Thus, it presents a significant opportunity for conducting more research.

In addition, the resulting graph would allow for the assessment of local forms of energy conservation, such as radiative equilibrium, where the total net flux entering/exiting an atmospheric layer must equal the energy captured in its heating or cooling. This spatiotemporal constraint could be incorporated into the model, making it even more physically consistent. Once such a model is trained, it could be used to infer fluxes from partially available observational data, leading to a more flexible and efficient RT parameterization.

In conclusion, while this approach poses significant challenges, it presents a promising opportunity to extend our model's capabilities, allowing for the incorporation of clouds and increasing model resolution. By modelling column-interdependency, we can improve our understanding of the atmosphere and enable more accurate climate predictions.

7 Conclusion

In this study, we have conducted an analysis of the ClimART dataset and its baseline models, which revealed some shortcomings in their physical validity. By incorporating physically motivated soft and Hard Constraints into the training procedure and architecture, we have developed models that exhibit better out-of-distribution generalization and physical validity with respect to the explored constraints than the ClimART baselines. Moreover, the proposed models require fewer computational resources while performing similarly in the original evaluation metrics.

To further validate the physical correctness of our proposed models, we suggest using external observational data. Additionally, future research should explore ways to incorporate clouds, handle different resolutions, and provide better interpretability by moving away from the widely used 2-stream approximation.

Overall, this study provides evidence supporting the advancement of Theory-guided data science in producing more efficient ML models. These models could directly be incorporated into existing Climate Models, like the CanESM5, to replace computational bottlenecks with dependable results. Globally, the development of lightweight, precise, and physically reasonable models could overcome the challenges faced by climate models and unleash their full potential. With well-developed climate models at hand, researchers and policy-makers can better comprehend the effects of climate change, resulting in well-considered adaptation and mitigation strategies. Given the current global situation and our society's course towards addressing climate change, these strategies are evidently crucial.

References

[1] V. Masson-Delmotte, P. Zhai, A. Pirani, *et al.*, *IPCC 2018: Summary for Policymakers.* T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds. Cambridge, United Kingdom; New York, NY, USA: Cambridge University Press, 2018.

[2] ——, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, Eds. Cambridge, United Kingdom; New York, NY, USA: Cambridge University Press, 2021.

[3] G. Flato, J. Marotzke, B. Abiodun, *et al.*, "Evaluation of climate models. in: Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change," in Jan. 2013.

[4] R. McSweeney. "Q&a: How do climate models work?" Carbon Brief. (2018).

[5] D. Rolnick, P. L. Donti, L. H. Kaack, *et al.*, "Tackling climate change with machine learning," *arXiv:1906.05433 [cs, stat]*, 2019.

[6] C. Irrgang, N. Boers, M. Sonnewald, *et al.*, "Towards neural earth system modelling by integrating artificial intelligence in earth system science," *Nature Machine Intelligence*, vol. 3, no. 8, pp. 667–674, 2021. DOI: 10.1038/s42256-021-00374-3.

[7] S. R. Cachay, V. Ramesh, J. N. S. Cole, H. Barker, and D. Rolnick, "Climart: A benchmark dataset for emulating atmospheric radiative transfer in weather and climate models," 2021. DOI: 10.48550/ARXIV.2111.14671.

[8] S. Mishra and R. Molinaro, "Physics informed neural networks for simulating radiative transfer," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 270, p. 107 705, 2021. DOI: 10.1016/j.jqsrt.2021.107705.

[9] P. Harder, Q. Yang, V. Ramesh, *et al.*, "Generating physically-consistent high-resolution climate data with hard-constrained neural networks," arXiv, arXiv:2208.05424, 2022, type: article.

[10] J. S. Read, X. Jia, J. Willard, *et al.*, "Process-guided deep learning predictions of lake water temperature," *Water Resources Research*, vol. 55, no. 11, pp. 9173–9190, 2109. DOI: 10.1029/2019WR024922.

[11] A. Daw, A. Karpatne, W. Watkins, J. Read, and V. Kumar, "Physics-guided neural networks (pgnn): An application in lake temperature modeling," arXiv, 2021, type: article.

[12] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, "Enforcing analytic constraints in neural networks emulating physical systems," *Physical Review Letters*, vol. 126, no. 9, p. 098 302, 2021. DOI: 10.1103/PhysRevLett.126.098302.

[13] H.-O. Pörtner, D. C. Roberts, M. Tignor, *et al.*, *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, B. Okem, Ed. Cambridge, UK and New York, NY, USA: Cambridge University Press, vol. 2022.

[14] P. R. Shukla, J. Skea, R. Slade, *et al.*, *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY, USA: Cambridge University Press, vol. 2022.

[15] V. Eyring, S. Bony, G. A. Meehl, *et al.*, "Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016. DOI: 10.5194/gmd-9-1937-2016.

[16] R. H. Moss, J. A. Edmonds, K. A. Hibbard, *et al.*, "The next generation of scenarios for climate change research and assessment," *Nature*, vol. 463, no. 7282, pp. 747–756, Feb. 2010.

[17] B. C. O'Neill, E. Kriegler, K. Riahi, *et al.*, "A new scenario framework for climate change research: The concept of shared socioeconomic pathways," *Climatic Change*, vol. 122, no. 3, pp. 387–400, Feb. 2014.

[18] K. Riahi, D. P. van Vuuren, E. Kriegler, *et al.*, "The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview," *Global Environmental Change*, vol. 42, pp. 153–168, 2017. DOI: https://doi.org/10.1016/j.gloenvcha.2016.05.009.

[19] B. Sanderson, "The end of the wait for climate sensitivity?" *Geophysical Research Letters*, vol. 46, no. 21, pp. 12 289–12 292, 2019. DOI: https://doi.org/10.1029/2019GL084685.

[20] V. Balaji, E. Maisonnave, N. Zadeh, *et al.*, "Cpmip: Measurements of real computational performance of earth system models in cmip6," *Geoscientific Model Development*, vol. 10, no. 1, pp. 19–34, 2017. DOI: 10.5194/gmd-10-19-2017.

[21] B. C. O'Neill, C. Tebaldi, D. P. van Vuuren, *et al.*, "The scenario model intercomparison project (scenariomip) for cmip6," *Geoscientific Model Development*, vol. 9, no. 9, pp. 3461–3482, 2016. DOI: 10.5194/gmd-9-3461-2016.

[22] D. Watson-Parris, Y. Rao, D. Olivié, *et al.* "ClimateBench: A benchmark dataset for data-driven climate projections," Earth and Space Science Open Archive. Archive Location: world Publisher: Earth and Space Science Open Archive Section: Climatology (Global Change). (2021).

[23] K. Zhang, X. Liu, M. Wang, *et al.*, "Evaluating and constraining ice cloud parameterizations in cam5 using aircraft measurements from the sparticus campaign," *Atmospheric Chemistry and Physics*, vol. 13, no. 9, pp. 4963–4982, 2013. DOI: 10.5194/acp-13-4963-2013.

[24] A. Ramadhan, J. Marshall, A. Souza, G. L. Wagner, M. Ponnapati, and C. Rackauckas, "Capturing missing physics in climate model parameterizations using neural differential equations," *arXiv:2010.12559 [physics]*, 2020.

[25] K. Heng, *Exoplanetary Atmospheres: Theoretical Concepts and Foundations*. Princeton University Press, 2017. DOI: doi:10.1515/9781400883073.

[26] K. Stamnes, G. E. Thomas, and J. J. Stamnes, *Radiative Transfer in the Atmosphere and Ocean*, 2nd ed. Cambridge University Press, 2017. DOI: 10.1017/9781316148549.

[27] N. C. Swart, J. N. S. Cole, V. V. Kharin, *et al.*, "The canadian earth system model version 5 (canesm 5.0.3)," *Geoscientific Model Development*, vol. 12, no. 11, pp. 4823–4873, 2019. DOI: 10.5194/gmd-12-4823-2019.

[28] K. von Salzen, J. F. Scinocca, N. A. McFarlane, *et al.*, "The canadian fourth generation atmospheric global climate model (CanAM4). part i: Representation of physical processes," *Atmosphere-Ocean*, vol. 51, no. 1, pp. 104–125, 2013. DOI: 10.1080/07055900.2012.755610.

[29] H. Barker, J. Cole, J.-J. Morcrette, *et al.*, "The monte carlo independent column approximation: An assessment using several global atmospheric models," *Quarterly Journal of the Royal Meteorological Society*, vol. 134, pp. 1463–1478, Jul. 2008. DOI: 10.1002/qj.303.

[30] D. A. Randall, "Cloud parameterization for climate modeling: Status and prospects," *Atmospheric Research*, vol. 23, no. 3, pp. 345–361, 1989. DOI: https://doi.org/10.1016/0169-8095(89)90025-2.

[31] S. Salcedo-Sanz, P. Ghamisi, M. Piles, *et al.*, "Machine learning information fusion in earth observation: A comprehensive review of methods, applications and data sources," *CoRR*, vol. abs/2012.05795, 2020.

[32] M. Chantry, S. Hatfield, P. Dueben, I. Polichtchouk, and T. Palmer, "Machine learning emulation of gravity wave drag in numerical weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 13, no. 7, Jul. 2021. DOI: 10.1029/2021ms002477.

[33] A. Gettelman, D. J. Gagne, C.-C. Chen, *et al.*, "Machine learning the warm rain process," *Journal of Advances in Modeling Earth Systems*, vol. 13, no. 2, e2020MS002268, 2021, e2020MS002268 2020MS002268. DOI: https://doi.org/10.1029/2020MS002268.

[34] T. Bolton and L. Zanna, "Applications of deep learning to ocean data inference and subgrid parameterization," *Journal of Advances in Modeling Earth Systems*, vol. 11, no. 1, pp. 376–399, 2019. DOI: https://doi.org/10.1029/2018MS001472.

[35] S. Rasp, M. S. Pritchard, and P. Gentine, "Deep learning to represent subgrid processes in climate models," *Proceedings of the National Academy of Sciences*, vol. 115, no. 39, pp. 9684–9689, 2018. DOI: 10.1073/pnas.1810286115.

[36] A. Karpatne, G. Atluri, J. H. Faghmous, *et al.*, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, 2017. DOI: 10.1109/TKDE.2017.2720168.

[37] S. Cuomo, V. S. D. Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli, "Scientific machine learning through physics-informed neural networks: Where we are and what's next," *CoRR*, vol. abs/2201.05624, 2022.

[38] C. Rackauckas, Y. Ma, J. Martensen, *et al.*, *Universal differential equations for scientific machine learning*, 2020. DOI: 10.48550/ARXIV.2001.04385.

[39] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, Jun. 2021.

[40] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," *CoRR*, vol. abs/1806.07366, 2018.

[41] T. Beucler, M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine, "Enforcing analytic constraints in neural networks emulating physical systems," *Phys. Rev. Lett.*, vol. 126, p. 098 302, 9 Mar. 2021. DOI: 10.1103/PhysRevLett.126.098302.

[42] P. Harder, D. Watson-Parris, D. Strassel, N. R. Gauger, P. Stier, and J. Keuper, "Emulating aerosol microphysics with machine learning," *CoRR*, vol. abs/2109.10593, 2021.

[43] J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar, "Integrating scientific knowledge with machine learning for engineering and environmental systems," *arXiv:2003.04919 [physics, stat]*, 2021.

[44] P. Márquez-Neila, M. Salzmann, and P. Fua, "Imposing hard constraints on deep networks: Promises and limitations," arXiv, arXiv:1706.02025, 2017, type: article.

[45] M. P. McCormick, L. W. Thomason, and C. R. Trepte, "Atmospheric effects of the mt pinatubo eruption," *Nature*, vol. 373, no. 6513, pp. 399–404, Feb. 1995. DOI: 10.1038/373399a0.

[46] V. M. Krasnopolsky, M. S. Fox-Rabinovitz, Y. T. Hou, S. J. Lord, and A. A. Belochitski, "Accurate and fast neural network emulations of model radiation for the ncep coupled climate forecast system: Climate simulations and seasonal predictions," *Monthly Weather Review*, vol. 138, no. 5, pp. 1822–1842, 2010. DOI: https://doi.org/10.1175/2009MWR3149.1.

[47] A. G. Ivakhnenko and V. G. Lapa, *Cybernetic Predicting Devices*. CCM Information Corporation, 1965.

[48] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and Cooperation in Neural Nets*, S.-i. Amari and M. A. Arbib, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 1982, pp. 267–285, ISBN: 978-3-642-46466-9.

[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *CoRR*, vol. abs/1609.02907, 2016.

[50] P. W. Battaglia, J. B. Hamrick, V. Bapst, *et al.*, *Relational inductive biases, deep learning, and graph networks*, 2018. DOI: 10.48550/ARXIV.1806.01261.

[51] J. Zhou, G. Cui, S. Hu, *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020. DOI: https://doi.org/10.1016/j.aiopen.2021.01.001.

[52] K. Oono and T. Suzuki, "On asymptotic behaviors of graph cnns from dynamical systems perspective," *CoRR*, vol. abs/1905.10947, 2019.

[53] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, May 2016. DOI: 10.1186/s40537-016-0043-6.

[54] J. G. Virgin, C. G. Fletcher, J. N. S. Cole, K. von Salzen, and T. Mitovski, "Cloud feedbacks from canesm2 to canesm5.0 and their influence on climate sensitivity," *Geoscientific Model Development*, vol. 14, no. 9, pp. 5355–5372, 2021. DOI: 10.5194/gmd-14-5355-2021.

[55] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996. DOI: 10.1007/BF00058655.

[56] G. Cesana, D. E. Waliser, D. Henderson, T. S. L'Ecuyer, X. Jiang, and J.-L. F. Li, "The vertical structure of radiative heating rates: A multimodel evaluation using a-train satellite observations," *Journal of Climate*, vol. 32, no. 5, pp. 1573–1590, 2019. DOI: ttps://doi.org/10.1175/JCLI-D-17-0136.1.

[57] E. Rossi, F. Frasca, B. Chamberlain, D. Eynard, M. M. Bronstein, and F. Monti, "SIGN: scalable inception graph neural networks," *CoRR*, vol. abs/2004.11198, 2020.

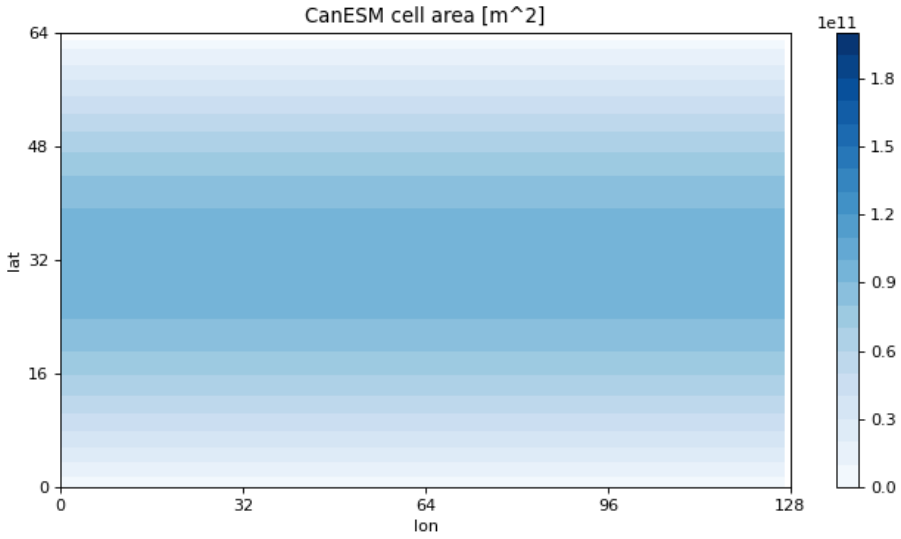


Figure 17: Cell area size over each grid cell at longitude and latitude coordinates. Due to the spherical shape of the earth and the Gaussian gridding method used in [7], grid cells appear to cover a bigger area around the equatorial belt (darker colours indicate larger area).

8 APPENDIX

8.1 Weighting by Column Contribution

Due to the spherical shape of the earth, most conventional gridding methods result in different-sized cells, whereas grid cells located near the equatorial belt are usually oversized and cells of the polar regions usually cover quite small areas. The ClimART data is mapped on two dimensions by a Gaussian gridding method, meaning that each grid point can be uniquely accessed by one-dimensional latitude and longitude arrays (i.e. the coordinates are orthogonal), whereas longitudes are equally spaced while the latitudes are unequally spaced according to the Gaussian quadrature. The resolution of ClimART is 3 degrees, resulting in grid cell areas ranging from around $300km^2$in the polar region up to $900km^2$ near the equator, as can be seen in Figure 17.
This uneven distributions of cell sizes in turn means that in any column-wise error computations, columns located near the equator contribute far more than columns located near the poles. To counterbalance this effect, we weigh the retrieved column-wise ERR_c error by the column's respective covered area size stored in $A \in \mathbb{R}^C$ with C being the total number of columns and divide by the total land area:

$$ERR_w = \frac{ERR_c \times A}{\sum_i^C A_i}$$

8.2 Constraint Analysis

As described in [12] [9] [10], terms expressing how well physical constraints are satisfied by a model's output can be added to the model's loss term while training. This is referred to as soft constraining. In that way, the model learns to minimize the error connected to a certain physical property that the output should maintain. Formally, we will write this loss constrain as follows:

$$\mathcal{L}(\vec{y}|\vec{x}) = \ sim(\vec{y}, \vec{t}) + \mathcal{C}(\vec{y})$$

Hereby, $\mathcal{L}$ is the loss of some prediction y given an input x and the target t using some similarity function sim and an instance of a constraint error function $\mathcal{C}(y)$.

In the following, we will explore several candidates of constraints that can be formulated in the context of radiative transfer and explore their significance and implementability in ClimART models.

8.2.1 Energy Conservation Constraint (ECC)

Considering that Radiative Transfer describes the transfer of energy, Energy Conservation seems like a natural choice to formulate a constraint.

8.2.1.1 Column-wise

Following the assumption, that atmospheric columns are independent of each other, as it is made in [7], the conservation of energy must be fulfilled by a single column. This means that the incoming energy at TOA must equal the energy absorbed by the atmosphere plus the energy absorbed by the surface plus the energy leaving the atmosphere at TOA.

One would expect to have to handle the energy conservation calculation differently for the two wavelength forms, shortwave and longwave, as there is no incoming longwave radiation at TOA. However, if one takes a wavelength of roughly 4 microns as the dividing line between shortwave and longwave radiation, there is about $12W/m^2$ of incoming solar radiation at wavelengths greater than 4 microns. The treatment of incoming radiation at the TOA for a longwave is particular to the radiative transfer model. Most radiative transfer models put this incoming radiation into their shortwave calculations, albeit in different manners, e.g., adding the additional incoming to the wavelength interval close to 4 microns. In the RT code underlying the parameterization data from ClimART however, the incoming solar radiation was explicitly accounted for in the longwave radiative transfer calculation. This is very specific to the exact climate mod and allows us to treat the two types in the same way in our analysis.

Thus, considering a single column c, N layers and $N+1$ levels with index 0 describing the top of the atmosphere (TOA) and N being the surface, up- $\vec{F_{\uparrow}}$ and downwelling $\vec{F_{\downarrow}}$ fluxes defined over levels, and the two wavelength-types given in the data $w \in sw, lw$, we can formulate this error mathematically as follows:

$$A^w_{atm} = ((F^{w,\downarrow}_{c_0} - F^{w,\uparrow}_{c_0}) - (F^{w,\downarrow}_{c_{N+1}} - F^{w,\uparrow}_{c_{N+1}}))$$

$$A^w_s = (F^{w,\downarrow}_{c_{N+1}} - F^{w,\uparrow}_{c_{N+1}})$$

$$ECC_c = \sum_{w \in \{sw,lw\}} F^{w,\downarrow}_{c_0} - A^w_{atm} - A^w_s - F^{w,\uparrow}_{c_0}$$

This error function can be calculated directly in each training step and added to the loss term.

8.2.1.2 Global

To retrieve an average of the global energy conservation mismatch, we can sum and average the mismatch over single columns given a number of total columns C:

$$ECC_{global} = \sum_{c=1}^{C} ECC_c$$

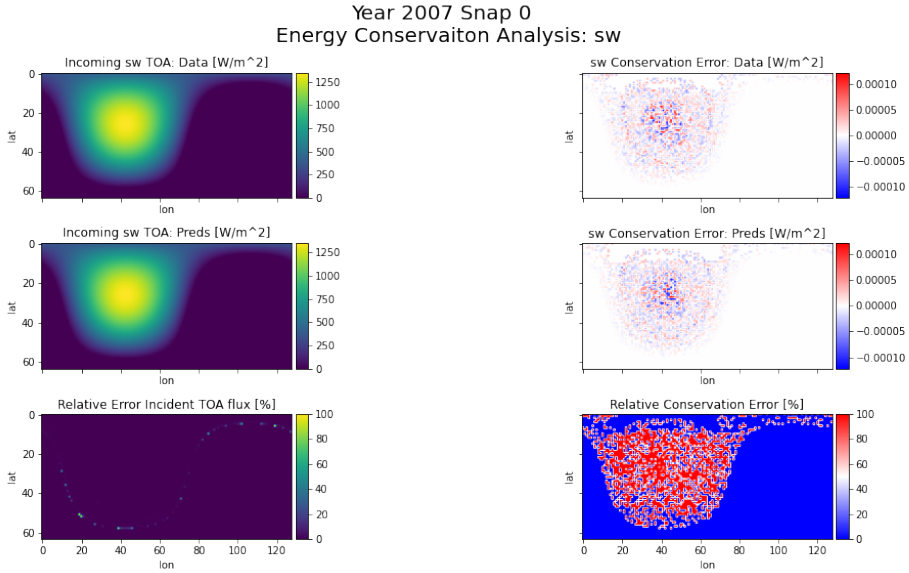


Figure 18: Energy Conservation Constraint Analysis: Comparison of the upholding of the energy conservation constraint for the test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only shortwave radiative fluxes are considered.

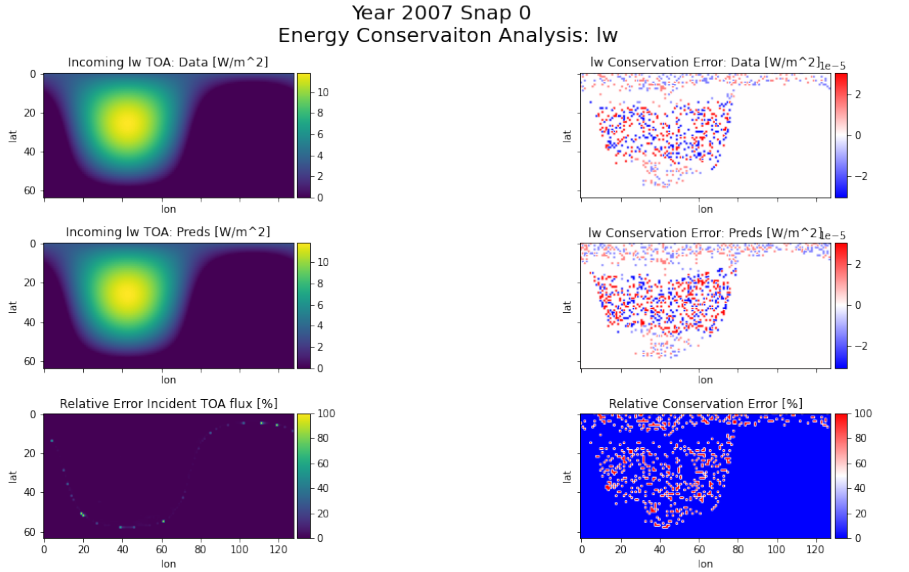


Figure 19: Energy Conservation Constraint Analysis: Comparison of the upholding of the energy conservation constraint for the test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only longwave radiative fluxes are considered.

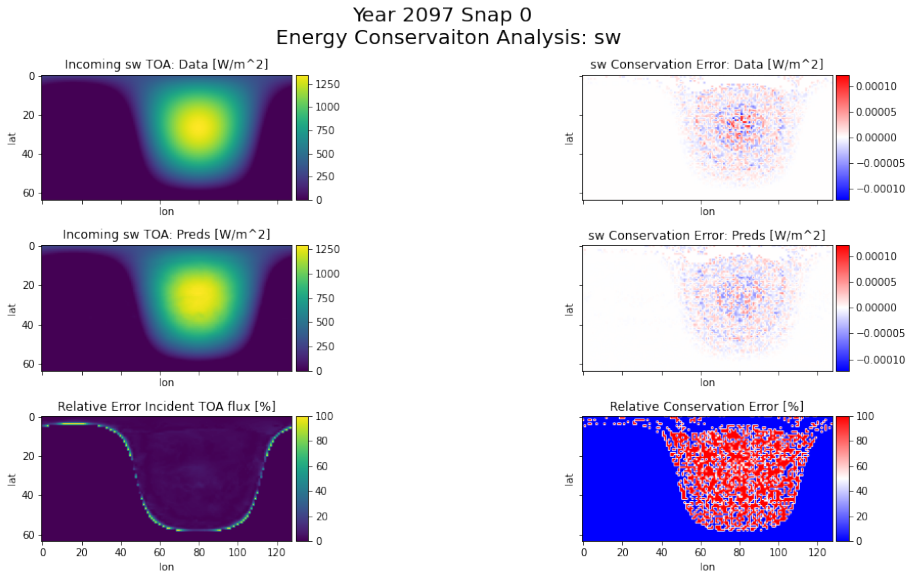


Figure 20: Energy Conservation Constraint Analysis: Comparison of the upholding of the energy conservation constraint for the out of distribution test year 2097 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only shortwave radiative fluxes are considered.

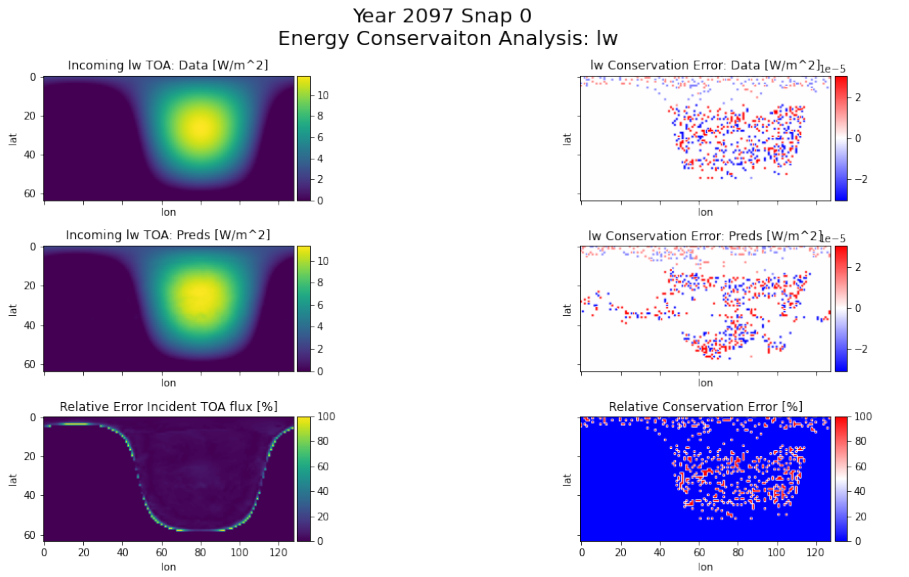


Figure 21: Energy Conservation Constraint Analysis: Comparison of the upholding of the energy conservation constraint for the out of distribution test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only longwave radiative fluxes are considered.

8.2.2 Analysis of Significance in ClimART

As examined by 18, energy conservation for shortwave radiation is accounted for in the original RT parameterization that makes up the data almost perfectly. A simple MLP trained on the full training dataset is able to mimic the original parameterization's behaviour in terms of energy conservation. Although the relative error between original parameterization data and model predictions seem big, the errors in itself are insignificantly big. The same holds for longwave radiation as can be seen in 19.

For years taken from the out of distribution years, energy conservation errors are still very small for shortwave radiation 20. Although similarly small for longwave radiation, a slight shift in distribution of the error can be observed and the simple model seems to exhibit some artefacts for this case 21.

All in all, though present, trying to minimize conservation constraint does not look like a promising way to impact the performance and generalization ability of the machine learning parameterization model. Possibilities to accelerate the training procedure overall using this constraint remain open, for example by directly enforcing it rather than having the model learn it on its own, which risks errors. In the following work the energy conservation constraint will be implicitly considered in the desgin of transfer models (see Section 4.3.2).

8.2.3 Energy Budget Constraint (EBC)

EEach climate represents the energy exchange between the Earth's climate system and space and how that affects the surface and atmospheric temperatures. Therewith, it holds that the energy entering the atmosphere and the energy that the Earth emits back into space must balance each other. This equilibrium state is known as the Earth's Energy Budget and reaching it drives interactions within the Earth's system, a process which radiative transfer describes.

For the incoming atmosphere, the Earth's internal heat can be neglected as the fracture is insignificant compared to the incoming energy from the sun at the top of the atmosphere (TOA) expressed by incoming (downwelling) shortwave fluxes. Incidentally, around $\cong 340 Wm^2$ of the solar radiation arrives at the Earth's TOA. Some of this incidental flux gets reflected by clouds, aerosols in the atmosphere and the surface albedo. We refer to the remaining absorbed fraction as A. The resulting incoming flux is affected by regional and seasonal variations, thus we will continue referring to the incoming flux as yearly averages.

Considering radiation reflected by clouds, atmosphere and the surface albedo, the absorbed solar radiation (ASR_{TOA}) moves in the range of $\cong 240 Wm^{-2}$ on a yearly average. Considering reflection by clouds only, the yearly average is approximate to $\cong 290 Wm^{-2}$
We define the outgoing energy as the outgoing longwave radiation (OLR_{TOA}, expressed by upwelling longwave fluxes at the TOA.

Thus, radiative transfer calculations must fulfil the Earth's energy budget constraint formulated as follows:

$$ASR_{TOA} \cong (1 - A) \times 340 Wm^{-2} = OLR_{TOA} \tag{10}$$

As the Earth's Energy Budget is an important driver in the climate system and can be computed directly from the information given by the RT predictions, it forms a second obvious choice to formulate a constraint to incorporate in training models emulating RT parameterizations.

However, such a budget only makes sense on a global scale and also is time sensitive. The considered time window should at least be spread over a full year. Thus, considering a number of snaps S and all atmospheric columns C, we can compute a global-spatial and temporal average of the Energy Budget Constraint as follows:

$$EBC_{temp,global} = |\frac{\sum_{s=0}^{S} \sum_{c=1}^{C} F_{sw_{c_0}}^{\downarrow} - F_{lw_{c_0}}^{\uparrow}}{S}| \tag{11}$$

8.2.3.1 Analysis of Significance in ClimART

We analyse the energy on a sample of the main test set and a sample of the OOD test set, taking the year 2007 and 2097, respectively. We plot temporal averages over the global grid and global averages over time for both the target RT parameterization (taken from the ClimART dataset) and the predictions done by a simple MLP baseline model provided by the authors of [7]. In Figure 22, we can see that the MLP is able to mimic the target behaviour surprisingly well. Although the error sumps up to around $72W/m^2$, we have to consider that there are a lot of factors neglected by this very simplified version of the energy budget calculation, such as reflection of aerosols and clouds, hence we conclude that the stability of this error is more important than its actual value.
For the OOD year 2097 shown in Figure 23, we can note more fluctuation on the temporal scale as well as a significant shift in the overall error from 70 to around $50W/m^2$ between target and model behaviour, all the exhibited pattern look the same.

In conclusion, it could be interesting to factor in energy balance into the training procedure, however, the temporal and spatial dynamics make it difficult, as the ClimART pipeline so far operates column-wise. Also, there are many uncertainties linked to this formulation of the energy balance constraint, such as clarity of what factors (aerosols, clouds...) are considered by the target data and whether a stable error given significant shifts in earth climatem, as it is expected to happen in the following decade due to ongoing climate change, is something physically valid or not.

In the remainder of this work, we will thus omit the energy budget constraint.

8.2.4 Radiative Layer Equilibrium

The relationship between heating rates and fluxes associated with each layer must obey the following principle: In the absence of any atmospheric dynamics, a radiative layer equilibrium should occur, meaning that the total net flux entering/exiting an atmospheric layer must equal the energy captured in its heating or cooling [25].

One can assess this local form of energy conservation by driving models with no atmospheric dynamics until layer temperatures cease to evolve. However, calculating the violation of this equilibrium by models trained with the current ClimART pipeline poses several problems: Models are trained column-wise, meaning they do not model any horizontal relationships between columns. While GNNs in theory provide the opportunity to account for the full integrity of the data structure, further investigation of the data shows an unequal distribution of layer heights per column present in a global snapshot (Figure **??**). This is due to the gridding method used for the vertical dimension of the atmosphere, which divides the total distance of the atmosphere between the surface and TOA into n layers irrespective of the actual height w.r.t sea level of the surface. This results in layers of different heights and thicknesses across columns. Thus, building a GNN allowing passage of information between points of the same height across columns would require further interpolation of the input data to provide a meaningful physical picture.

In summary, testing for radiative layer equilibrium in the absence of atmospheric dynamics could be a very interesting physical constraint to test for, especially when moving away from a 2-stream approximation to modelling horizontal flows of energy. Nevertheless, we will disregard it in the remainder of this work for the sake of simplicity.

8.2.5 Integral Balance Constranit (IBC)

Another constraint that can be viewed columns-wise, but also on more global scales, is balancing the relationship between heating rates, pressures and fluxes. The two outputs from the radiative

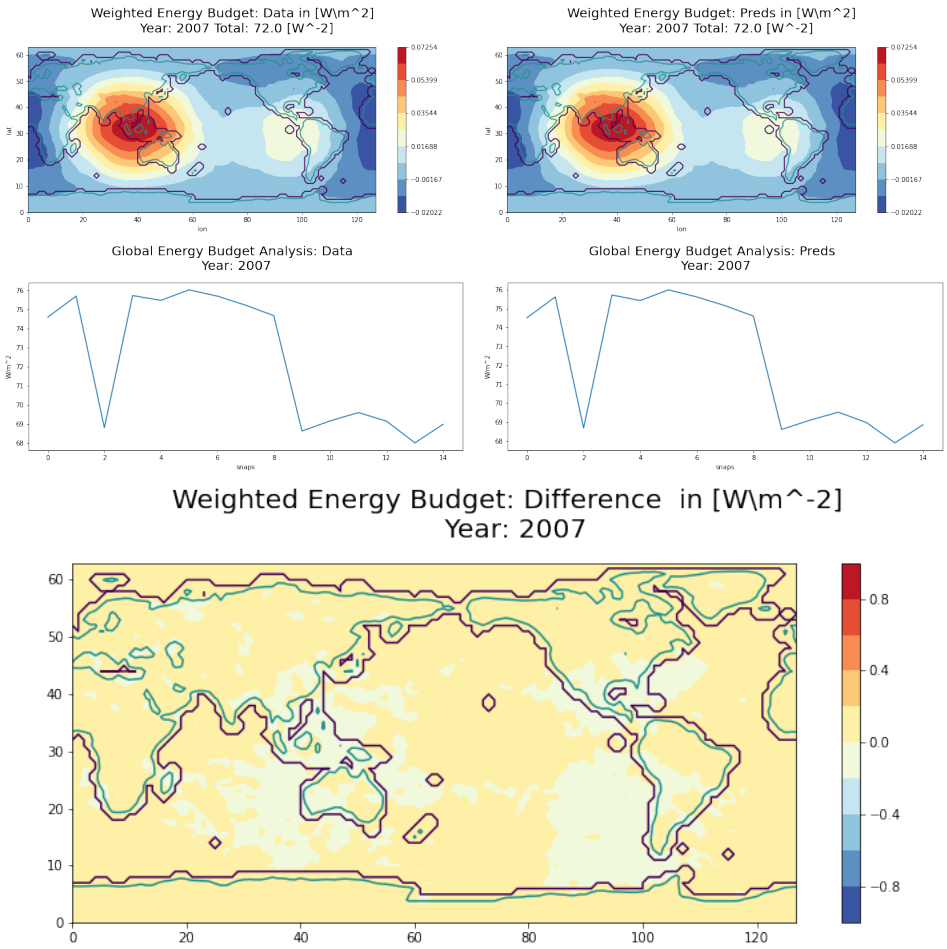


Figure 22: **Energy Budget Constraint** *EBC* **test year 2007**: On left side of the first two rows, we show an analysis of the data of the target RT parameterization in ClimART, whereas on right side we show an analysis of the predictions done by a simple MLP ClimART baseline model. The first line shows the error over the grid, whereas the actual error is weight ed by the area each grid cell covers. Below, we plot the difference of data versus model predictions weighted by cell area over the grid.

transfer parameterization are thus not completely independent of each other, as there exists an integral relationship between them, as expressed by equation 12. Here $k = 1..N$ stand for the vertical levels, with 0 being TOA and N being the surface level. $\alpha_k = (p_k - p_{k-1})G^{-1}$ describes the pressure difference between levels with the constant $G = (86.400 * 9.81/1005.5)$ set to seconds per day times the gravitational constant divided by the specific heat constant. G is necessary to make the units consistent with each other. h_k stand for the heating rate at a given level k. Φ 13 is normally computed differently for the two cases, short wave and long wave radiation, as there is no natural incident longwave radiation at the top of the atmosphere. However, as mentioned before in the energy conservation formulation the underlying radiative transfer parameterization in the CanESM5 uses a trick to account for solar radiation with a wavelength slightly below the dividing line between shortwave and longwave radiation, simply factoring it in as incoming radiation in the longwave radiative transfer calculations. Hence, in our case, Φ is actually equal to the SWR case for both types of radiation.

In ClimART, the predicted heating rates are directly computed from the fluxes, as discussed in 3.6. As this computation is fixed, such a balance constraint would not directly improve the model's predic-

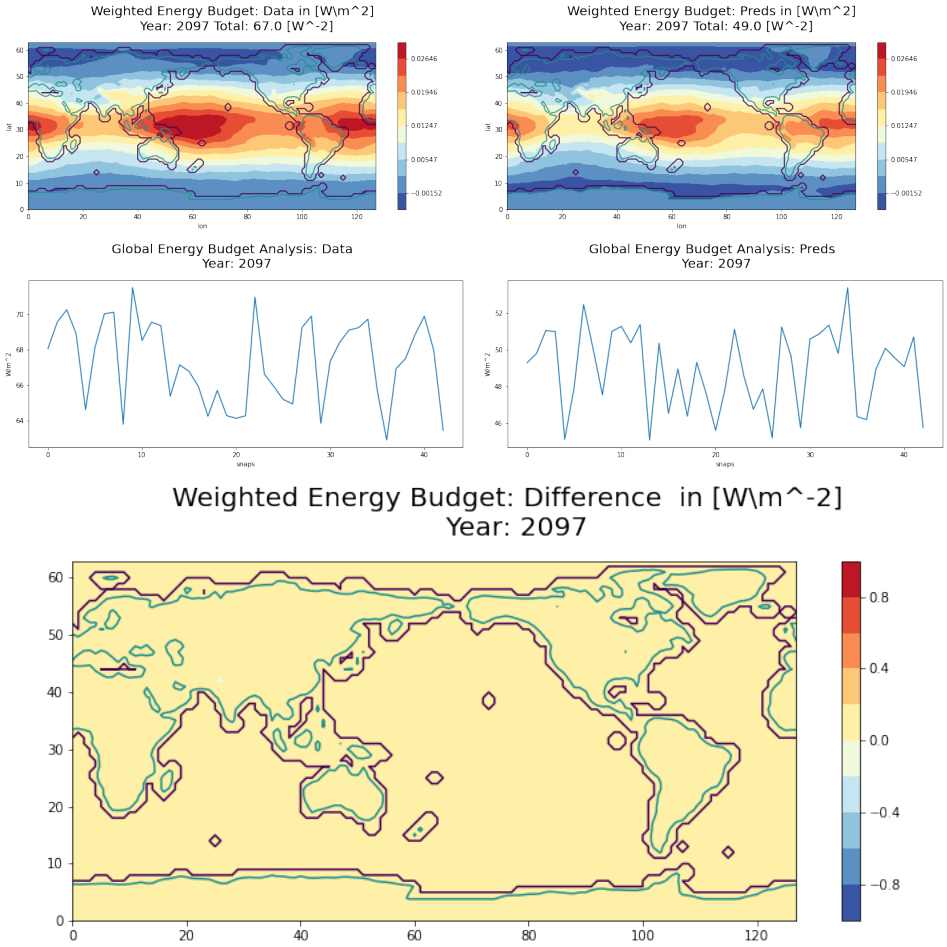


Figure 23: **Energy Budget Constraint** *EBC* **OOD test year 2097**: On left side of the first two rows, we show an analysis of the data of the target RT parameterization in ClimART, whereas on right side we show an analysis of the predictions done by a simple MLP ClimART baseline model. The first line shows the error over the grid, whereas the actual error is weight ed by the area each grid cell covers. Below, we plot the difference of data versus model predictions weighted by cell area over the grid.

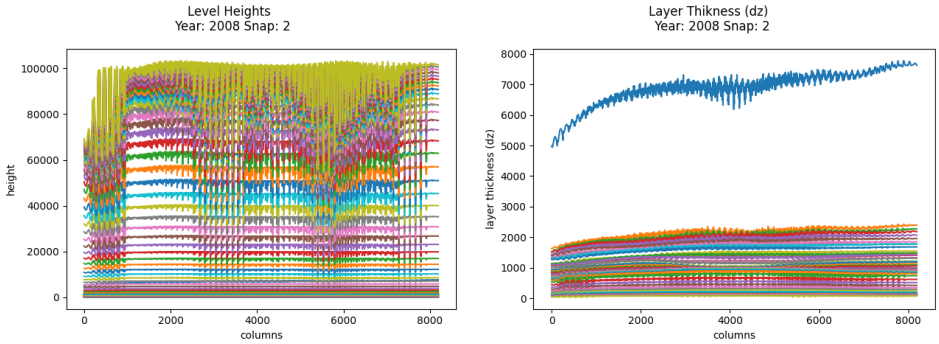


Figure 24: **Level heights and layer thicknesses:** Level heights (left) and layer thicknesses (right) as shown for the test year 2008 (second global snapshot) are unequally distributed across columns as the gridding method of CanESM5 divides the atmosphere into levels/layers with respect to the surface and not with respect to level.

tion of the fluxes. Instead, we fix the heating rates to being the heating rates of the parameterization, accepting these as ground truth and aiming to improve the flux prediction by enforcing an integral balance, which can be interpreted as improving the vertical distribution of fluxes.

$$BC = \frac{\sum_{k=1}^{N} \alpha_k h_k}{\sum_{k=1}^{N} \alpha_k} + \frac{\Phi}{\sum_{k=1}^{N} \alpha_k} = 0 \tag{12}$$

$$\Phi = \begin{cases} F_0^{\uparrow} - F_N^{\uparrow} + F_N^{\downarrow}, LWR \\ F_0^{\uparrow} - F_0^{\downarrow} - F_N^{\uparrow} + F_N^{\downarrow}, SWR \end{cases} \tag{13}$$

8.2.5.1 Analysis of Significance in ClimART

We again compare the integral balance error of both the data and predictions by a simple MLP baseline, as presented by [7]. To get a grasp of how significance the error is, we also plot it in relation to the heating rate averaged over all layers for the respective snap. We can see in 25 and 19 that error is present in both data and model, but rather small for the year taken from the normal test set, whereas the model seems to have problems imitating the data in the regions that define the edges of the sun's radiation imprint. As seen in 25 and 26, the integral balance constraint becomes more significant for out of distribution cases. Not only do the error rates become larger, also the emerging pattens diverse drastically between data and model predictions.

All in all, the integral balance constraint seems like a promising candidate to incorporate in the training design of deep learning methods to help improve physical consistency and generalization ability. In the remainder of the work, we will thus concentrate on this constraint only to test our methods.

8.3 Results

8.3.1 Quantitative

Base Models Training Convergence We note that both base models, meaning the GNN and the MLP architecture, show a fast convergence in terms of training loss, which is the RMSE between the models' predictions and the target parameterization of shortwave up and down-welling fluxes (Figure 29). Similar to the behaviour reported in the ClimART [7], the GNN converges faster, and yields a better and an overall more stable performance than the MLP. This becomes evident as well when looking at the convergence behaviour of the evaluation metric (upwelling shortwave RMSE only on

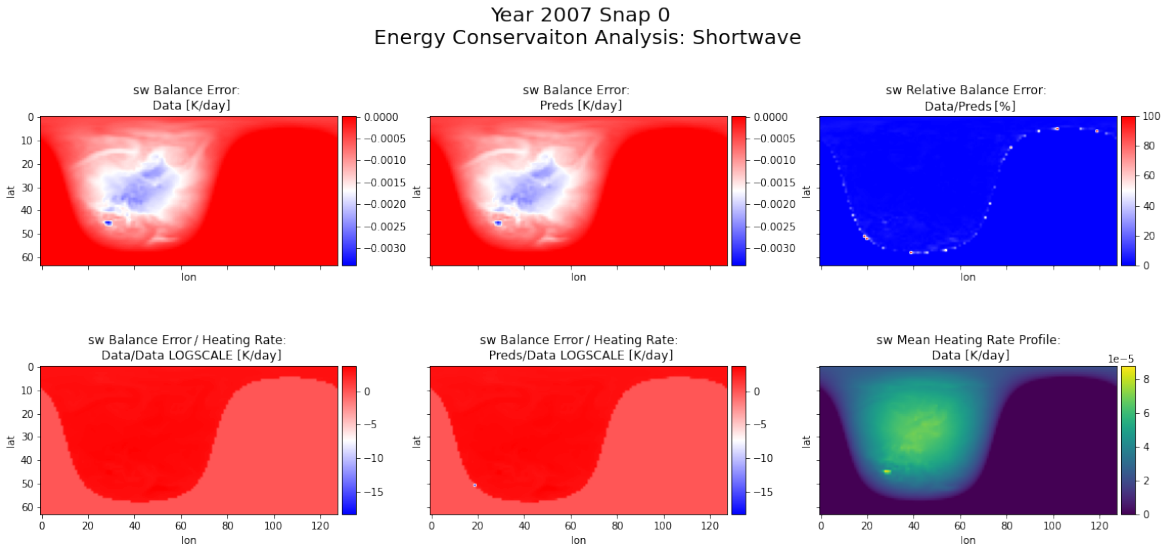


Figure 25: Integral Balance Constraint Analysis: Comparison of the upholding of the integral balance constraint for the test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only shortwave radiative fluxes are considered. Relative Errors are plotted on a log scale.

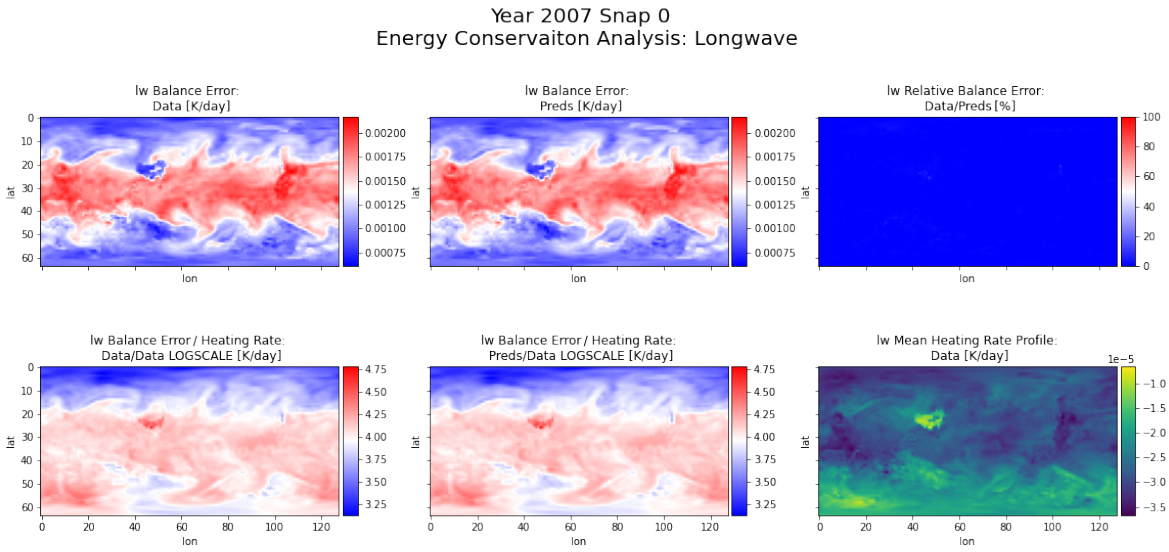


Figure 26: Integral Constraint Analysis: Comparison of the upholding of the integral constraint for the test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only longwave radiative fluxes are considered. Relative Errors are plotted on a log scale.

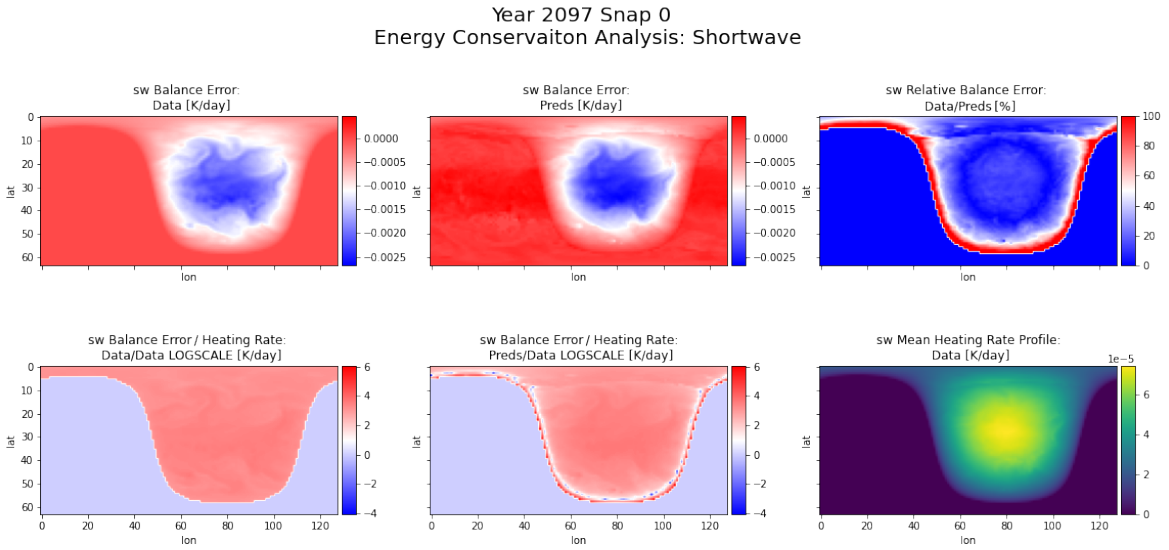


Figure 27: Integral Balance Constraint Analysis: Comparison of the upholding of the integral balance constraint for the out of distribution test year 2097 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only shortwave radiative fluxes are considered. Relative Errors are plotted on a log scale.

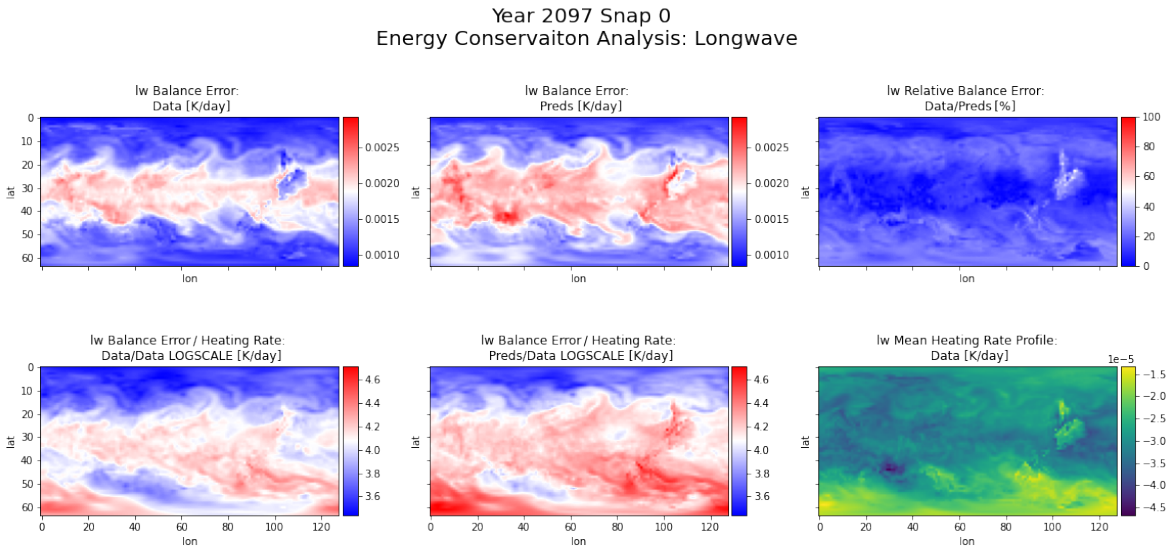


Figure 28: Integral Constraint Analysis: Comparison of the upholding of the integral balance constraint for the out of distribution test year 2007 between the data and the predictions of a simple MLP model which was trained on the full training dataset. In this plot, only longwave radiative fluxes are considered. Relative Errors are plotted on a log scale.

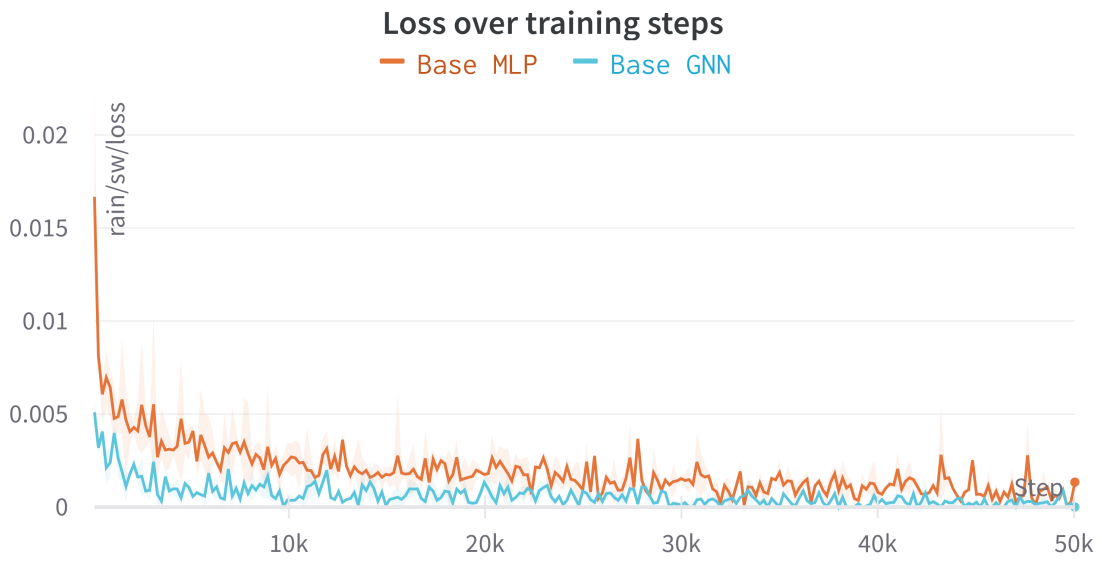


Figure 29: **Base Models Loss Training Convergence:** The Figureshows the convergence of the flux-based training loss (a weighted combination of up- and downwelling RMSE between predicted and target parameterization fluxes, y-axis) over training steps

validation set) (Figure 30), which is the metric early stopping of training is conditioned on - since it is evaluated on the validation test set which is not part of the training data, overfitting can be avoided. Early stopping means, stopping the training process after a certain number of epochs, given that the evaluation metric did not improve significantly over the last n epochs. We can justify that the chosen number of maximum epochs 65 seems to be enough to achieve a satisfactory degree of convergence while not overfitting.

Two-Head Models Training Convergence All Two-Head models plateau in terms of the evaluation metric pretty quickly, which is the RMSE balance error of the shortwave flux prediction computed over the validation test set, and terminate their training stages before the 'deadline' of a maximum of 35 epochs (Figure 32). Looking at the convergence plots suggests, that the combined loss approach makes it too hard for the models to decrease the evaluation metric, whereas at least some improvement in early training can be made out for models trained with the balance-based loss only (Figure 32).

Flux RMSE and MBE - Base Models We report all flux prediction statistics in Table 2, meaning the root-mean-squared-error and bean-bias-error averages over all vertical levels and averaged over all years belonging to each of the test sets, thus the main test set, the OOD test year, the historic and the future test set. We can note, that although the base models were trained on a smaller subset of the training data, they can measure up to the performance reported for the ClimART baselines for most cases.

In ClimART, the baseline MLP is reported to have an RMSE mean over both up- and downwelling fluxes of 0.701 ± 0.04 and a mean MBE of -0.16 ± 0.07 [7]. We report slightly worse and less stable Base MLP behaviour when training with data bagging with an RMSE 01.984 ± 0.331 and MBE 0.042 ± 0.325 (see Table 7). The GNN in ClimART is reported to have an average test RMSE over up- and downwelling fluxes of 0.648 ± 0.04 and an MBE of -0.142 ± 0.03 on the main test set, which is even exceeded by our base models' performance with an RMSE of 0.648 ± 0.004 and an MBE of

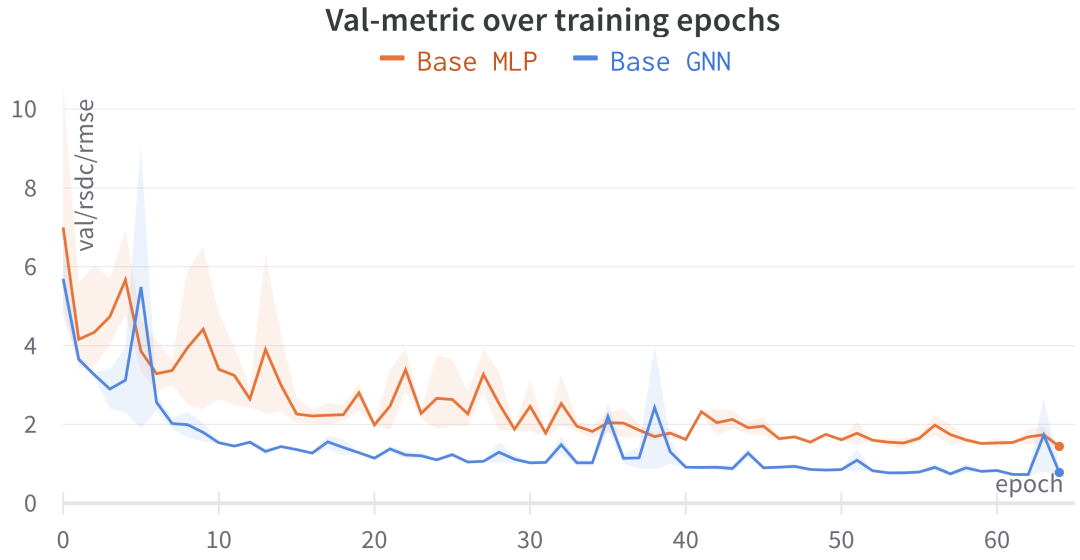


Figure 30: **Base Models Val-Metric Training Convergence:** The Figureshows the convergence of the evaluation metric (RMSE between the models predictions and the target parameterization of the upwelling shortwave, y-axis, computed over the validation set) over training epochs.

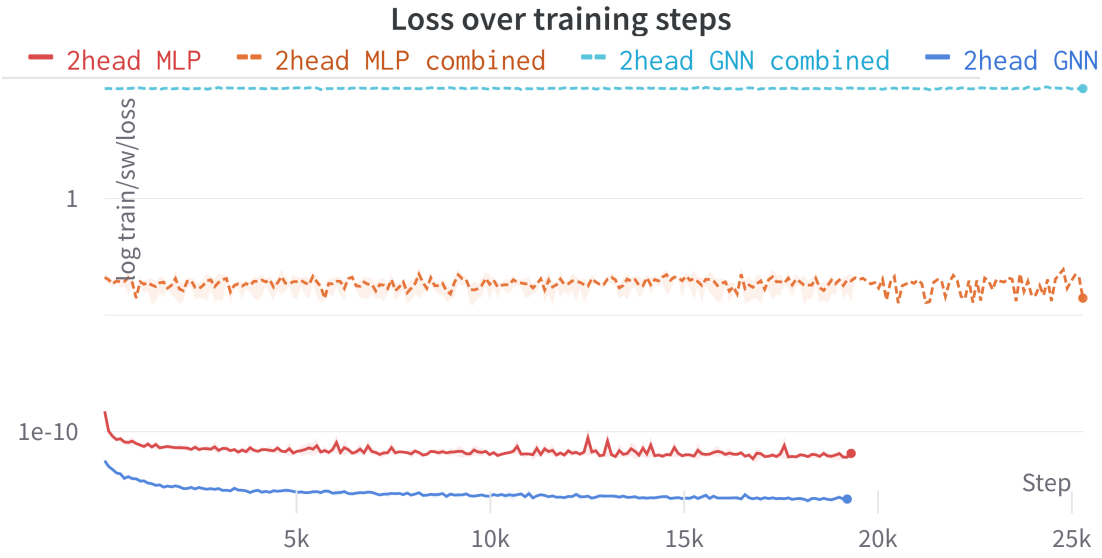


Figure 31: **Two-Head Models Loss Training Convergence:** The Figure shows the convergence of the balance-based (and flux-based in case of combined models)training loss (the squared error of the balance constraint violation error of the predictions up- and downwelling fluxes and target heating rates for the standard Two-Head models or a weighted sum of this loss and the flux-based rmse, log scale, y-axis) over training epochs.

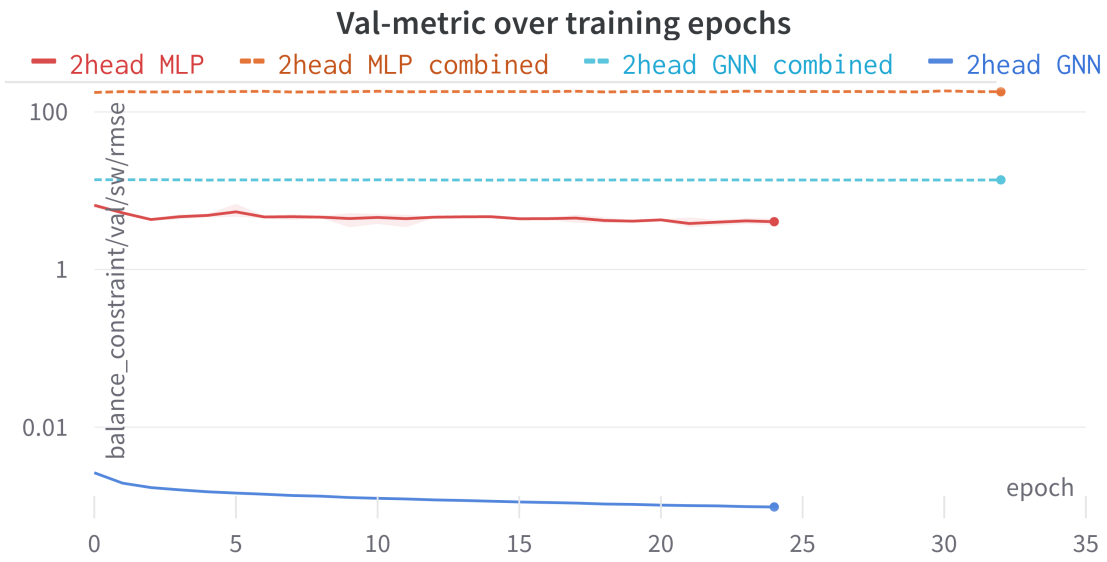


Figure 32: **Two-Head Models Val-Metric Training Convergence:** The Figureshows the convergence of the evaluation metric (the squared error of the balance constraint violation error of the predictions up- and downwelling fluxes and target heating rates computed on the validation set, log-scale, y-axis) over training epochs.

-0.0142 ± 0.03. This thus provides strong evidence that inductive bias, such as the structural bias by the GNN model, improves convergence behaviour even when less training data is present, as it is the case in our data bagging approach.

In terms of generalization, we notice that all base models perform worse, the further away the test year is in time from the training years (see Table 8). This behaviour and the severity of errors match the report by the authors of ClimART.

In conclusion, even though the training set was limited, our base models show similar performance and generalization behaviour as the reported baselines.

Flux RMSE and MBE / Scores - All Models As expected, all models diverge significantly from the base models' flux predictions in terms of flux performance measures when loss and structure in training is changed (see tables 3-6). Surprisingly, we can report that the Two-Head MLP models that were trained on a combined loss, meaning equal contribution of flux (shortwave, upwelling) and balance error RMSE, results in a similar or ever better generalization performance with less variation over the ensemble than the base models it is build upon. This becomes especially evident when looking at combined flux scores, meaning averages over years, layers and up- and downwelling streams (tables 7 and 8). This, however, can probably be attributed to the fact, that the base models were stopped early in their training process and due to data bagging, can learn from novel unseen data in the second training stage.

Mean Balance Score per test set - All Models Measuring the violation of the integral balance error (see Section 8.2.5), measured in an RMSE fashion averaged over ensemble members and

Model	test/rsuc/rmse	test/rsuc/mbe	test/rsdc/rmse	test/rsdc/mbe
Base MLP	2.13 ± 0.39	-0.16 ± 0.64	1.84 ± 0.28	0.24 ± 0.25
Base GNN	0.71 ± 0.02	0.02 ± 0.08	0.82 ± 0.06	0.11 ± 0.1

Model	test OOD/rsuc/rmse	test OOD/rsuc/mbe	test OOD/rsdc/rmse	test OOD/rsdc/mbe
Base MLP	14.55 ± 8.3	0.24 ± 2.11	26.5 ± 24.08	-4.68 ± 5.42
Base GNN	0.99 ± 0.26	-0.01 ± 0.05	1.18 ± 0.26	-0.04 ± 0.04

Model	test historic/rsuc/rmse	test historic/rsuc/mbe	test historic/rsdc/rmse	test historic/rsdc/mbe
Base MLP	12.09 ± 2.24	-1.67 ± 6.85	7.13 ± 2.22	-1.21 ± 2.83
Base GNN	2.32 ± 0.46	0.13 ± 0.34	1.95 ± 0.21	0.02 ± 0.31

Model	test future/rsuc/rmse	test future/rsuc/mbe	test future/rsdc/rmse	test future/rsdc/mbe
Base MLP	53.6 ± 13.6	1.1 ± 24.9	95.76 ± 73.82	-52.47 ± 47.58
Base GNN	37.05 ± 33.8	0.01 ± 9.13	67.16 ± 87.03	-29.11 ± 48.94

Table 2: **Base Flux Statistics:** Base models flux statistics averaged over years of each of the test sets and over vertical levels.

Model	test/rsuc/rmse	test/rsuc/mbe	test/rsdc/rmse	test/rsdc/mbe
Base MLP	2.13 ± 0.39	-0.16 ± 0.64	1.84 ± 0.28	0.24 ± 0.25
Base GNN	0.71 ± 0.02	0.02 ± 0.08	0.82 ± 0.06	0.11 ± 0.1
2head MLP	52.7 ± 0.29	0.14 ± 0.0	269.41 ± 2.24	-0.19 ± 0.0
2head MLP combined	5.03 ± 0.18	0.14 ± 0.0	2.32 ± 0.16	-0.02 ± 0.0
2head GNN	54.77 ± 0.06	0.2 ± 0.0	226.82 ± 1.0	0.05 ± 0.0
2head GNN combined	4.71 ± 0.2	0.2 ± 0.0	1.8 ± 0.06	0.21 ± 0.0

Table 3: **Main Flux Statistics:** All models flux statistics averaged over years of the main test set and vertical levels.

Model	test OOD/rsuc/rmse	test OOD/rsuc/mbe	test OOD/rsdc/rmse	test OOD/rsdc/mbe
Base MLP	14.55 ± 8.3	0.24 ± 2.11	26.5 ± 24.08	-4.68 ± 5.42
Base GNN	0.99 ± 0.26	-0.01 ± 0.05	1.18 ± 0.26	-0.04 ± 0.04
2head MLP	53.45 ± 0.28	0.05 ± 0.0	265.1 ± 2.2	-2.49 ± 0.0
2head MLP combined	12.28 ± 0.07	0.05 ± 0.0	13.38 ± 0.16	-2.33 ± 0.0
2head GNN	54.88 ± 0.06	0.02 ± 0.0	226.19 ± 0.99	-0.2 ± 0.0
2head GNN combined	4.62 ± 0.18	0.02 ± 0.0	1.93 ± 0.09	-0.05 ± 0.0

Table 4: **OOD Flux Statistics:** All models flux statistics averaged over years of the OOD test set and vertical levels.

Model	test historic/rsuc/rmse	test historic/rsuc/mbe	test historic/rsdc/rmse	test historic/rsdc/mbe
Base MLP	12.09 ± 2.24	-1.67 ± 6.85	7.13 ± 2.22	-1.21 ± 2.83
Base GNN	2.32 ± 0.46	0.13 ± 0.34	1.95 ± 0.21	0.02 ± 0.31
2head MLP	52.89 ± 0.28	-6.89 ± 0.0	265.47 ± 2.18	-1.91 ± 0.0
2head MLP combined	13.03 ± 0.06	-6.89 ± 0.0	6.16 ± 0.13	-1.72 ± 0.0
2head GNN	60.05 ± 0.06	0.62 ± 0.0	229.66 ± 1.02	0.15 ± 0.0
2head GNN combined	5.05 ± 0.17	0.62 ± 0.0	3.21 ± 0.02	0.3 ± 0.0

Table 5: **Historic Flux Statistics:** All models flux statistics averaged over years of the historic test set and vertical levels.

Model	test future/rsuc/rmse	test future/rsuc/mbe	test future/rsdc/rmse	test future/rsdc/mbe
Base MLP	53.6 ± 13.6	1.1 ± 24.9	95.76 ± 73.82	-52.47 ± 47.58
Base GNN	37.05 ± 33.8	0.01 ± 9.13	67.16 ± 87.03	-29.11 ± 48.94
2head MLP	53.8 ± 0.17	9.31 ± 0.0	243.1 ± 2.74	-13.76 ± 0.0
2head MLP combined	41.49 ± 0.04	9.31 ± 0.0	41.21 ± 0.77	-13.61 ± 0.02
2head GNN	56.43 ± 0.05	10.62 ± 0.0	234.38 ± 1.04	7.26 ± 0.01
2head GNN combined	21.98 ± 0.04	10.62 ± 0.0	24.26 ± 0.38	7.45 ± 0.0

Table 6: **Future Flux Statistics**All models flux statistics averaged over years of the future test set and vertical levels.

Model	test/flux score/rmse/mean	test/flux score/mbe/mean
Base MLP	1.984 ± 0.331	0.042 ± 0.325
Base GNN	0.764 ± 0.038	0.063 ± 0.087
2head MLP	161.055 ± 1.255	-0.029 ± 0.003
2head MLP combined	3.672 ± 0.171	0.058 ± 0.003
2head GNN	140.794 ± 0.524	0.125 ± 0.003
2head GNN combined	3.254 ± 0.133	0.204 ± 0.002

Model	test OOD/flux score/rmse/mean	test OOD/flux score/mbe/mean
Base MLP	20.525 ± 16.189	-2.219 ± 1.665
Base GNN	1.085 ± 0.263	-0.027 ± 0.008
2head MLP	159.275 ± 1.229	-1.219 ± 0.002
2head MLP combined	12.828 ± 0.046	-1.138 ± 0.001
2head GNN	140.532 ± 0.523	-0.089 ± 0.003
2head GNN combined	3.273 ± 0.131	-0.012 ± 0.001

Table 7: **Average Flux Scores:** Flux Score RMSE and MBE (average over up- and downwelling flux over) of all models averaged over all years of each test set and averaged over all vertical levels.

Model	test historic/flux score/rmse/mean	test historic/flux score/mbe/mean
Base MLP	9.611 ± 1.882	-1.439 ± 4.73
Base GNN	2.132 ± 0.285	0.077 ± 0.321
2head MLP	159.178 ± 1.219	-4.399 ± 0.003
2head MLP combined	9.597 ± 0.096	-4.304 ± 0.002
2head GNN	144.856 ± 0.541	0.386 ± 0.003
2head GNN combined	4.13 ± 0.075	0.456 ± 0.001

Model	test future/flux score/rmse/mean	test future/flux score/mbe/mean
Base MLP	74.681 ± 35.46	-25.686 ± 33.572
Base GNN	52.1 ± 60.381	-14.553 ± 28.735
2head MLP	148.449 ± 1.452	-2.223 ± 0.003
2head MLP combined	41.351 ± 0.374	-2.151 ± 0.011
2head GNN	145.405 ± 0.539	8.942 ± 0.003
2head GNN combined	23.118 ± 0.169	9.036 ± 0.001

Table 8: **Average Flux Scores** Flux Score RMSE and MBE (average over up- and downwelling flux over) of all models averaged over all years of each test set and averaged over all vertical levels.

years belonging to a test set, we can report that our base models and the target Radiative Transfer parameterization (short Physics RT) perform equally bad (see Table 9).

We note that the Two-Head approaches in general decrease the balance error values compared to their base model counterparts significantly, whereas models trained with a combined loss show a less drastic improvement. Especially, the combined Two-Head MLP shows only small improvements in terms of the balance errors. Most models equally perform worse on test years further away from training, except for the Two-Head GNN models, that are able to satisfy the balance constraint next to perfectly on every test set. Notably, for years belonging to the future condition, thus furthest away from the training years, Base models and the combined Two-Head MLP models perform significantly worse than the target Physics RT.

Mean overall score - All models Although building a score measure that summarizes flux scores and balance scores reported in previous Sections is theoretically unphysical, due to their different units ($Wm^{-}2$ vs. $KDay^{-}1$), it gives us a good estimate of the tradeoff capabilities of the models between the two tasks of trying to approximate the target parameterizations flux prediction while minimizing the constraint error. For the overall score, we average the squared individual overall balance and flux scores and take a square root subsequently in order to punish very large errors.

As evident from previous Sections, the Two-Head models, meaning building a rescaling readout layer on top of the base model that is trained on an instance of a constraint-based loss, archive the better tradeoff performance (see Table 10). However, posing the question what model delivers the best tradeoff, requires weighting of the different errors against each other. The overall score computed in 10 assumes equal contribution and is somewhat unphysical, as it combines errors that are in principle measured in different units, which results in the simple Two-Head models being the best models. However, one could argue that actually the Two-Head models trained on a combined loss show the best tradeoff performance as, especially in the case of the Two-Head GNN trained with a combined loss, flux predictions were kept close to the base models, even being slightly better than the base models for future conditions, while the balance errors were reduced below target RT.

8.3.2 Qualitative

8.3.2.1 Balance Error Heat Maps

Following the analysis of significance 8.2.5, we report heating maps of the balance constraint violation as exhibited by the models versus the target data, weighted by the cell area coverage of the respective column for which the balance constraint error is determined (see Appendix section **??** for details). Although this is not done in training, and it does not change the exhibited patterns, it gives us a grasp of significance. As the evaluation of the integral balance error depends on the heating rates, we also plot the errors against an average of target heating rates over all vertical layers. We also plot the vertical averaged heating rates over land coverage, to note any patterns that may be influenced by geographical factors.

Base Models: As evident from figures 33 and 34, violation of the balance constraint by the target parameterization can be found in mild and less mild cases in all test years, without any significant difference for years later away from training years. We can note that our base models imitate the target parameterization behaviour in great detail for years closer to training years, but start to diverge for far-future projections. In all cases, but especially in the future scenarios, differences between data and the model's predictions allocate in the outer region of the suns' imprint it's incoming solar flux projects on the heating rate.

Two-Head Models: Most Two-Head models make local adjustments to satisfy the balance constraint, which can be seen in the difference between the subplots target parameterization versus model flux predictions related to the target heating rates (figures 35, 37 and 38). These local adjustments

Model	test/balance score/mean	test OOD/balance score/mean
Physics RT	1.285 ± 0.049	3.72 ± 0.141
Base MLP	63.992 ± 0.125	187.015 ± 14.271
Base GNN	63.894 ± 0.394	175.973 ± 0.449
2head MLP	1.378 ± 0.129	8.742 ± 0.343
2head MLP combined	62.066 ± 0.747	180.437 ± 2.312
2head GNN	1.285 ± 0.049	3.72 ± 0.141
2head GNN combined	61.922 ± 0.157	171.854 ± 0.251

Model	test historic/balance score/mean	test future/balance score/mean
Physics RT	6.804 ± 0.288	16.918 ± 0.378
Base MLP	342.363 ± 28.257	1081.315 ± 499.512
Base GNN	351.602 ± 8.216	979.197 ± 523.294
2head MLP	21.768 ± 0.891	94.596 ± 0.302
2head MLP combined	341.31 ± 2.975	839.91 ± 8.544
2head GNN	6.804 ± 0.288	16.918 ± 0.378
2head GNN combined	331.263 ± 0.386	384.375 ± 3.056

Table 9: **Average Balance Scores** Balance Score (balance constraint rmse) of all models averaged over all years of each test set and averaged over all vertical levels.

Model	all/flux score/rmse/mean	all/balance score/mean	all/overall score/rmse
Base MLP	26.7 ± 12.597	418.671 ± 125.139	305.405 ± 0.0
Base GNN	14.021 ± 15.222	392.666 ± 131.954	288.224 ± 0.0
2head MLP	156.989 ± 1.278	31.621 ± 0.255	113.24 ± 0.0
2head MLP combined	16.862 ± 0.135	355.931 ± 2.0	251.966 ± 0.0
2head GNN	142.897 ± 0.532	7.182 ± 0.211	101.171 ± 0.0
2head GNN combined	8.444 ± 0.043	237.354 ± 0.761	167.941 ± 0.0

Table 10: **Average Overall Scores:** Scores averaged over all test years and vertical levels. The overall score is a mean of the overall balance and flux score.

cluster in the outer regions of the sun's imprint, thus where incoming solar flux should be not a strong. The only exception to this is the Two-Head MLP trained with a combined loss, showing almost no difference to its base models (36 vs 33).

8.3.2.2 Energy Conservation Error Heat Maps

Following the analysis of significance 8.2.1, we report heating maps of the energy conservation violation as exhibited by the models versus the target data, weighted by the cell area coverage of the respective column for which the energy conservation error is determined (see Appendix section **??** for details). Although this is not part of the training pipeline, neither of the base nor the 2 head models', it is a way to externally validate physical correctness of flux predictions. To check whether there is any violation of the incident TOA shortwave flux (see Appendix section **??**, we also plot the incident TOA shortwave fluxes as present in the target RT predictions, model predictions and difference thereof.

Base Models: As evident from figures 39 and 40, violation of energy conservation by the target parameterization is very low for all test years, without any significant difference for years later away from training years. We can note that our base models imitate the target parameterization behaviour very well.

Two-Head Models: All Two-Head models, although trained to adjust the flux predictions of their base models to diverge from the target RT parameterization, nevertheless inherit energy conserving properties for all training sets and incident shortwave TOA values stay the same (figures 41, 43 and 44). This suggests the success of the Hard Constraints employed in the two head models (flux rescaling and sum of energy conservation, see Section 4.3.2).

8.3.2.3 Height errors

To show where Two-Head models make adjustments to the flux predictions, we show mean deviations by height from the target predictions. The resulting height errors are averaged over columns, years and snaps belonging to each of the major test set, meaning main, OOD, historic and the future-condition test set. Adjustments are relatively consistent among each of these test sets, and are the most distinct for Two-Head models trained on the purely constraint-based loss. Up- and downwelling fluxes show the largest deviation around the mid-layer (level) area, with one big or two big "spikes" in deviation, respectively. It is worth noting that the number of layer or level does not translate directly to its atmospheric height as there is no even spacing between levels (for details see Section 8.2.4 Figure 24). One can see that there is no deviation for downwelling flux at TOA, due to the hard constraint regarding incident TOA flux (see Section **??**), however, upwelling flux at TOA diverge significantly for Two-Head models, interestingly being far less for future-conditions than for the other test set, could be interpreted as the long term heating up of the climate system. The largest deviations for heating rates occur near the top of the atmosphere. These observations are consistent, independent on the base model architecture used (MLP vs. GNN), suggesting that they are more than just mere statistical artefacts.

8.4 Glossary

Radiative Transfer (RT): The process by which electromagnetic radiation interacts with matter and is absorbed, scattered, or transmitted through a medium

Climate Model: A computer program that simulates the behaviour of the Earth's climate system, typically by incorporating components such as the atmosphere, oceans, land surface, and ice.

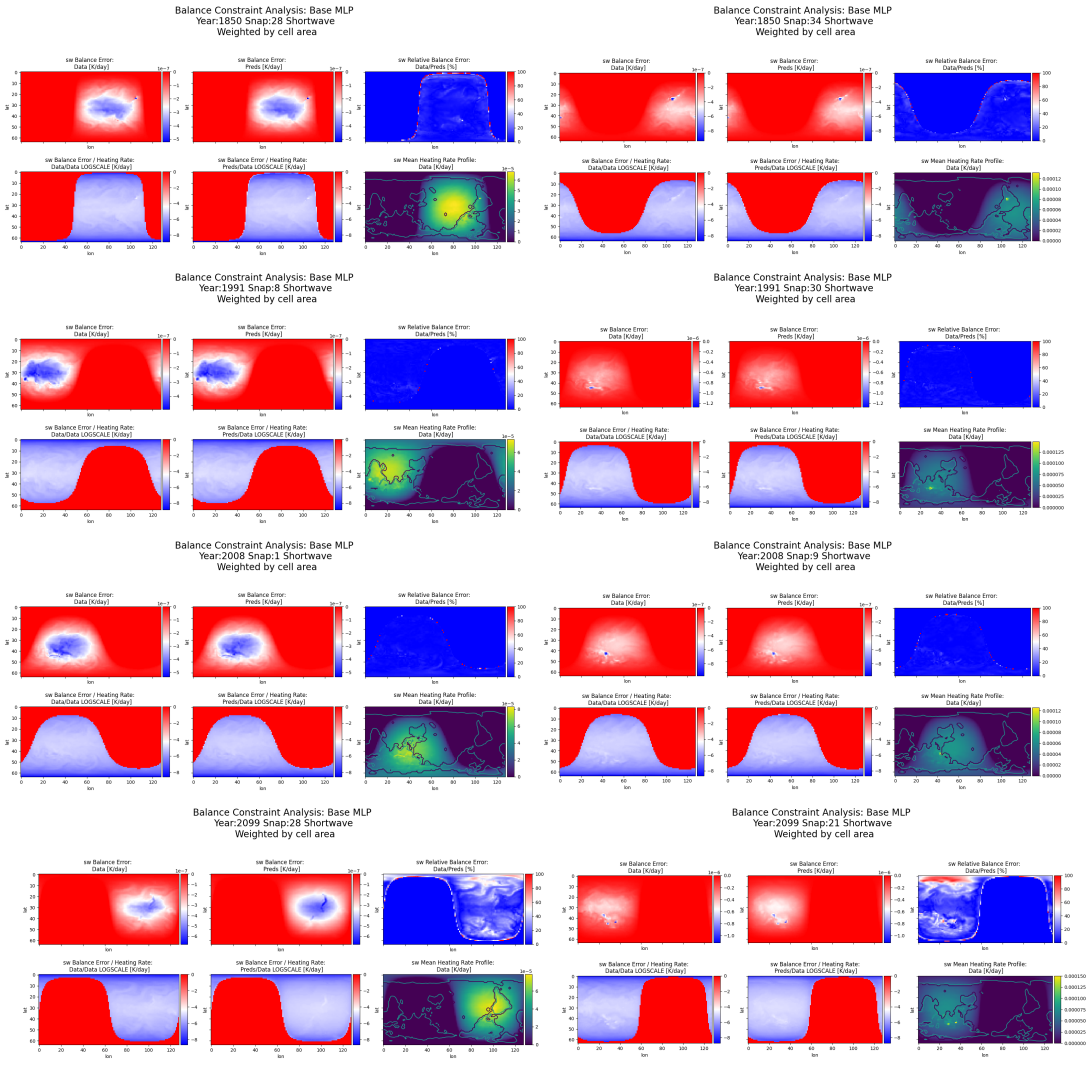


Figure 33: **Base MLP Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

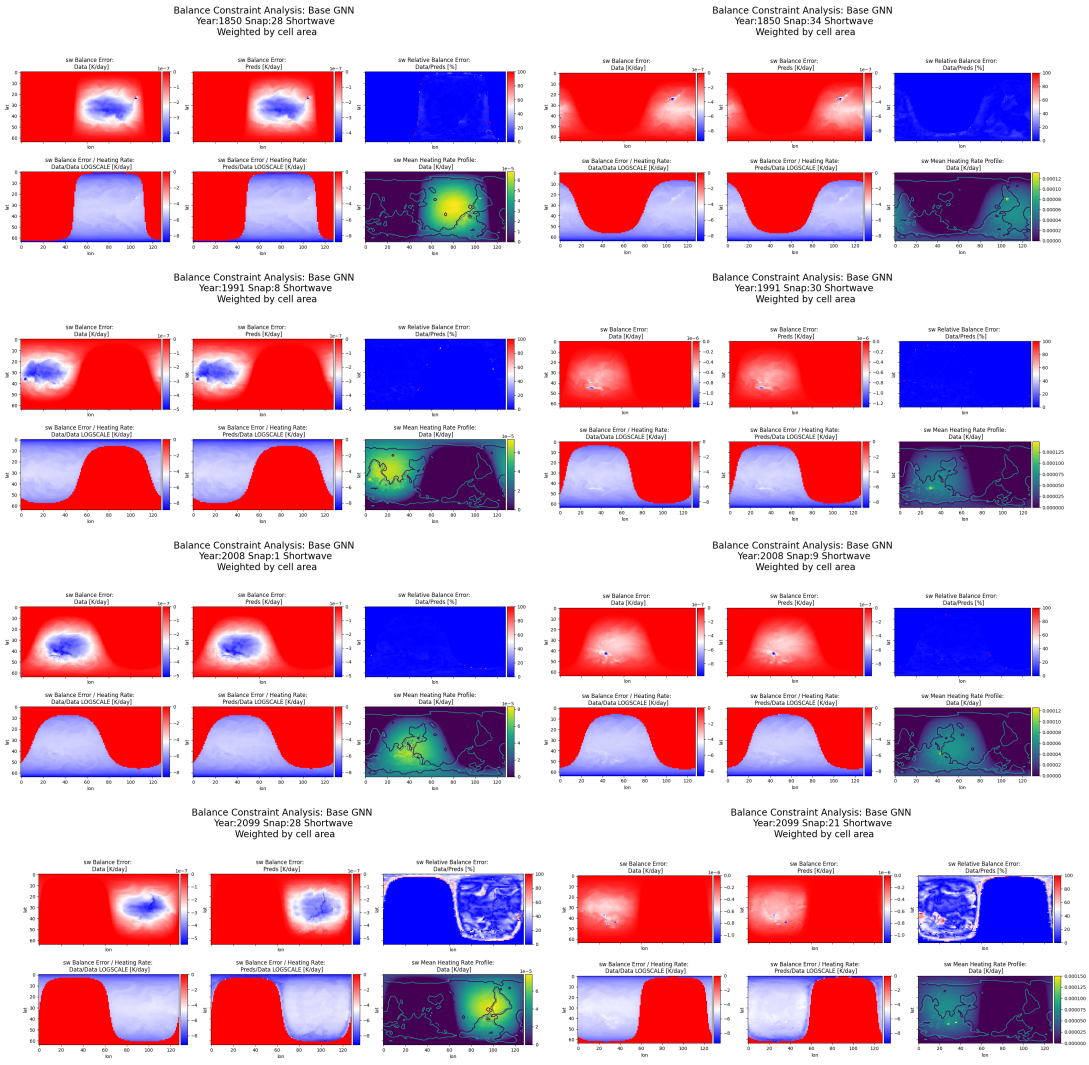


Figure 34: **Base GNN Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

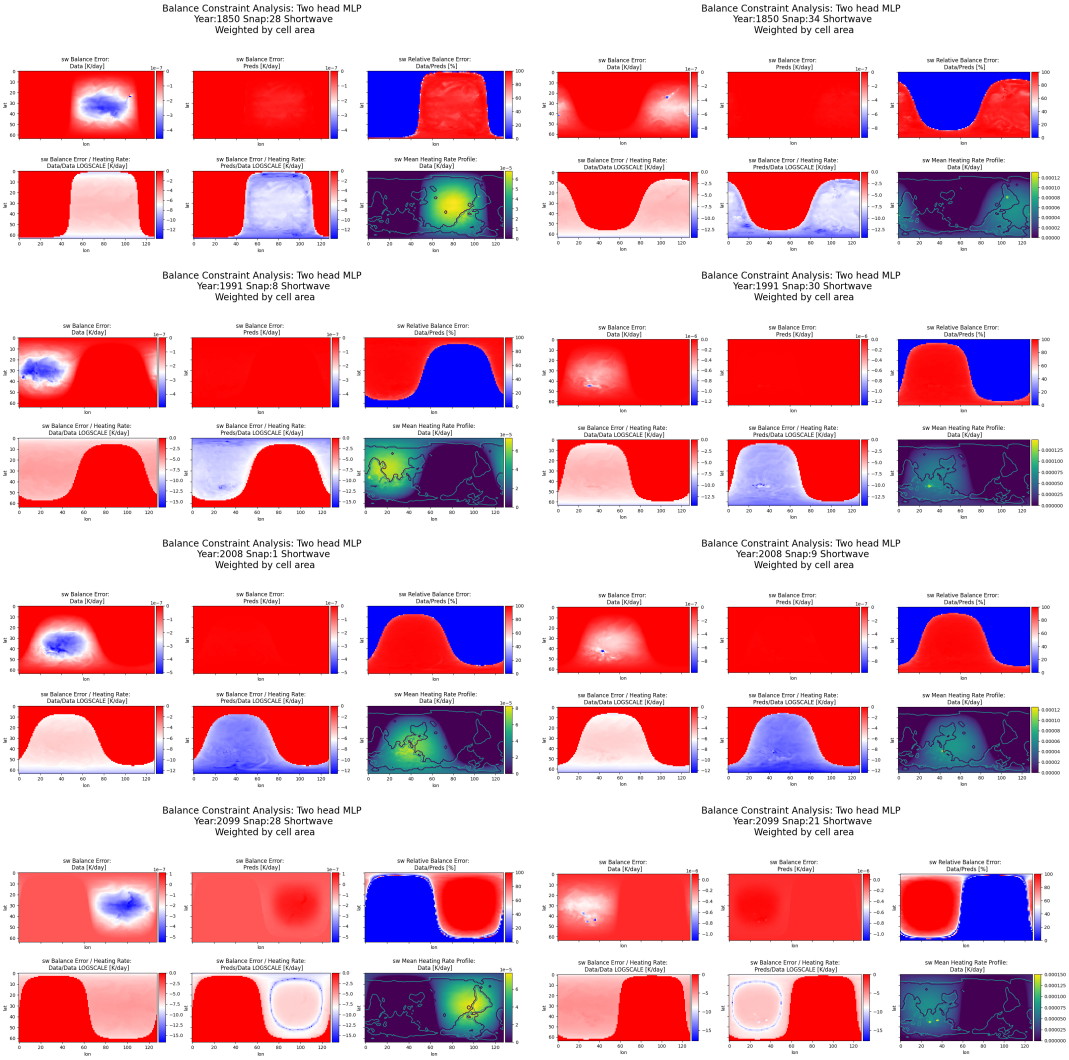


Figure 35: **Two-Head MLP Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

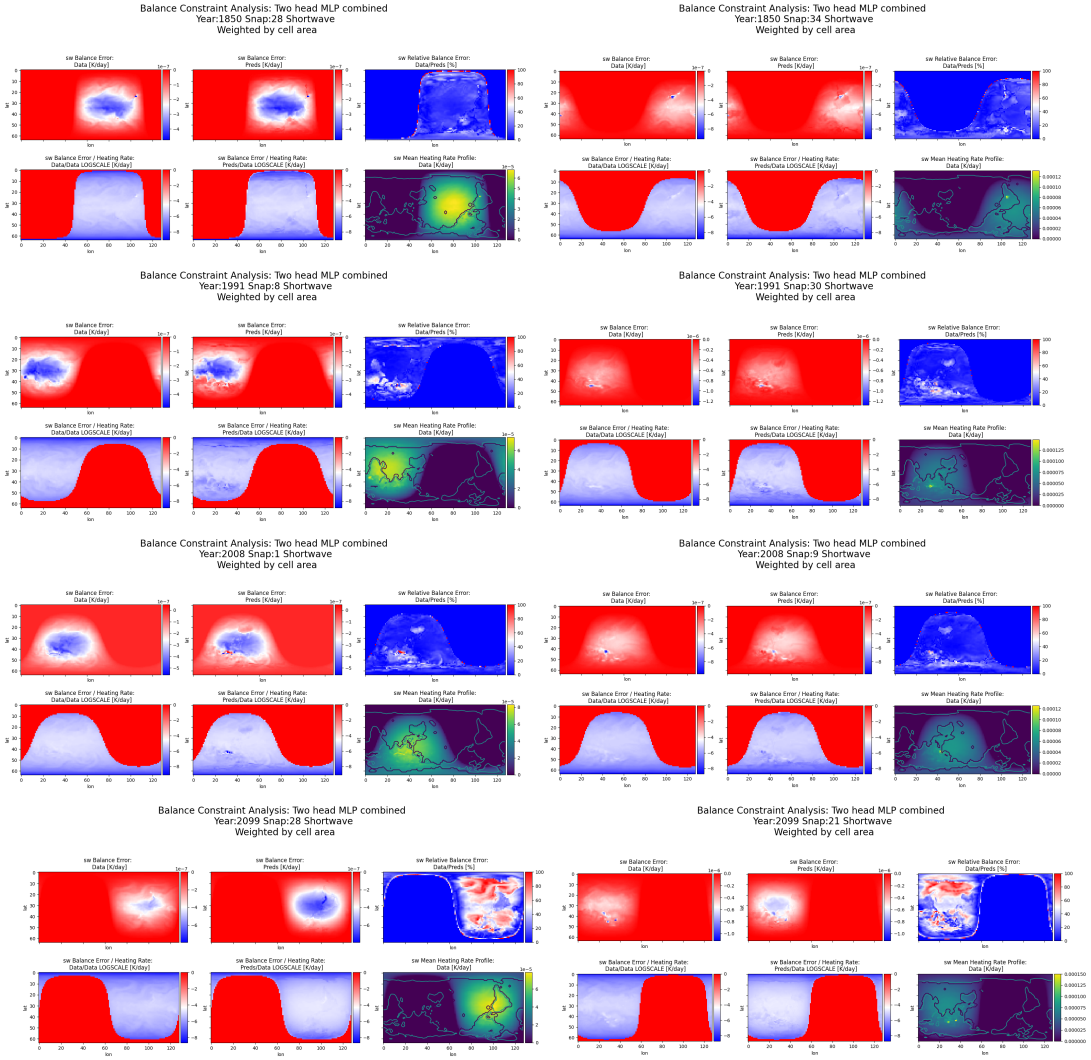


Figure 36: **Two-Head MLP combined Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

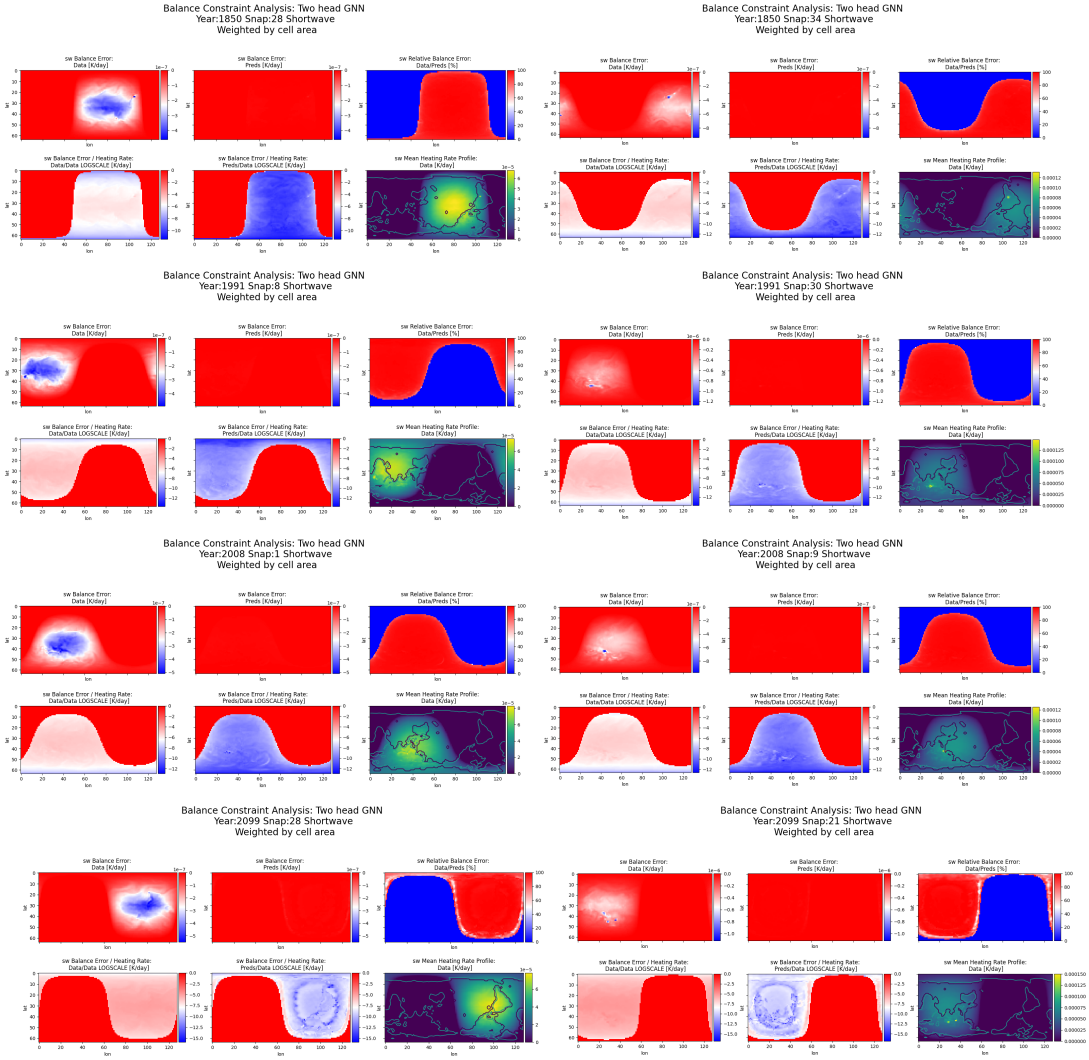


Figure 37: **Two-Head GNN Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

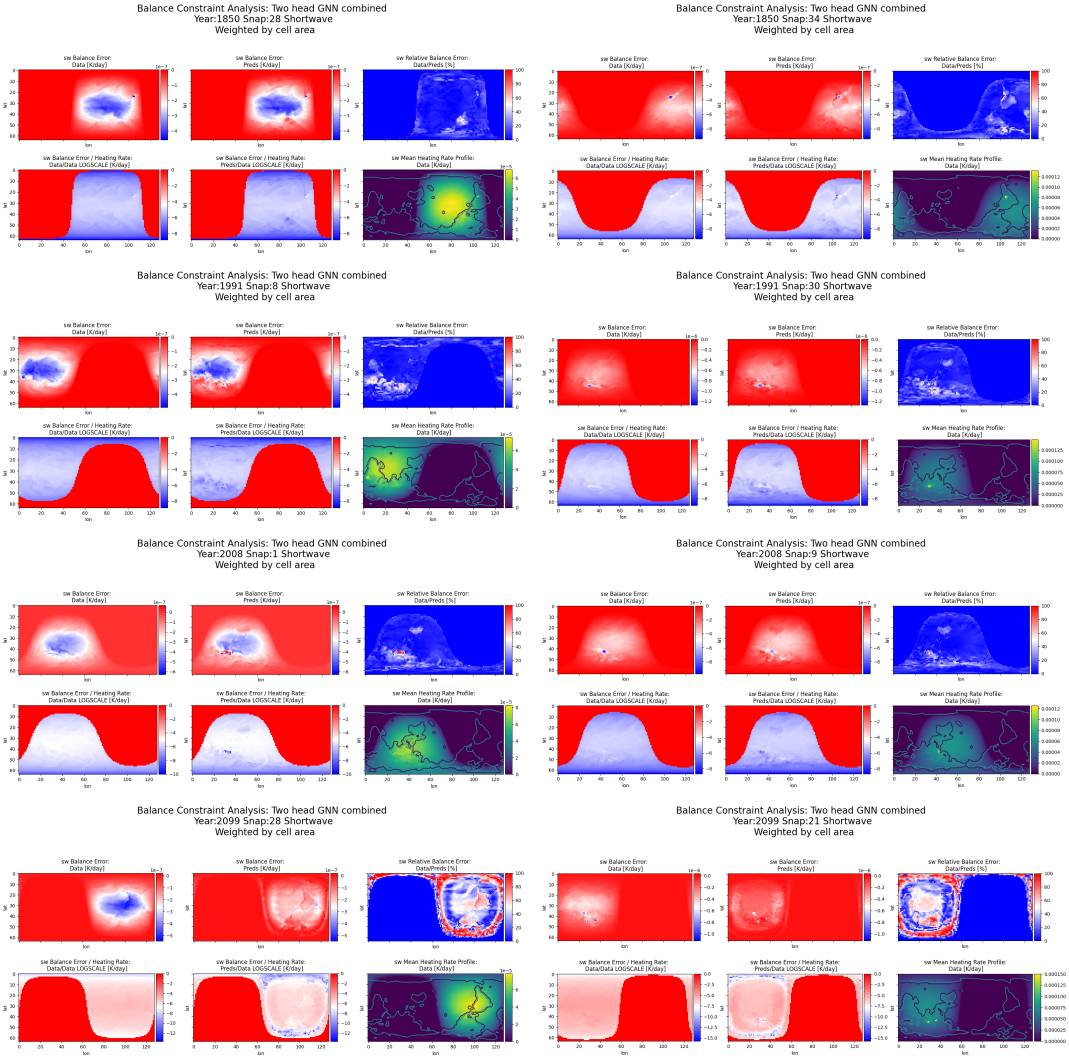


Figure 38: **Two-Head GNN combined Balance Constraint Violation Heat Maps:** This plot shows the comparison between the violation of the integral balance constraint as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. The first column shows snaps that exhibit more violation in general, whereas the second column shows cases with less overall violation. In each plot, the first column shows the heat map of the error for the target data, followed by the relation of that error to the target heating rates averaged over all layers. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the heating target heating rates averaged over all layers over a contour plot of the underlying land area. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

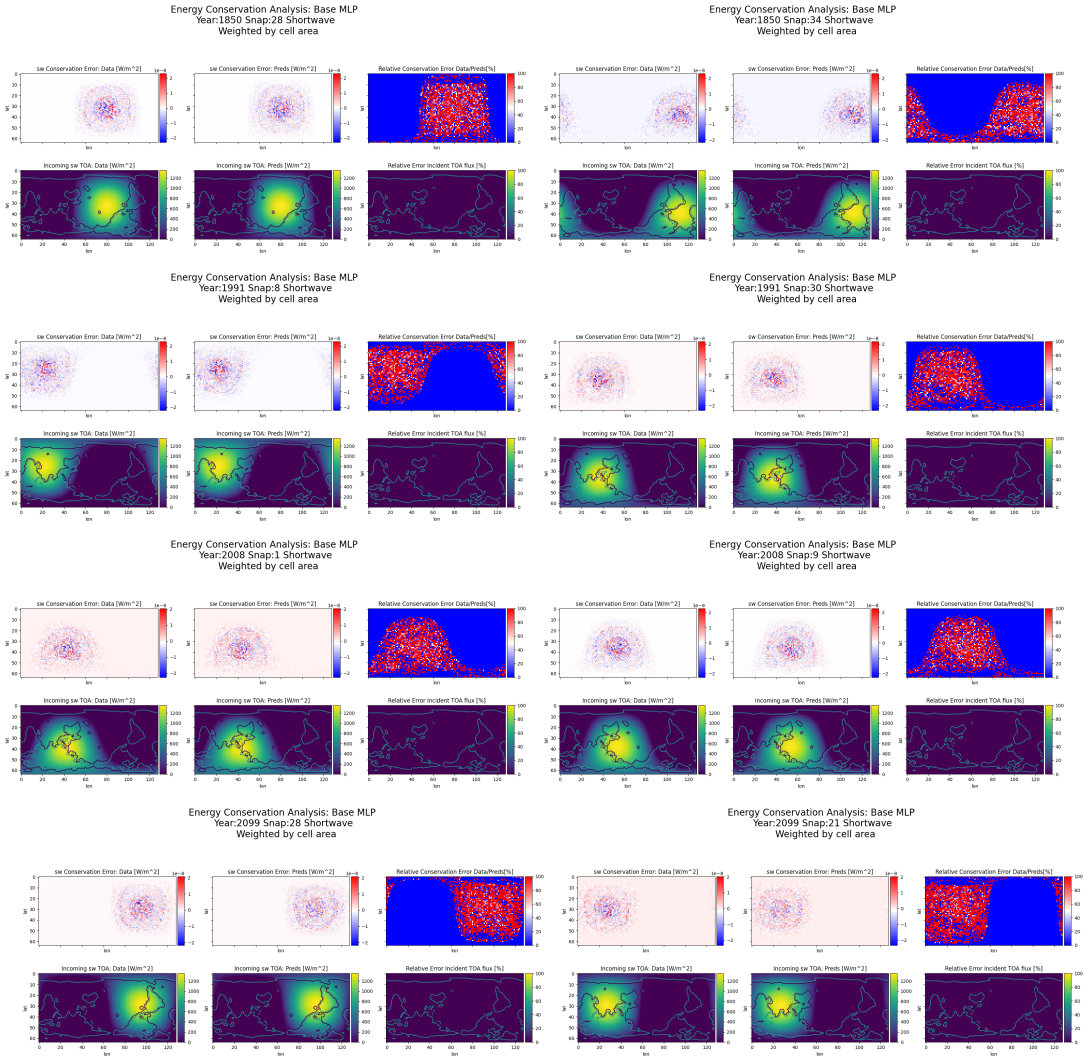


Figure 39: **Base MLP Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

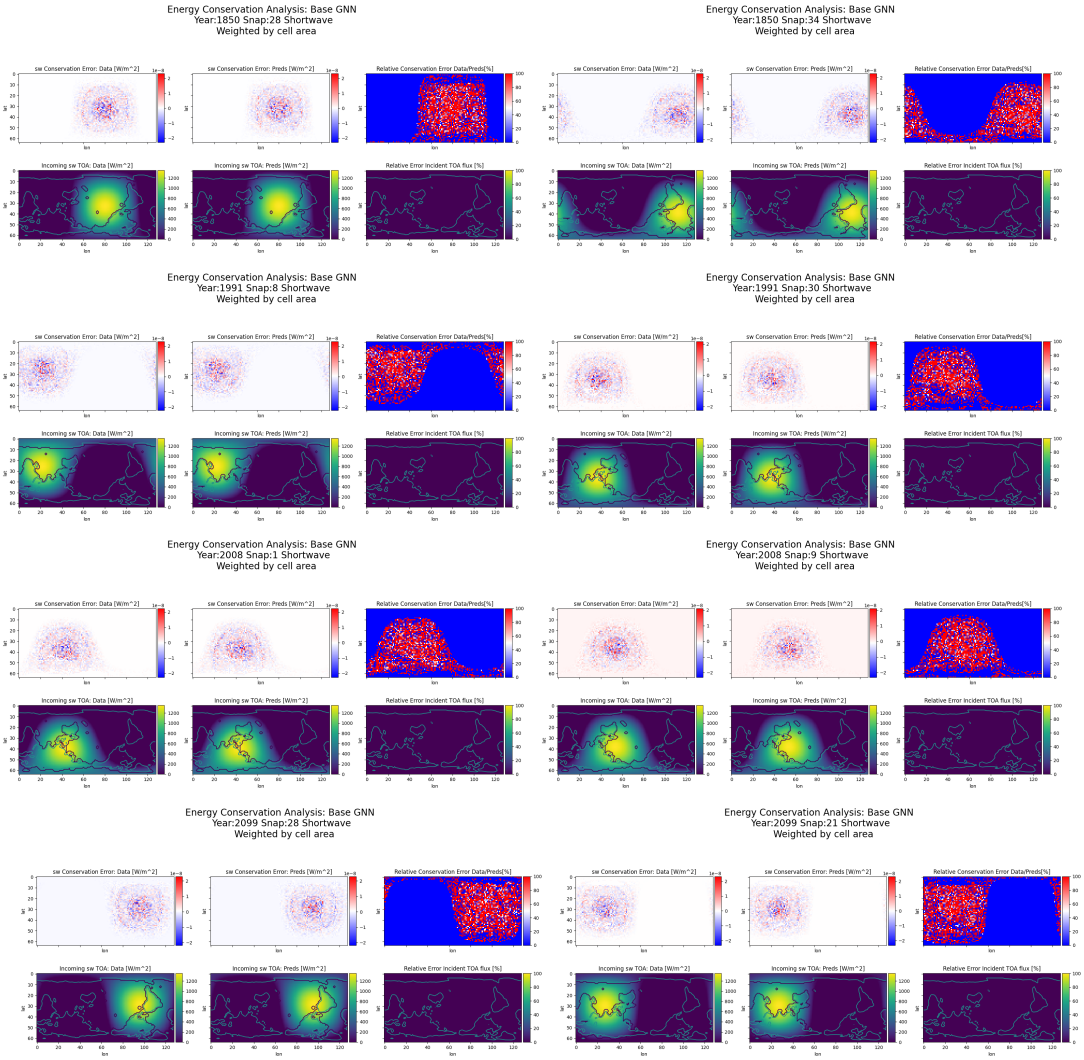


Figure 40: **Base GNN Energy Conservation Violation Heat Maps:** TThis plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

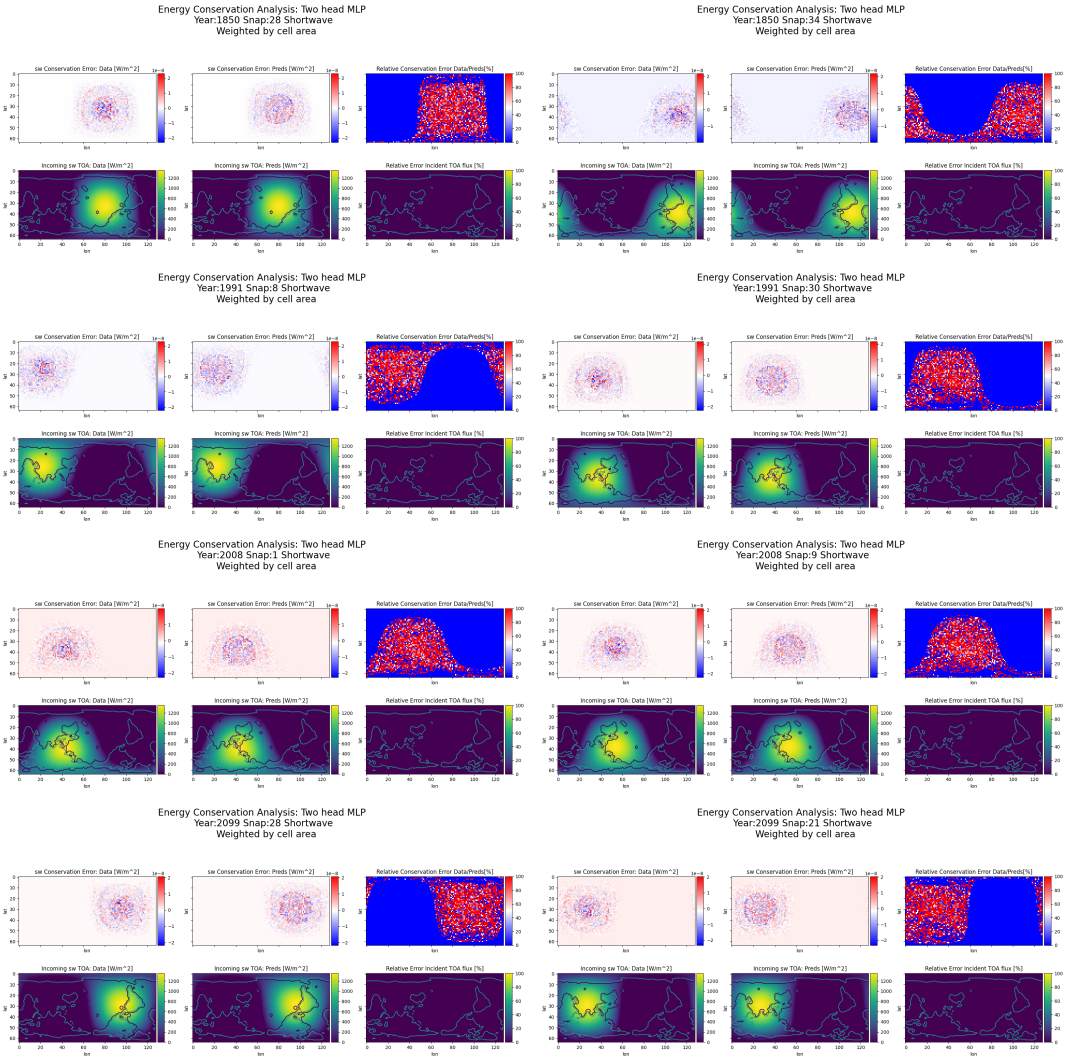


Figure 41: **Two-Head MLP Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

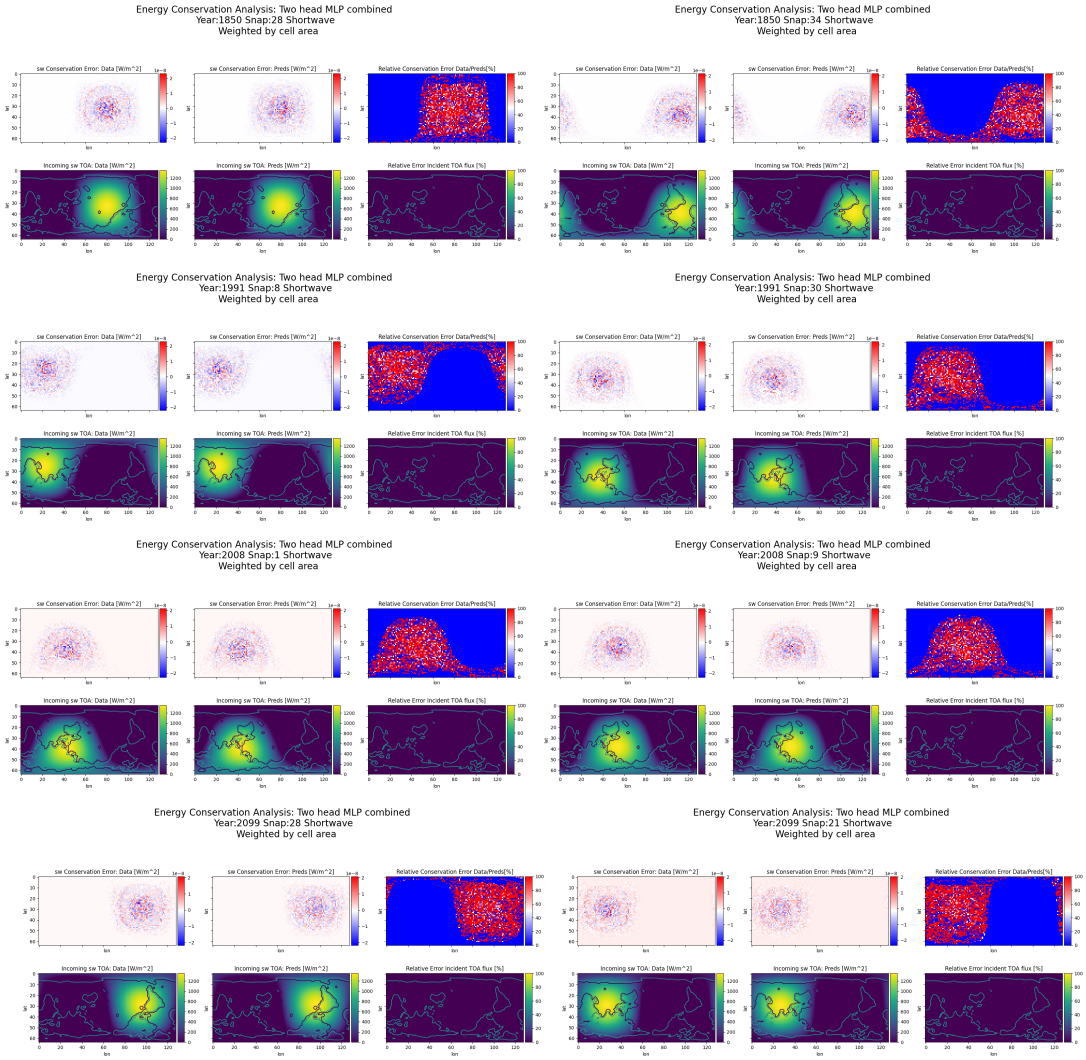


Figure 42: **Two-Head MLP combined Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

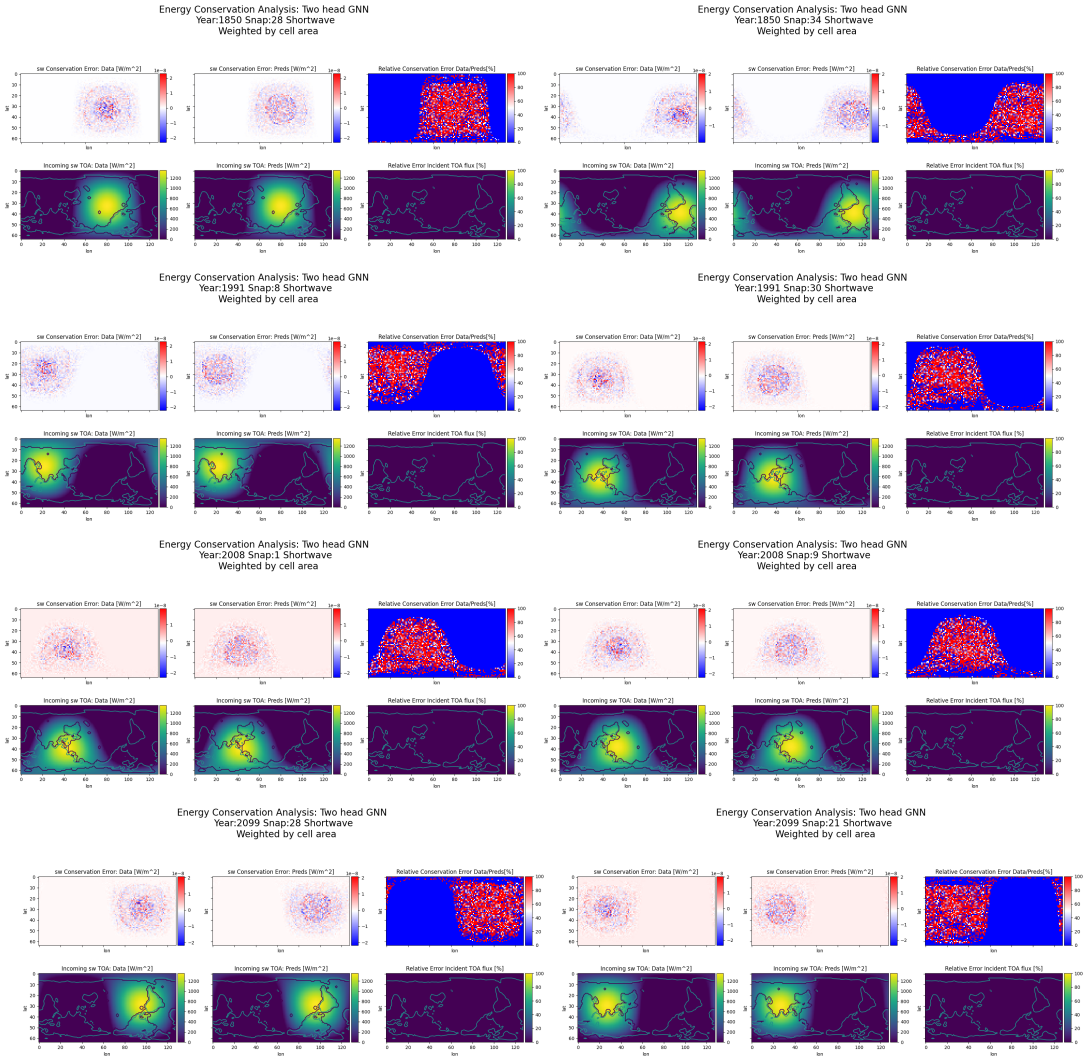


Figure 43: **Two-Head GNN Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

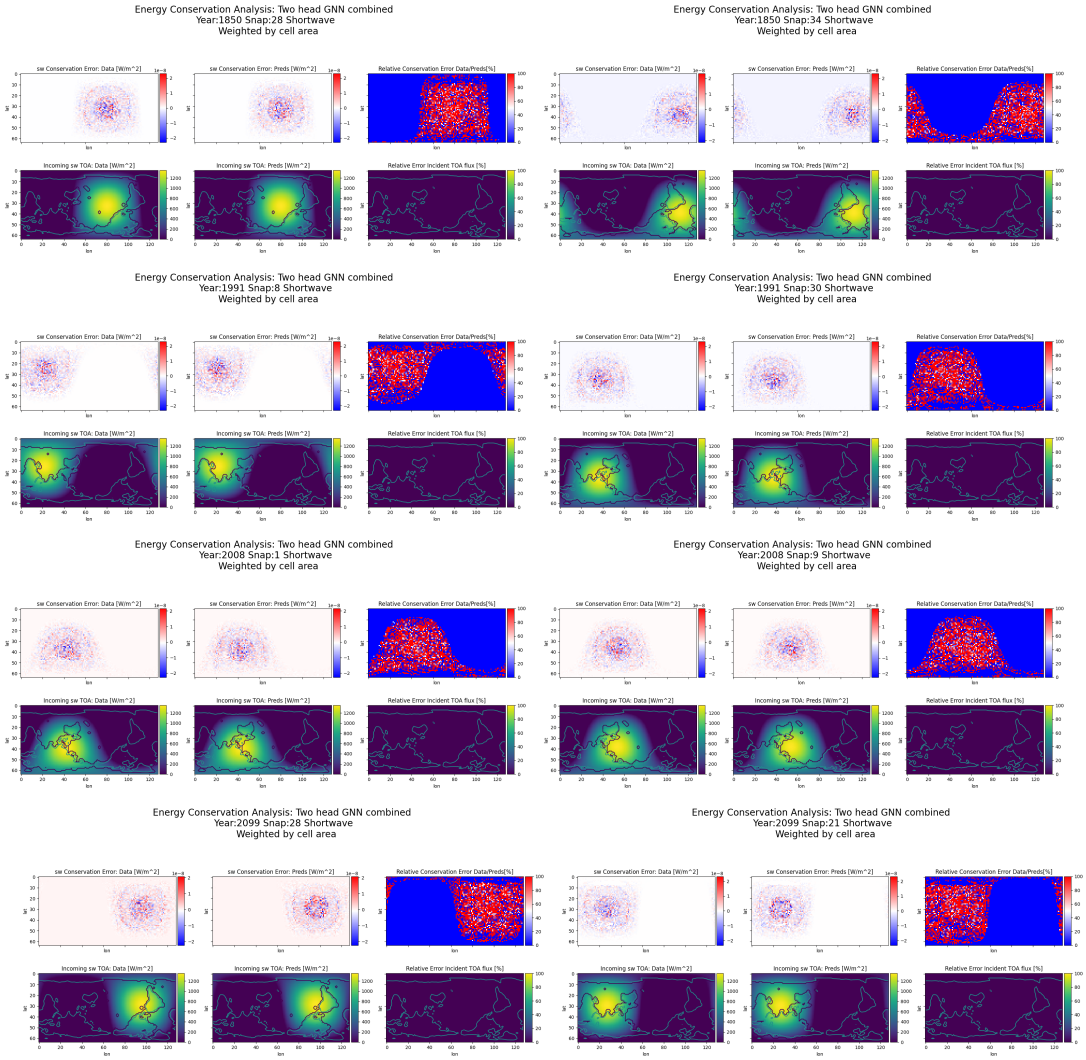


Figure 44: **Two-Head GNN combined Energy Conservation Violation Heat Maps:** This plot shows the comparison between the violation of energy conservation as exhibited in the target parameterization versus a trained model. Each row represents two snaps of a year belonging to one of the four test sets, moving further away from the training data in time as we progress downwards. In each plot, the first column shows the heat map of the error for the target data, followed by a plot of the incident shortwave flux at TOA. The same is repeated for the trained model in the second column. In the third column, we first show the relative error between the target data and the predictions and in the second row, we plot the difference in incident shortwave fluxes. Each pixel represents a full atmospheric column, and values per column are multiplied by cell area covered by that column.

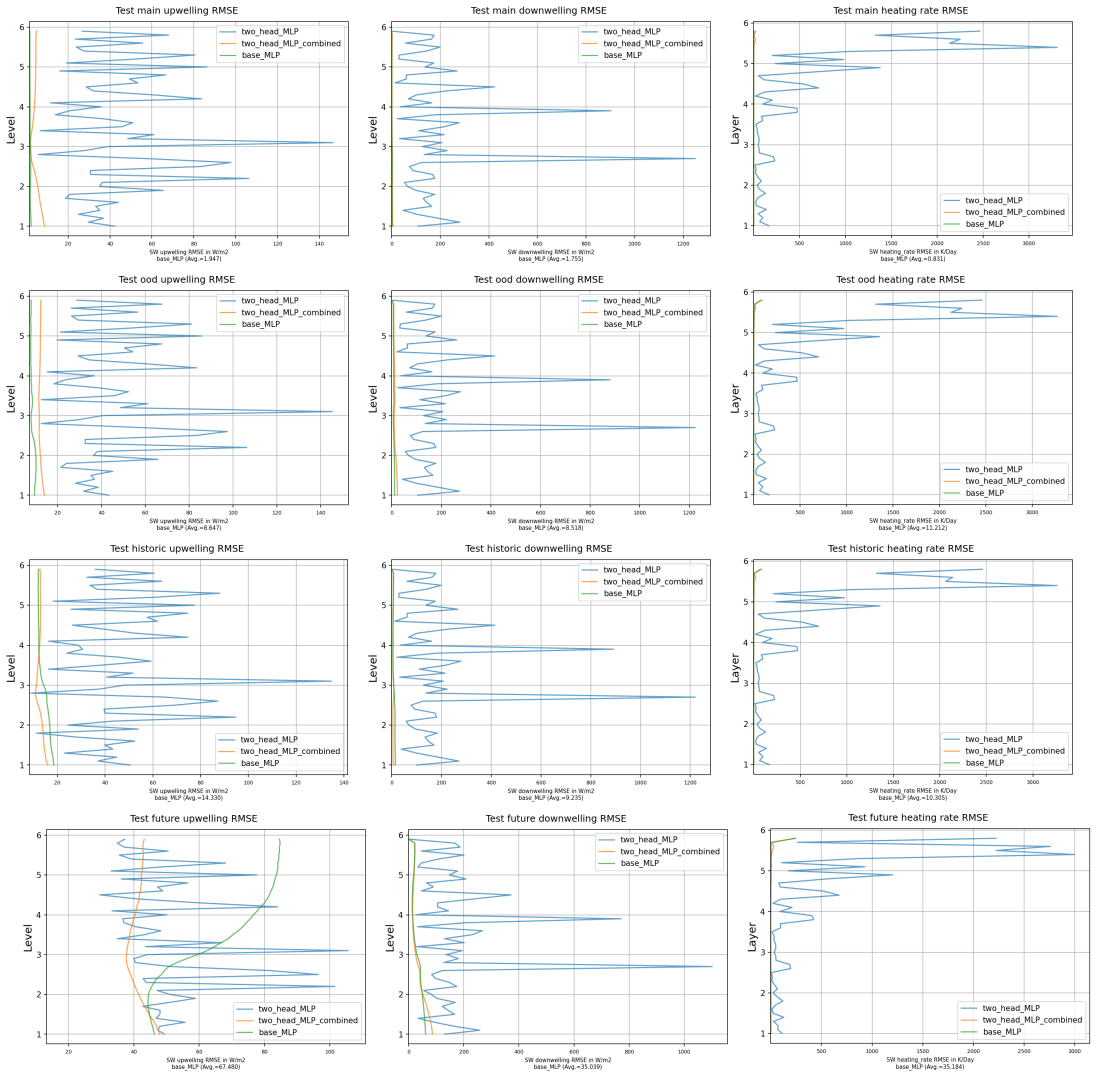


Figure 45: **Height Errors MLP-based models:** This plot shows the difference between model flux predictions and target RT (RMSE) over height in the atmosphere. The lines show errors averaged over years belonging to the main test set, the OOD test set, the historic and the future test set respectively. The first column shows errors of the upwelling flux predictions (y-axis equals levels), the second errors for downwelling flux predictions (y-axis equals levels) and the third shows the difference in resulting heating rates (y-axis equals atmospheric layers).

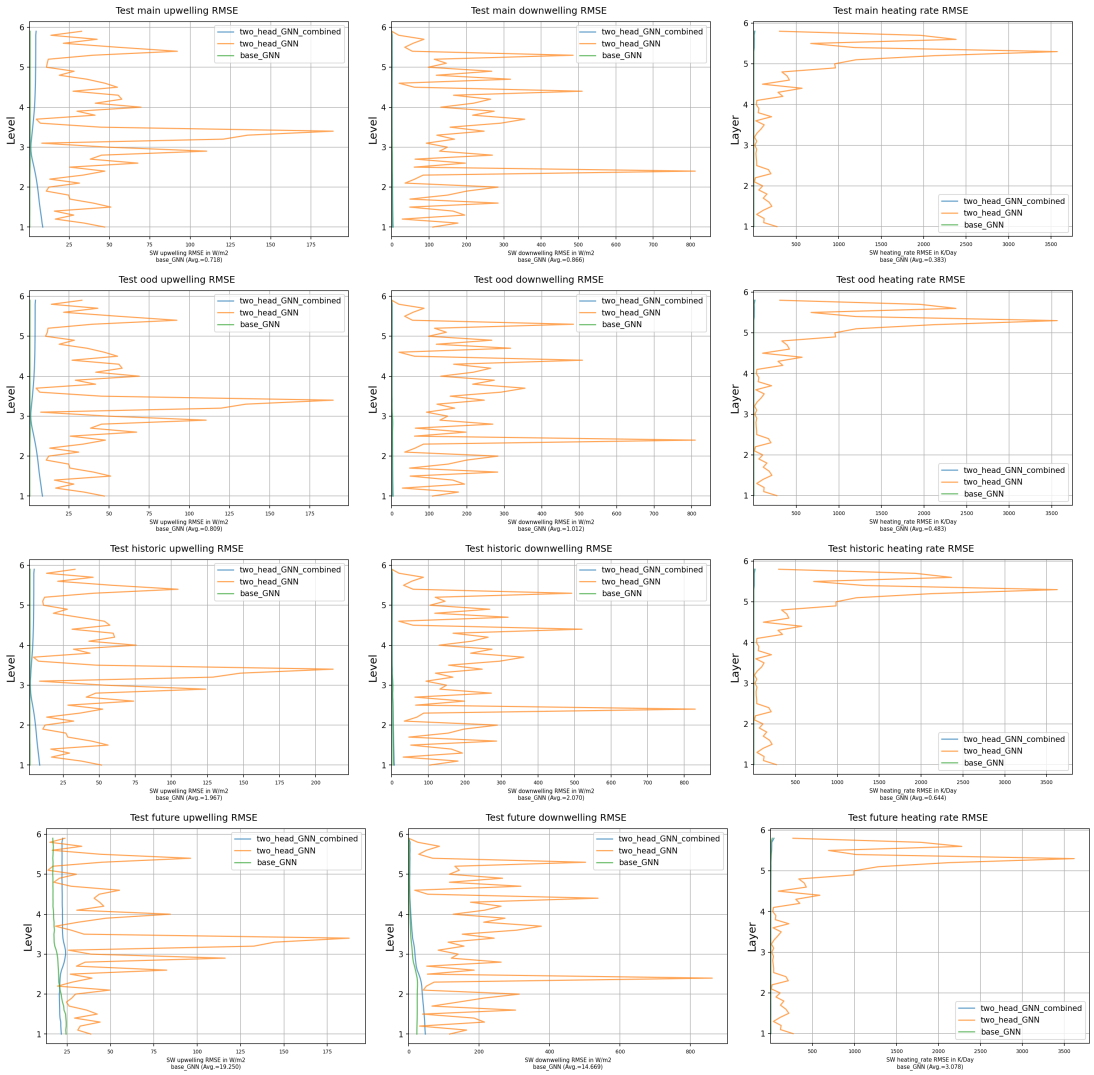


Figure 46: **Height Errors GNN-based models:** This plot shows the difference between model flux predictions and target RT (RMSE) over height in the atmosphere. The lines show errors averaged over years belonging to the main test set, the OOD test set, the historic and the future test set respectively. The first column shows errors of the upwelling flux predictions (y-axis equals levels), the second errors for downwelling flux predictions (y-axis equals levels) and the third shows the difference in resulting heating rates (y-axis equals atmospheric layers).

Radiative Flux: The amount of radiant energy that passes through a unit area per unit time. It is measured in watts per square meter (W/m^2) and is a crucial component of radiative transfer modelling in climate science.

Radiative Intensity: The amount of radiant energy that flows in a specific direction per unit area per unit solid angle. It is measured in watts per square meter per steradian ($W/m^2/sr$) and is an important parameter in radiative transfer models.

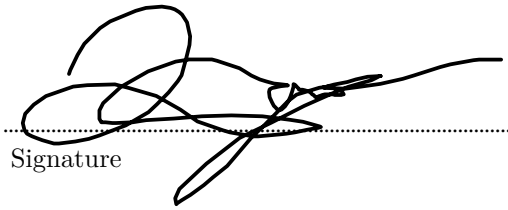
Parameterizations: Simplified mathematical representations of physical processes that are used in climate models to reduce the computational cost of simulating the climate system. In radiative transfer modelling, parameterizations are often used to represent the interaction of radiation with clouds, aerosols, and other atmospheric constituents.

Theory-guided data science: A methodology that combines domain knowledge and machine learning techniques to develop models that are physically interpretable, explainable, and accurate.

Soft Constraints: Constraints that are incorporated into machine learning models as a penalty term or regularization parameter to encourage the model to behave in a specific way. Hard Constraints: Constraints that are enforced on the model during training to ensure that the model output satisfies specific physical principles.

9 Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.



Signature

Osnabrück, 12th April 2023

City, Date