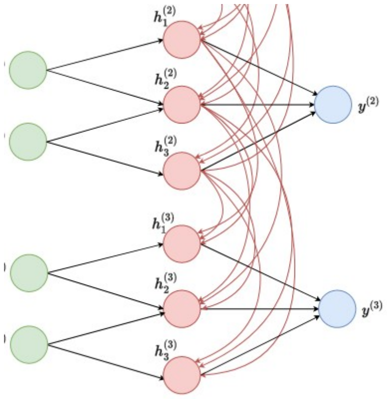
A physically informed recurrent neural network approach for emulating radiative transfer

Danmarks
Meteorologiske
Institut

Peter Ukkonen
Danish Meteorological Institute
peterukk@gmail.com

With help from Matthew Chantry,
Robin Hogan (ECMWF)



The art of approximation

Maxwell's equations in terms of fields $\mathbf{E}(\mathbf{x},t)$, $\mathbf{B}(\mathbf{x},t)$

3D radiative transfer in terms of monochromatic radiances $I(\mathbf{x},\Omega,\nu)$

$$\mathbf{\Omega} \cdot \nabla I(\mathbf{\Omega}) = -\beta_e I(\mathbf{\Omega}) + \frac{\beta_s}{4\pi} \int_{4\pi} p(\mathbf{\Omega}', \mathbf{\Omega}) I(\mathbf{\Omega}') \mathrm{d}\mathbf{\Omega}' + S(\mathbf{\Omega}).$$

1D radiative transfer in terms of two monochromatic fluxes $F\!\downarrow\!(z, \nu)$, $F\!\uparrow\!(z, \nu)$

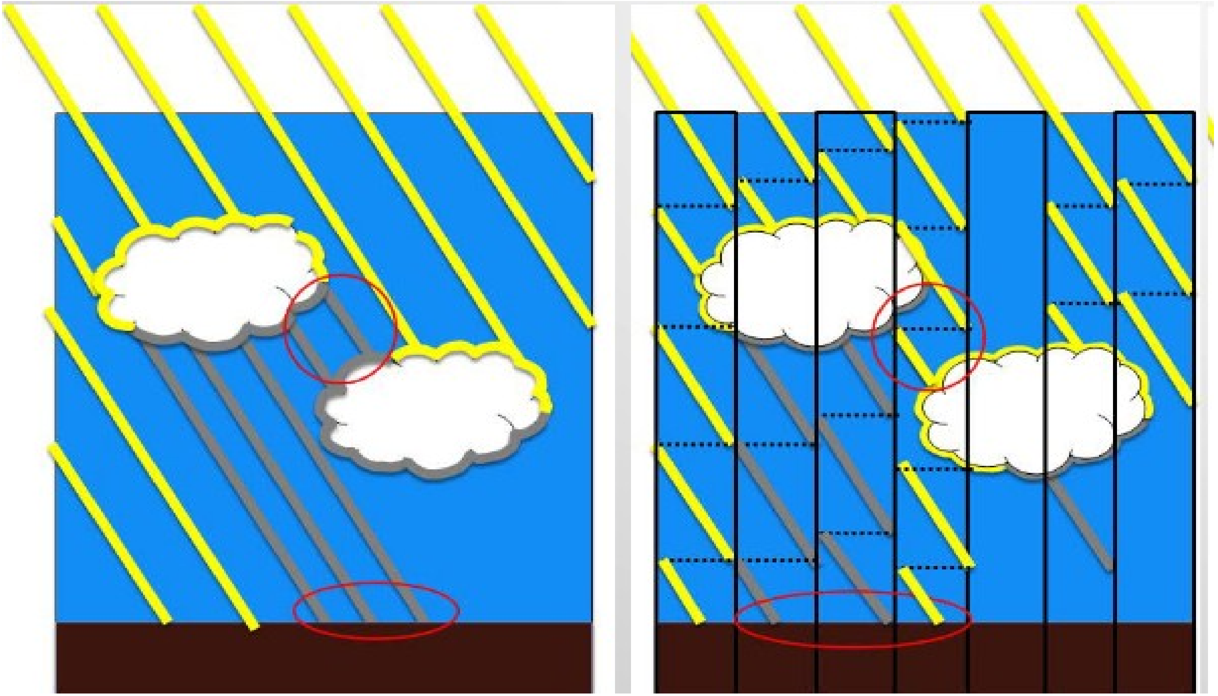
Atmospheric radiation is well-understood but approximated out of computational necessity

- ignore polarization
- group together frequencies
- atmosphere is horizontally homogenous within a grid column ("plane-parallel")
- consider radiation only in two directions, up and down ("two-stream")

Adapted from slides by Robin Hogan

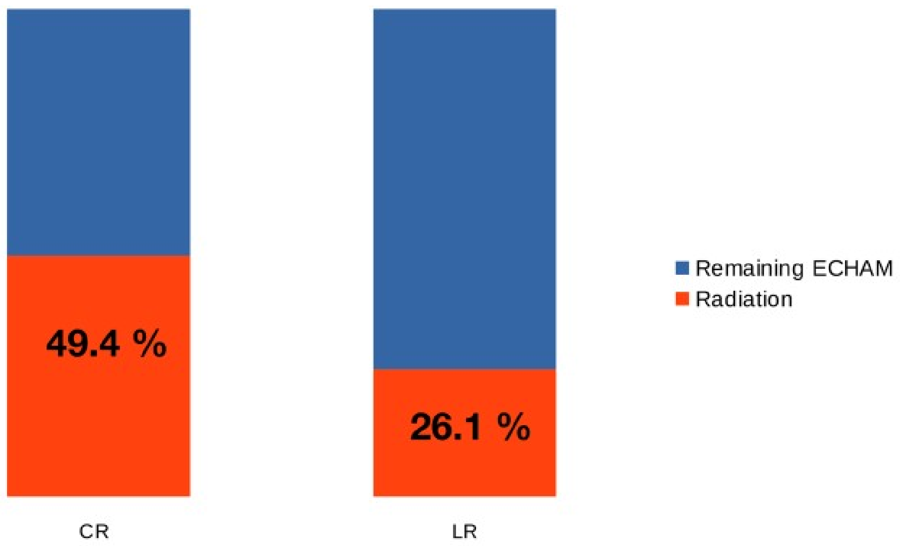
The art of approximation

3D radiation
(real atmosphere)



Weather/
climate model

The art of approximation

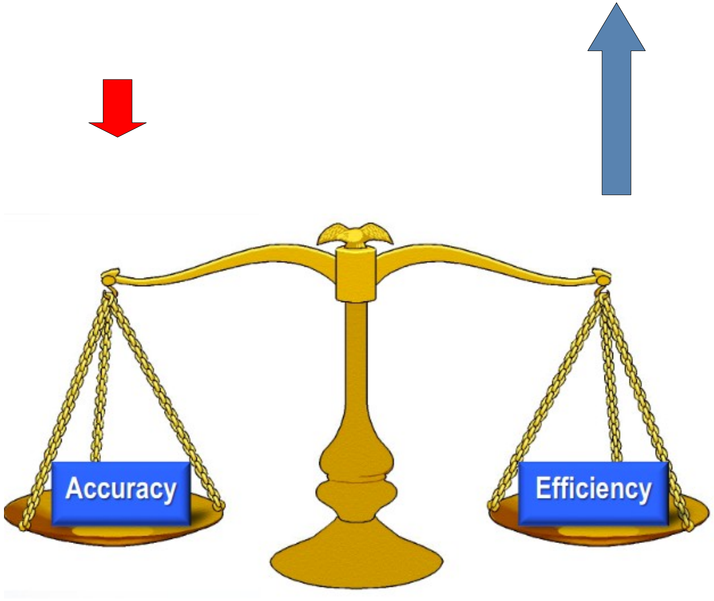


Coarse-res GCM (~460 km)

Low-res (170 km)

- Radiative transfer is an expensive component in coarse-resolution simulations especially
- This is despite using many approximations
- In the IFS, only a few % of model runtime, but radiation is called on a coarser grid and only every hour
- Since atmospheric radiation drives weather and climate, approximations and infrequent computations are consequential
- → **accuracy/speed trade-off** is important and should be improved

ML to the rescue?

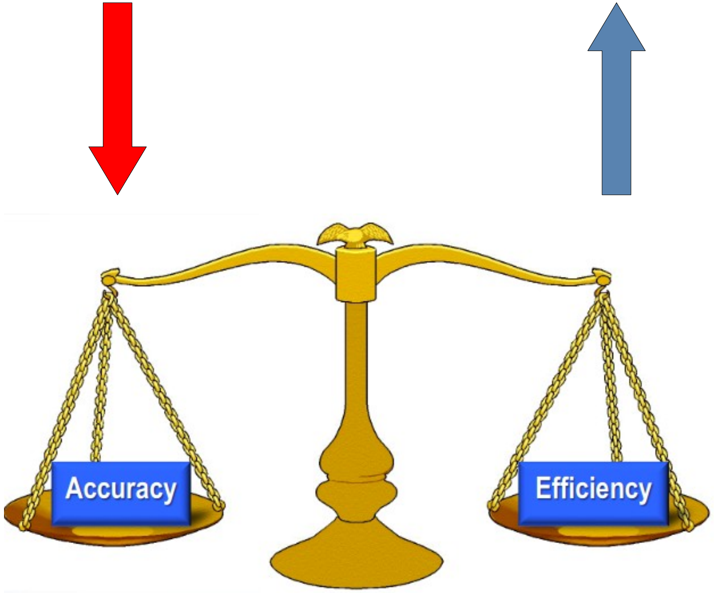


Key question:

Can machine learning actually "improve" the trade-off between accuracy and efficiency for radiation?

ML to the rescue?

Danmarks Meteorologiske Institut



Key question:

Can machine learning actually “improve” the trade-off between accuracy and efficiency for radiation?

Attempts so far using dense networks have given big speedups but at large costs in accuracy and generalization

How might machine learning be used for parametrised physics?

<u>Emulate existing model component</u>

Learn an operational scheme
Reduce computational cost
Port to GPUs
TL/Ad (see later)

<u>Examples</u>
Chevallier (Radiation 1990!)
Krasnoposky (Radiation + more)
Song & Roh (Radiation)
Chantry (NOGWD)
Espinosa (NOGWD)

<u>Emulate increased complexity model component</u>

Learn an unaffordable scheme
Reduce computational cost
…

<u>Examples</u>
Meyer (Radiation)
Gettelman (Cloud microphysics)

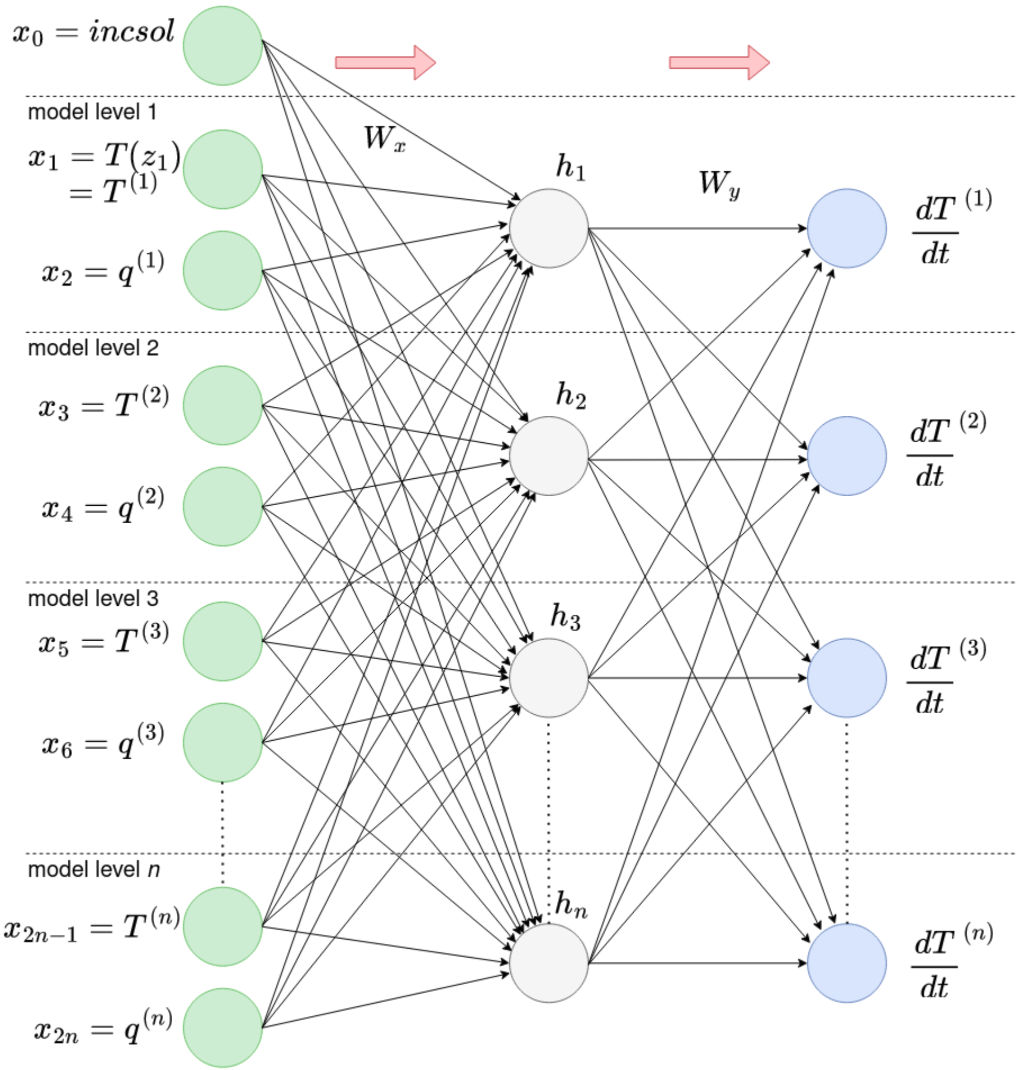
<u>Learn new parametrisation scheme</u>

Use data from high resolution simulations or observations
Greater challenges for model stability

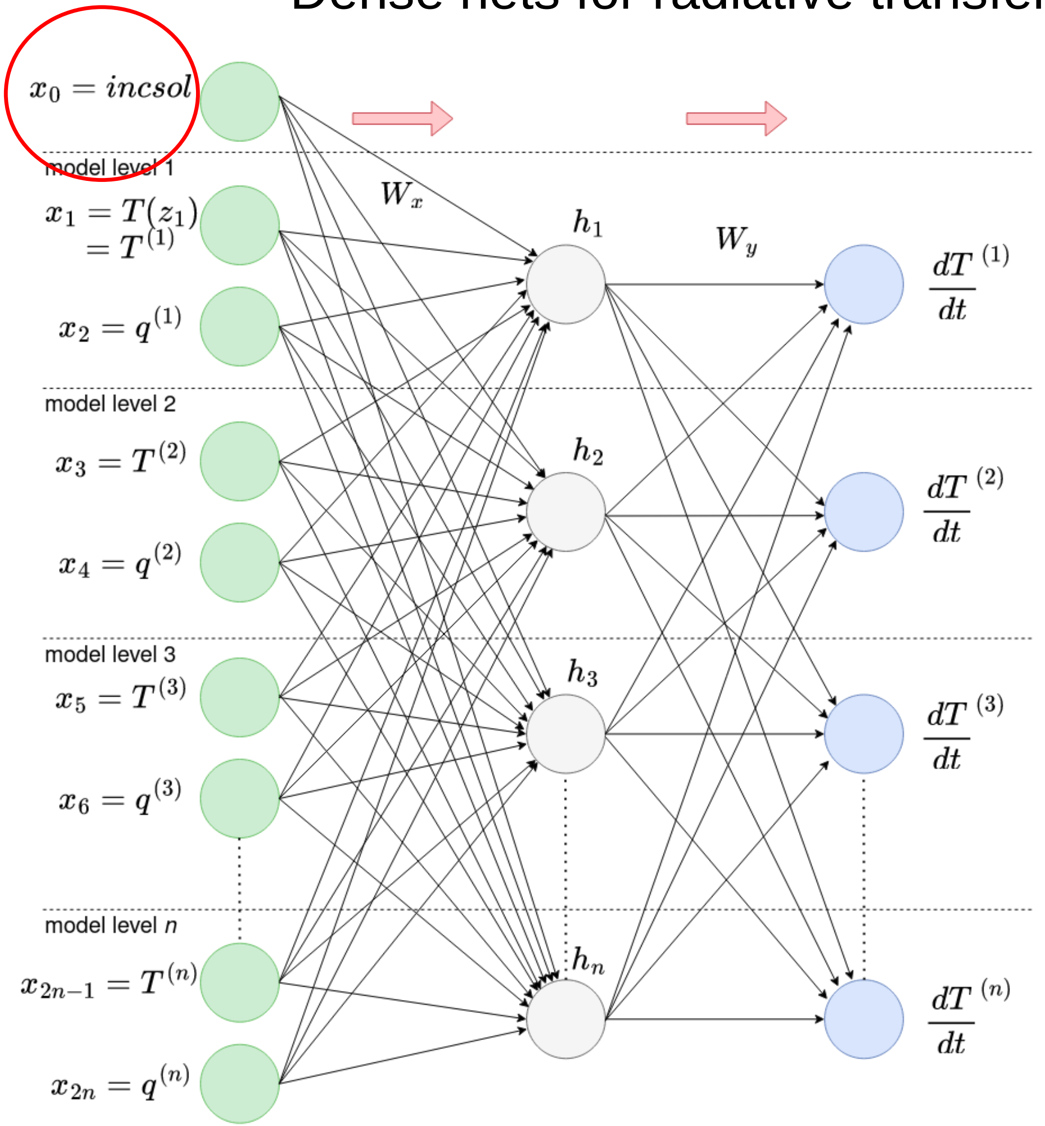
<u>Examples</u>
Yuval & O’Gorman (Convection, subgrid momentum)
Brenowitz & Bretherton (Radiation, convection, etc)
Beucler, Pritchard, Gentine, Rasp (Convection)

Slide by Matthew Chantry (ECMWF Annual Seminar 2022)

Dense nets for radiative transfer – the problem



Dense nets for radiative transfer – the problem



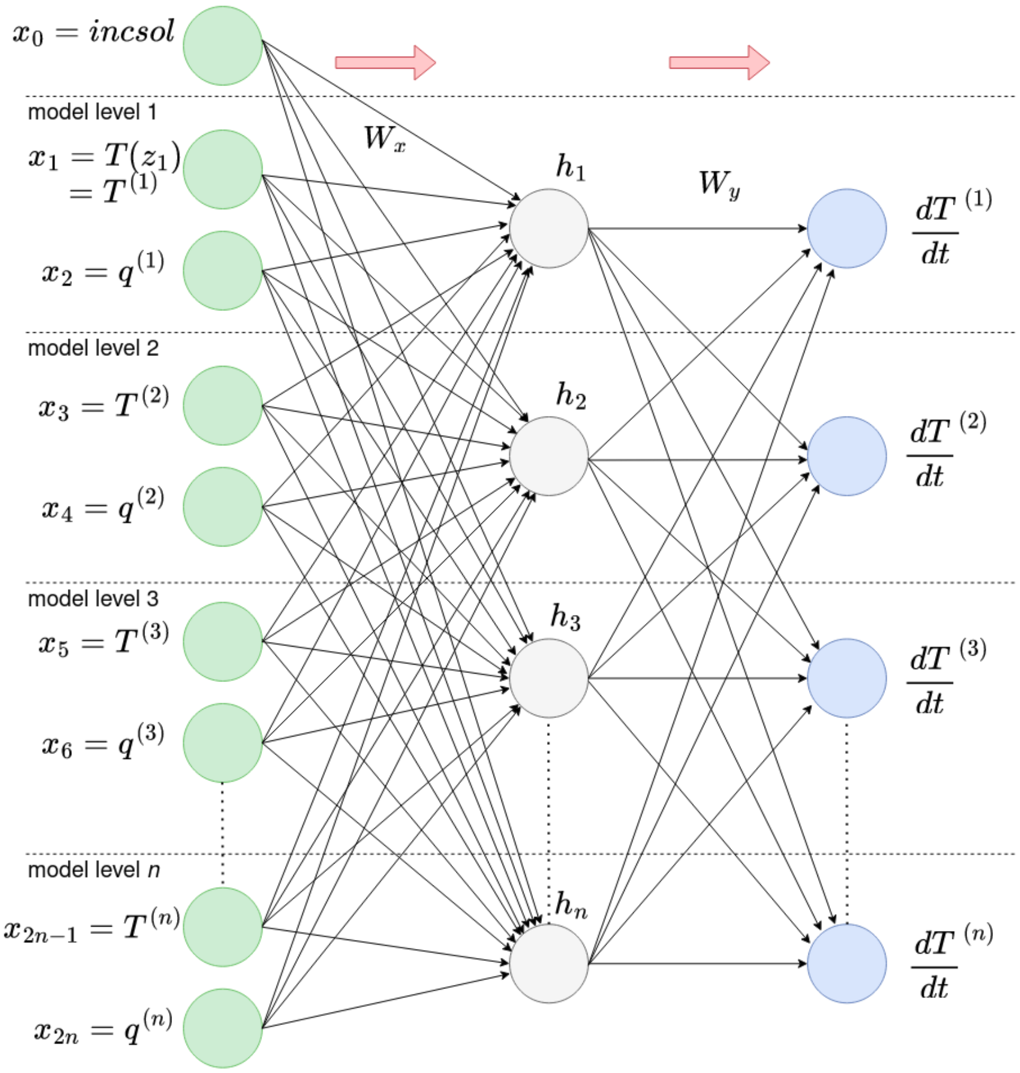
Inputs are profiles of pressure, temperature, gases, cloud water and ice, and a few scalar variables such as incident solar radiation (shortwave only)

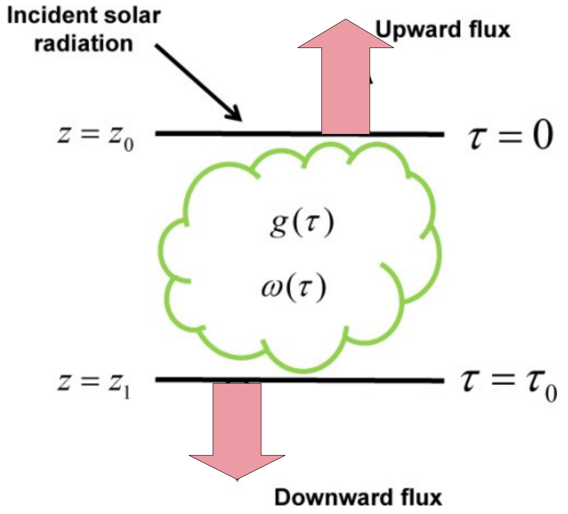
Outputs are profiles of **heating rates (HR) = dT/dt**

Radiation codes compute HR from upward and downward **fluxes,** but this approach gives noisy heating rates with dense NNs, so typically the outputs are HRs + surface and top-of-atmosphere fluxes

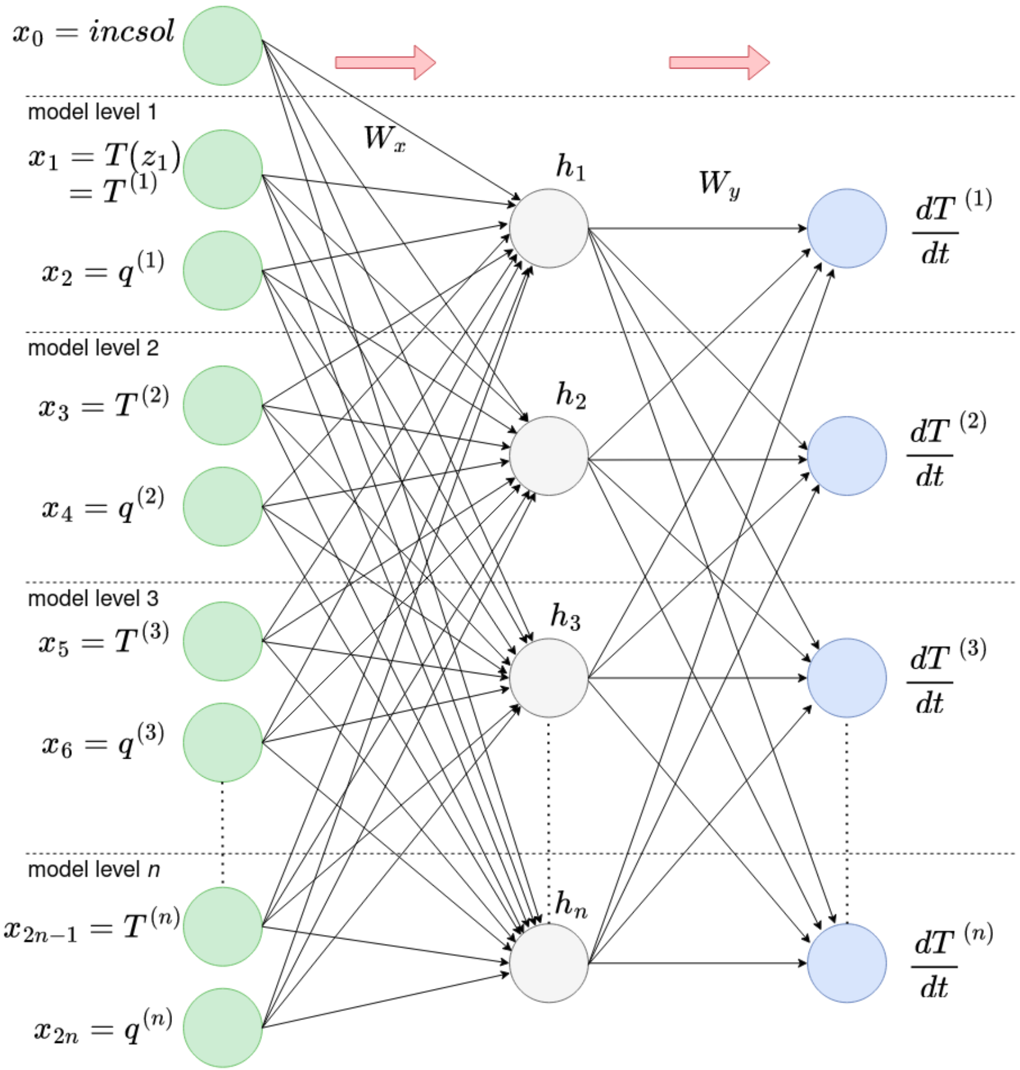
→ better estimate of HR but breaks energy conservation

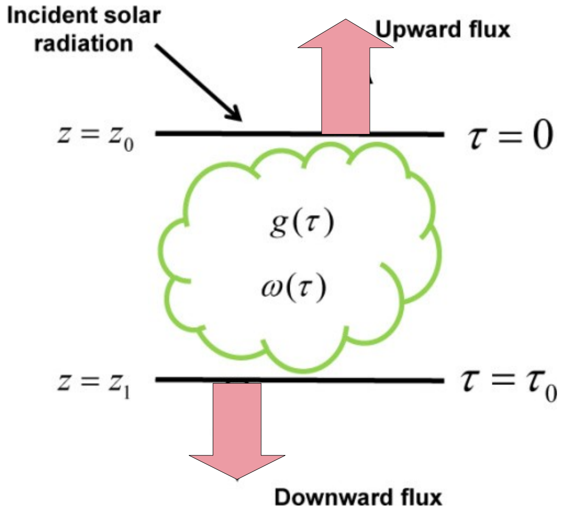
Dense nets for radiative transfer – the problem





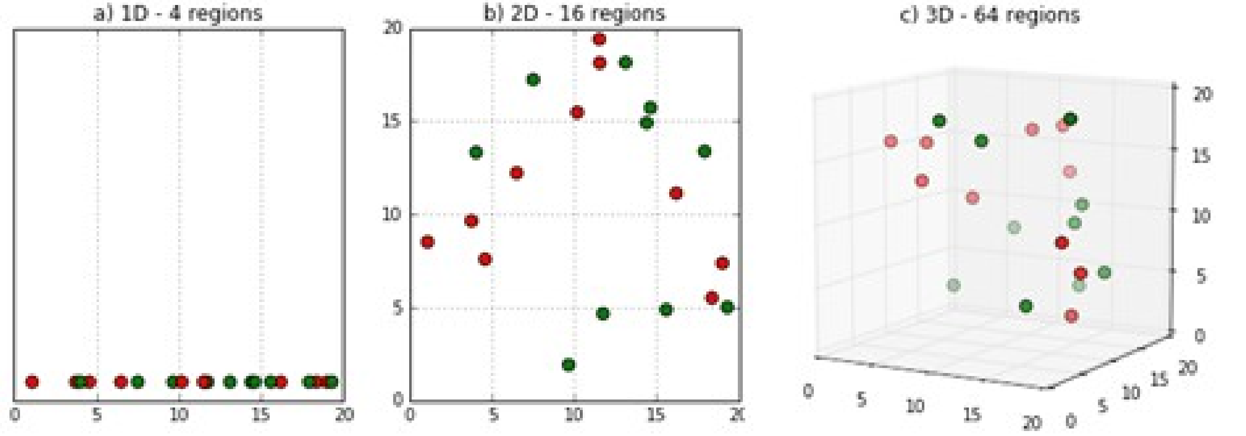
Dense nets for radiative transfer – the problem





Mismatch in the **direction of information flow** between the model and the process!

Curse of dimensionality

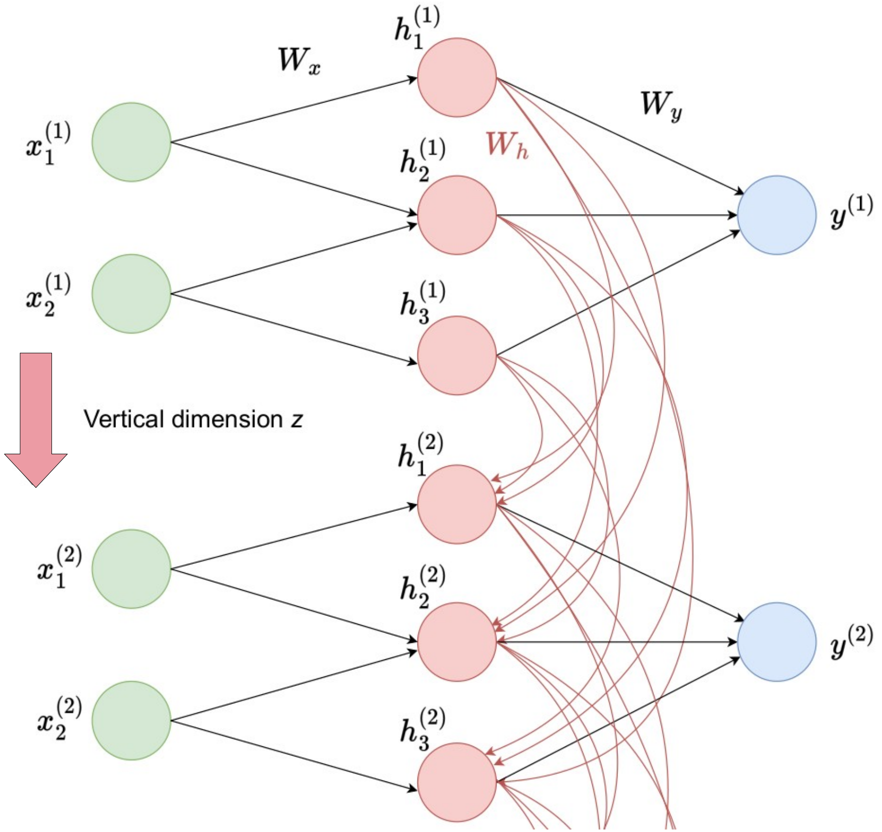


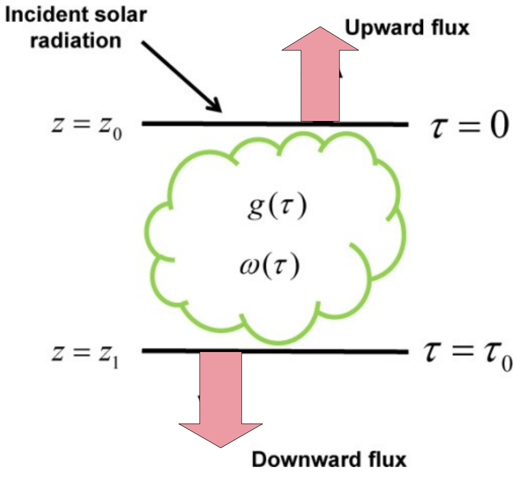
Given 137 levels in model column, 6 level-wise inputs:

DNN-based emulator has 137*6=822 inputs, weight matrix of input layer is BIG! (822 × num_neurons)

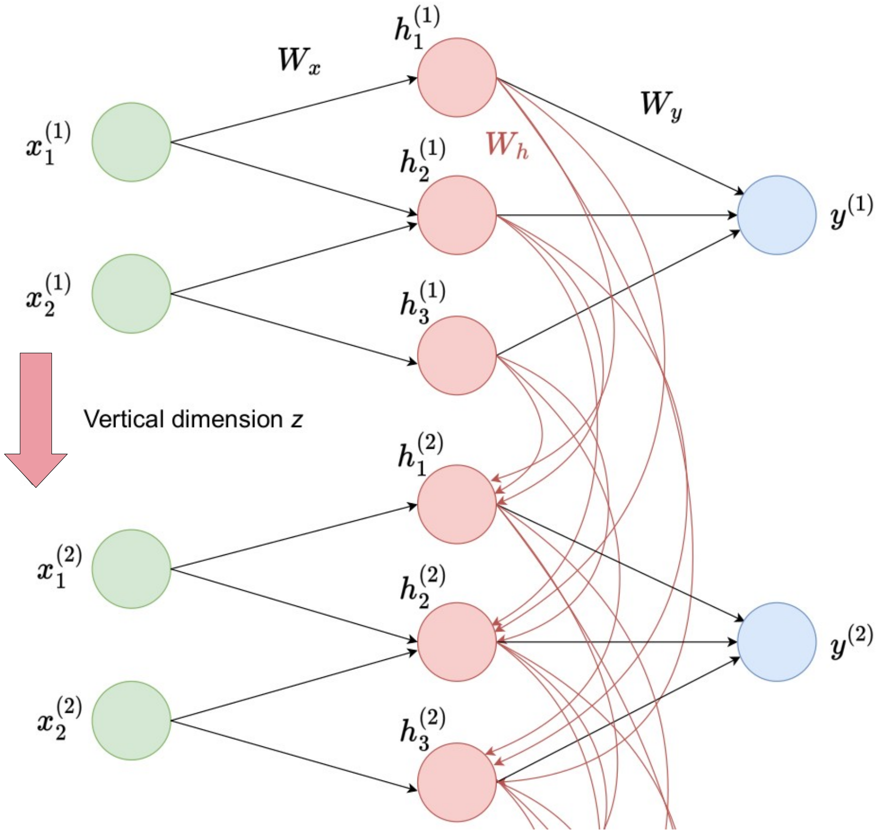
Recurrent neural networks for radiation (the solution?)

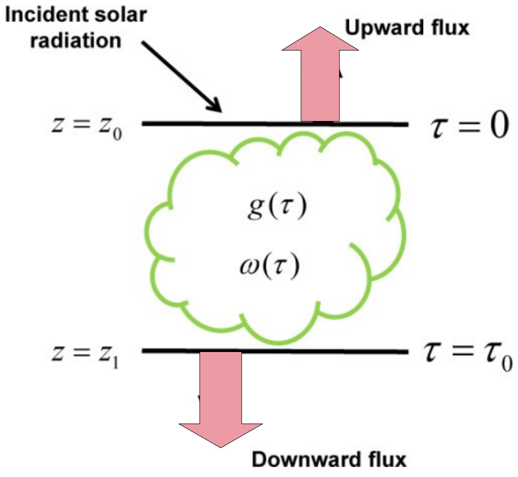
Danmarks Meteorologiske Institut





Recurrent neural networks for radiation (the solution?)





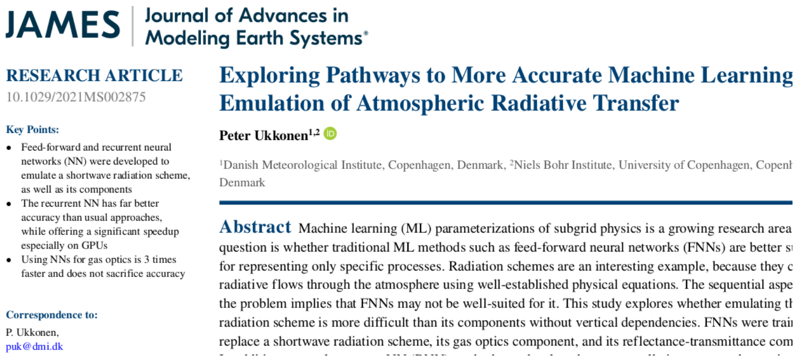
Characteristics of atmospheric radiative transfer respected by RNNs :

- **Correct directionality**, however radiation flows both upward and downward, so we need bidirectional RNNs (BiRNN)!
- **Sequential** from one level to the next – unlike DNN, which (unphysically) connects the top directly to the surface
- **Invariable** physical laws by height – unlike DNN, which (unphysically) uses level-specific weights

RNNs for shortwave radiation (JAMES 2022)

Danmarks Meteorologiske Institut

- RNNs were introduced for atmospheric radiation in Ukkonen (2022), where they were compared to other emulation approaches for SW radiation (including dense nets).
- Training data was generated offline using the RTE+RRTMGP scheme, using global CAMS data and including clouds, but not aerosols


JAMES Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2021MS002875

Key Points:
- Feed-forward and recurrent neural networks (NN) were developed to emulate a shortwave radiation scheme, as well as its components
- The recurrent NN has far better accuracy than usual approaches, while offering a significant speedup especially on GPUs
- Using NNs for gas optics is 3 times faster and does not sacrifice accuracy

Correspondence to:
P. Ukkonen,
puk@dmi.dk

Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer

Peter Ukkonen[1,2]

[1]Danish Meteorological Institute, Copenhagen, Denmark, [2]Niels Bohr Institute, University of Copenh
Denmark

Abstract Machine learning (ML) parameterizations of subgrid physics is a growing research area
question is whether traditional ML methods such as feed-forward neural networks (FNNs) are better s
for representing only specific processes. Radiation schemes are an interesting example, because they c
radiative flows through the atmosphere using well-established physical equations. The sequential aspe
the problem implies that FNNs may not be well-suited for it. This study explores whether emulating th
radiation scheme is more difficult than its components without vertical dependencies. FNNs were trai
replace a shortwave radiation scheme, its gas optics component, and its reflectance-transmittance com

RNNs for shortwave radiation (JAMES 2022)

Danmarks Meteorologiske Institut

- RNNs were introduced for atmospheric radiation in Ukkonen (2022), where they were compared to other emulation approaches for SW radiation (including dense nets).
- Training data was generated offline using the RTE+RRTMGP scheme, using global CAMS data and including clouds, but not aerosols
- To ensure energy conservation, the NNs predict full flux profiles, from which heating rates are computed. A hybrid loss function is used to reduce HR errors
- For shortwave, a helpful trick is to normalize all the fluxes by the incoming solar radiation, so outputs range from 0..1 (physical scaling)

JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE
10.1029/2021MS002875

Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer

Peter Ukkonen[1,2]

[1]Danish Meteorological Institute, Copenhagen, Denmark, [2]Niels Bohr Institute, University of Copenhagen, Copenh
Denmark

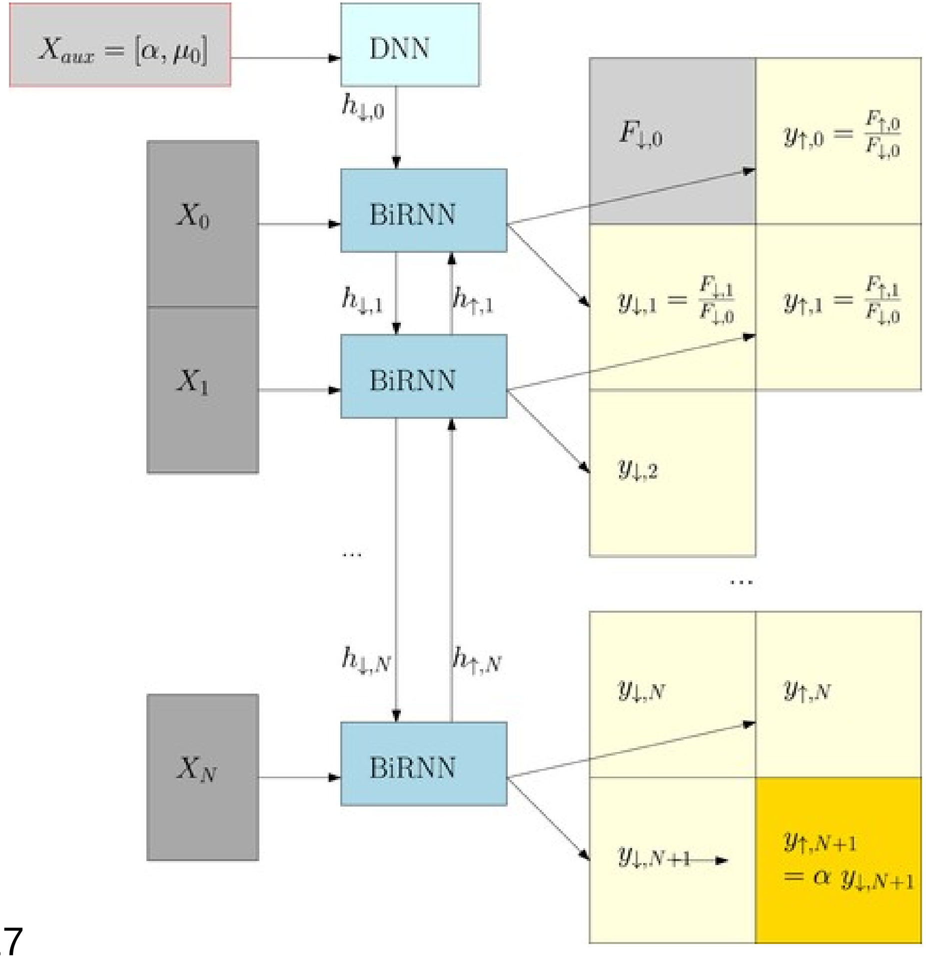
Key Points:

- Feed-forward and recurrent neural networks (NN) were developed to emulate a shortwave radiation scheme, as well as its components
- The recurrent NN has far better accuracy than usual approaches, while offering a significant speedup especially on GPUs
- Using NNs for gas optics is 3 times faster and does not sacrifice accuracy

Correspondence to:
P. Ukkonen,
puk@dmi.dk

Abstract Machine learning (ML) parameterizations of subgrid physics is a growing research area
question is whether traditional ML methods such as feed-forward neural networks (FNNs) are better s
for representing only specific processes. Radiation schemes are an interesting example, because they c
radiative flows through the atmosphere using well-established physical equations. The sequential aspe
the problem implies that FNNs may not be well-suited for it. This study explores whether emulating th
radiation scheme is more difficult than its components without vertical dependencies. FNNs were trai
replace a shortwave radiation scheme, its gas optics component, and its reflectance-transmittance com

RNNs for shortwave radiation (JAMES 2022)



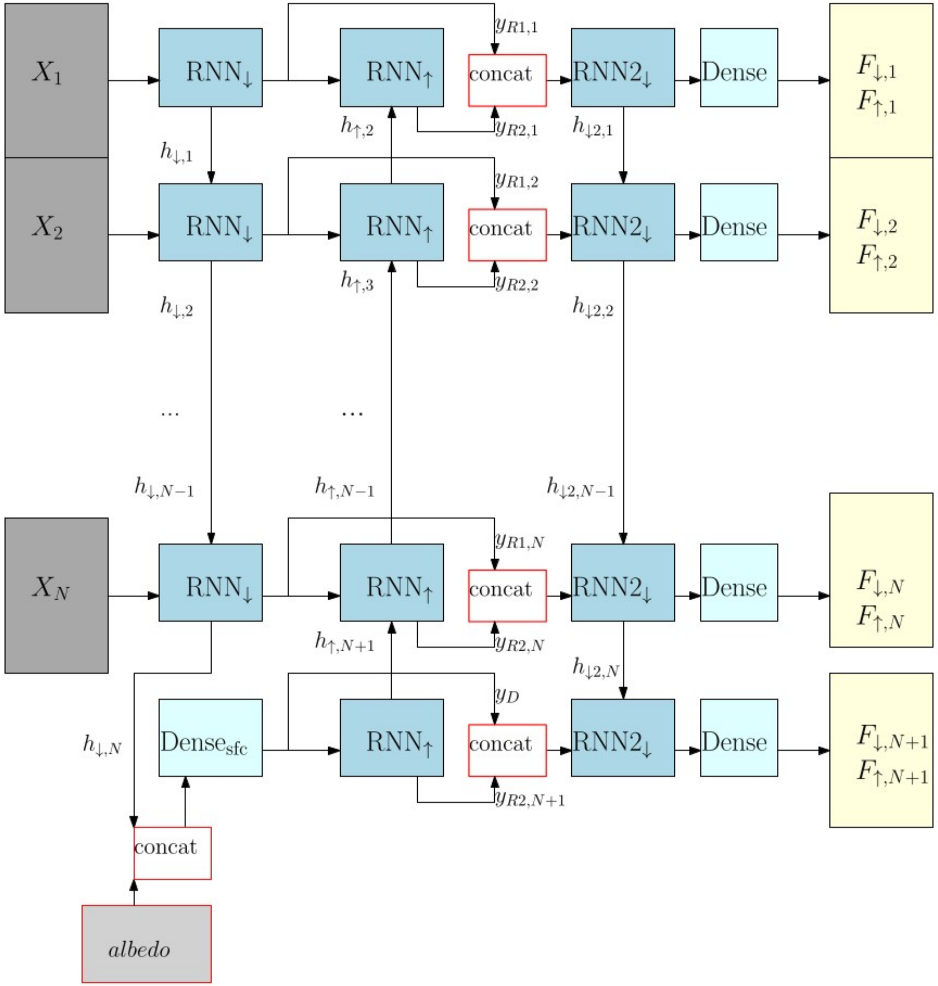
First attempt (simple approach)

A BiRNN iterates through *N* level-wise inputs and predicts the upward flux above and downward flux below that level

Works OK but has problems:

- The albedo **α** was incorporated to the BiRNN thorugh a DNN that predicts the initial state, but physically in the wrong place (top instead of surface)
- Upward flux at surface, F↑ (*N*+1) not predicted but computed as $F\downarrow_{pred}$ (*N*+1) × α
- Introduces a discontinuity at the surface and ignores any spectral variation of albedo

RNNs for shortwave radiation (JAMES 2022)

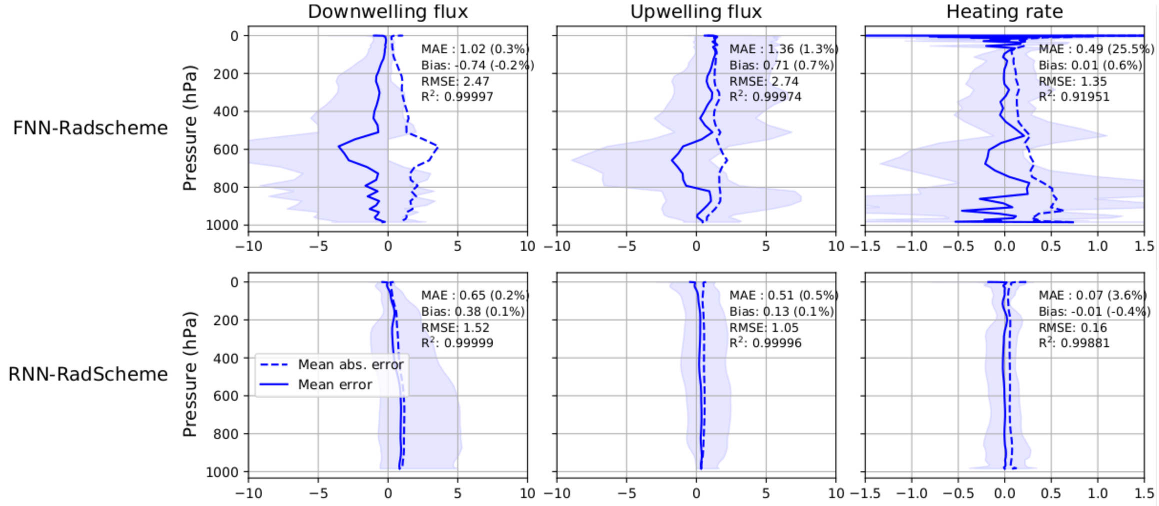


A better model by attempting to mimic the equations

- Three iterations (down, up and down again) as in the RTE shortwave solver
- From inputs defined at *N* levels (layers) to fluxes at *N+1* half-levels by concatenating the first RNN sequence with the output of a “surface” DNN – corresponds to how surface albedo is concatenated with level albedos
- Looks complicated but model complexity is low: **final model used only 5600 parameters** (3x GRUs with 8 neurons)

RNNs for shortwave radiation (JAMES 2022)

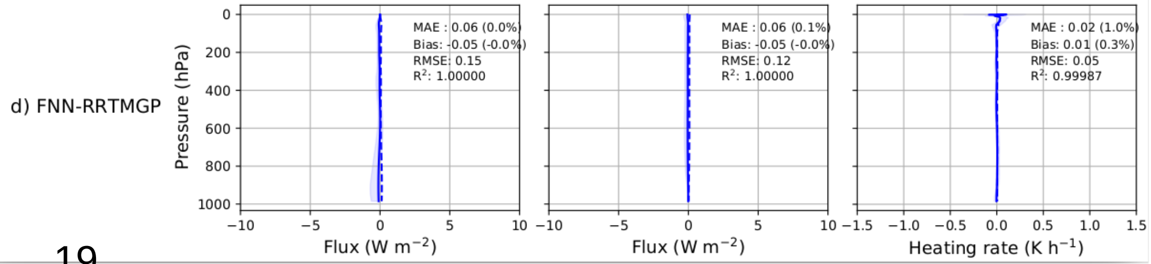
Danmarks Meteorologiske Institut



Dense networks:
RMSE 1.35 K / day
100,000 parameters
~50x speedup

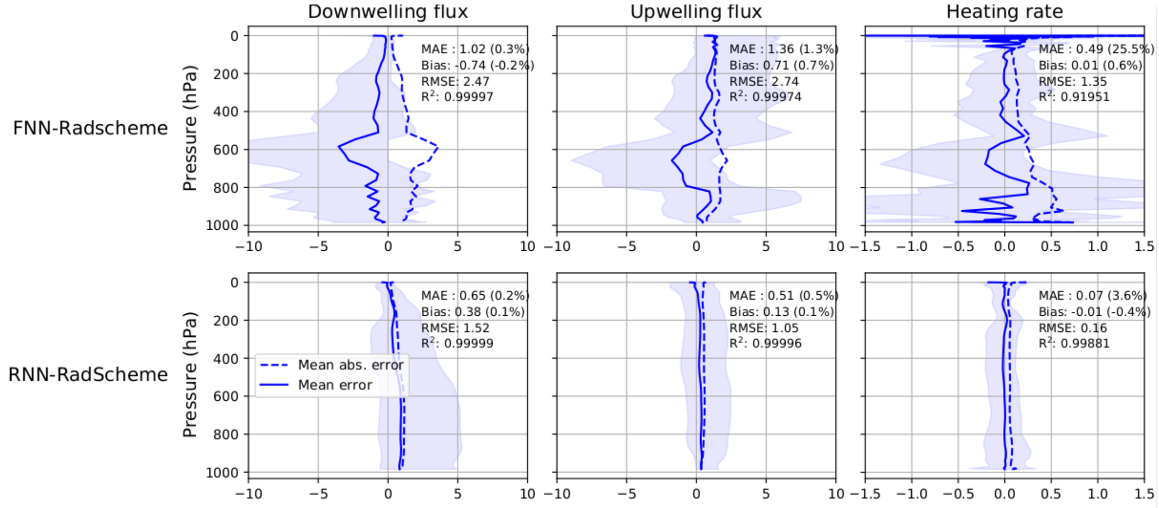
RNNs:
RMSE 0.16 K / day
5600 parameters
~5x speedup

! speed-ups are on CPU and relative to a modern but somewhat expensive radiation scheme with high spectral resolution (RTE+RRTMGP)



NNs only for predicting optical properties:
RMSE 0.05 K / day
4200 parameters
~1.3x speedup, but also better generalization and flexibility

RNNs for shortwave radiation (JAMES 2022)

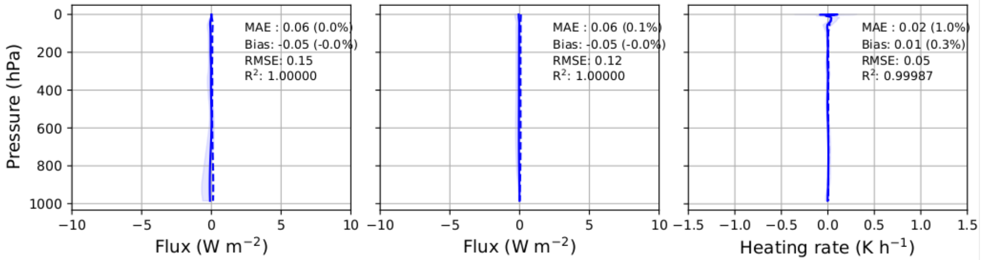


Dense networks:
RMSE 1.35 K / day
100,000 parameters
~50x speedup

RNNs:
RMSE 0.16 K / day
5600 parameters
~5x speedup

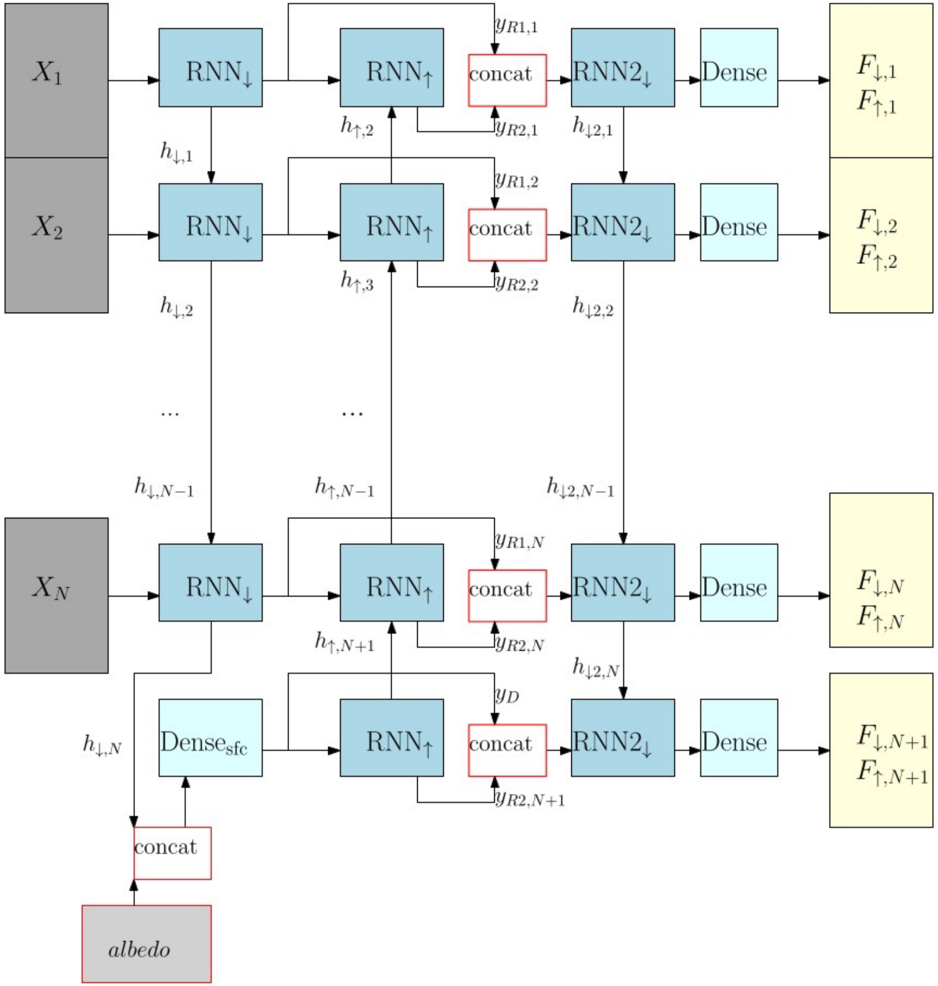
Dense networks produce noisy fluxes, which leads to inaccurate heating rates!

$$\left(\frac{dT}{dt}\right)_{\text{SW radiation}} = -\frac{g}{c_p}\frac{F_{i+1/2,\text{SW}} - F_{i-1/2,\text{SW}}}{p_{i+1/2} - p_{i-1/2}},$$



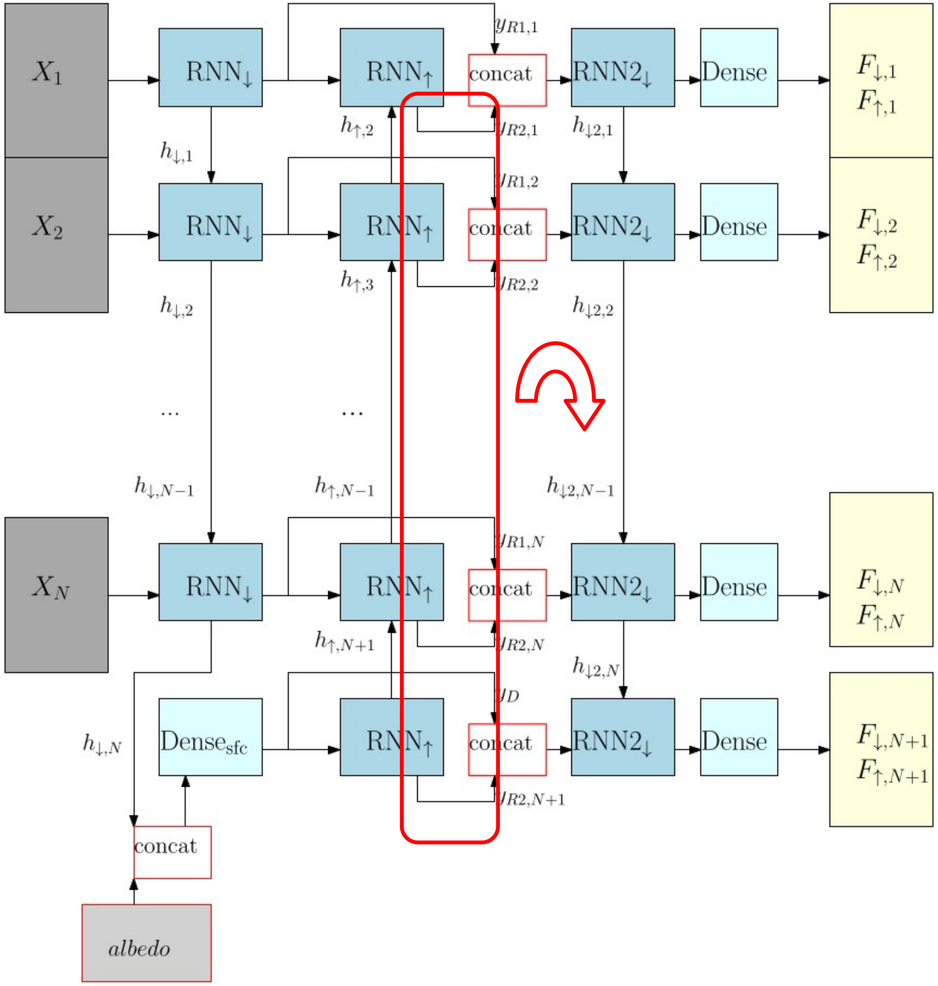
NNs only for predicting optical properties:
RMSE 0.05 K / day
4200 parameters
~1.3x speedup, but also better generalization and flexibility

A mistake



Need for three iterations not so intuitive. In practice, I used three because it gave better results than two.

A mistake

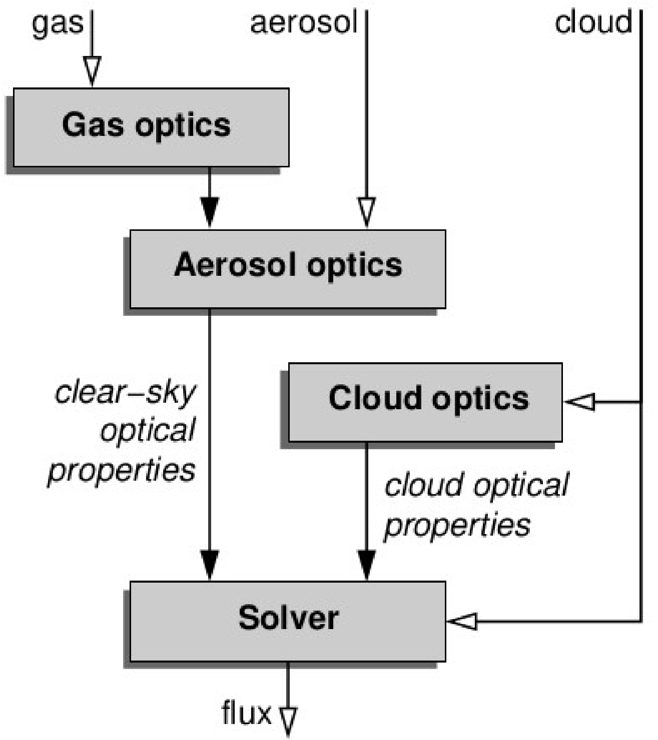


Need for three iterations not so intuitive. In practice, I used three because it gave better results than two.

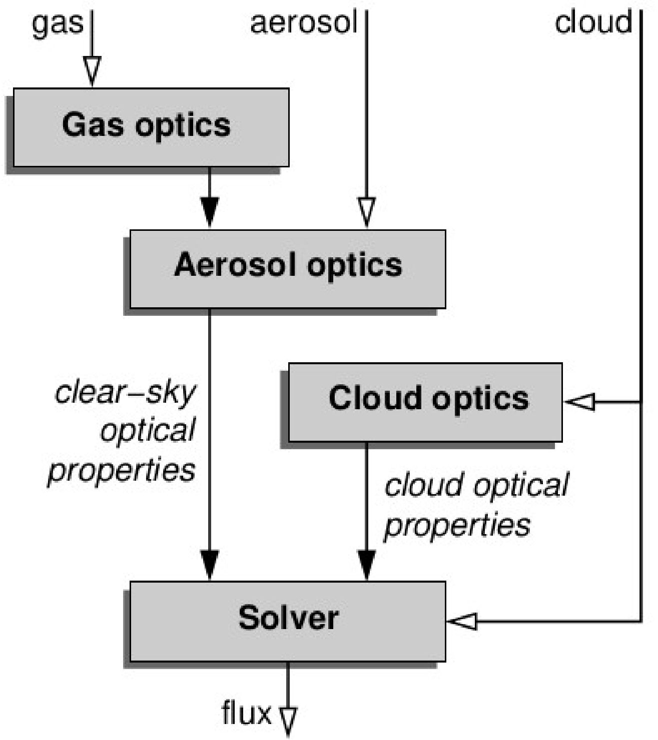
However, there was a mistake in the code: Keras apparently requires the output of RNNs with “backward=true” to be manually reversed, which wasn’t done in the paper.

Oops.

A closer look at radiation schemes (ecRad)



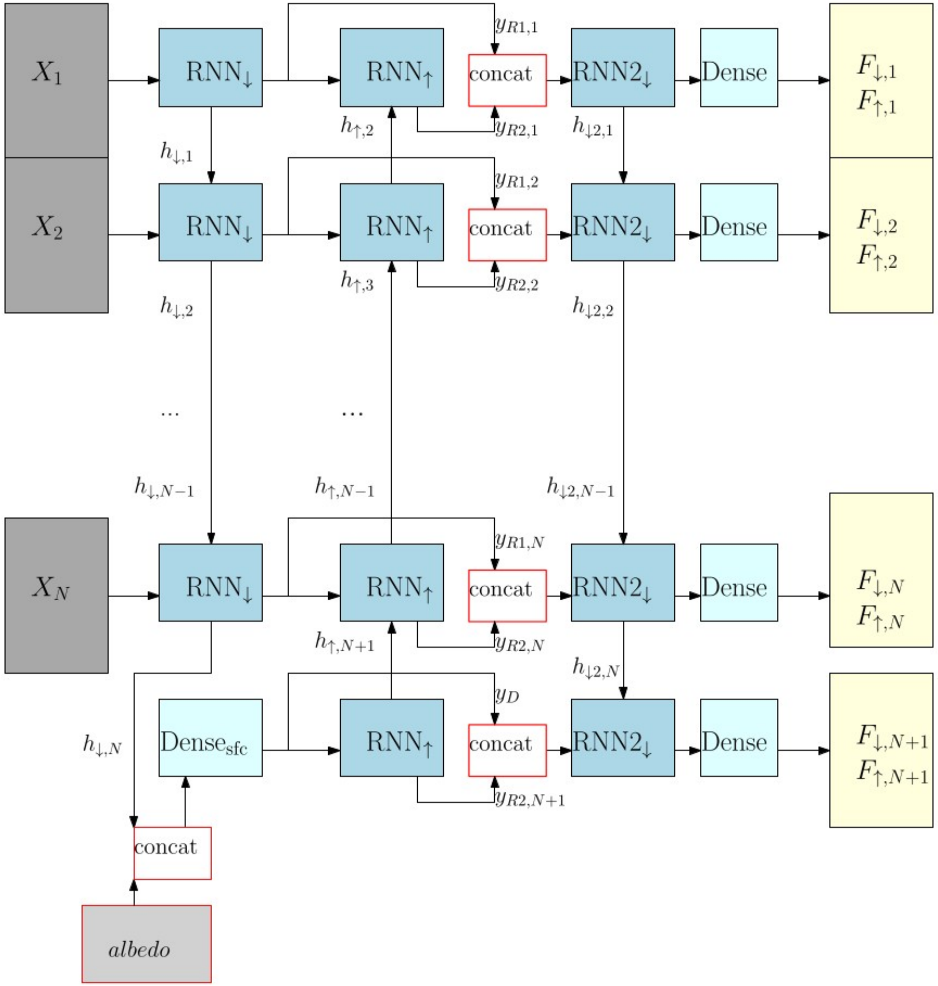
A closer look at radiation schemes (ecRad)



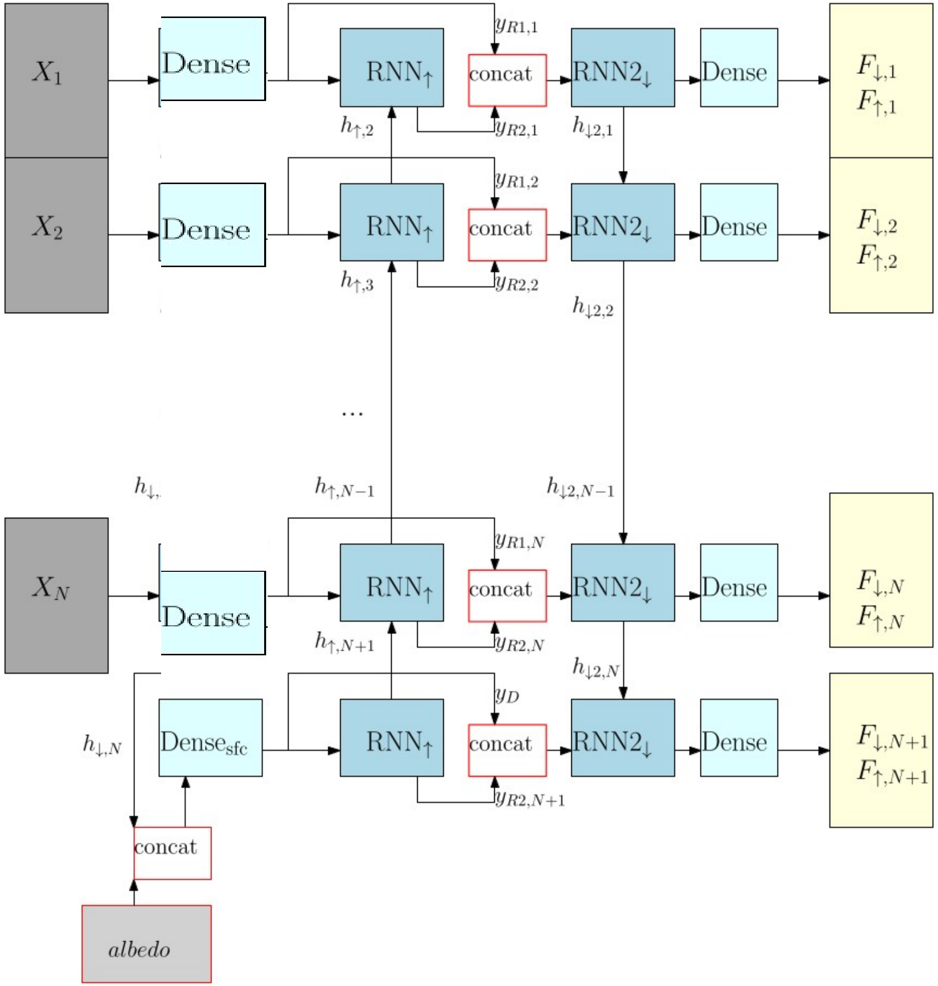
1. Compute layer-wise optical properties such as optical depth given gas concentrations, clouds, aerosols **(no vertical dependencies)**

2. The radiative transfer solver takes the spectrally defined optical properties, cloud overlap assumptions, etc, and:

1) Compute layer-wise reflectances and transmittances **(no vertical dependencies)**
2) Starting at the surface, **iterate upwards to compute total albedos** and sources (LW only)
3) Starting at TOA, **iterate downwards to compute fluxes** (spectral, then average for broadband flux)



X_1
X_2
X_N
$\mathrm{RNN}_\downarrow$
$\mathrm{RNN}_\uparrow$
concat
$\mathrm{RNN2}_\downarrow$
Dense
$F_{\downarrow,1}$ $F_{\uparrow,1}$
$F_{\downarrow,2}$ $F_{\uparrow,2}$
$F_{\downarrow,N}$ $F_{\uparrow,N}$
$F_{\downarrow,N+1}$ $F_{\uparrow,N+1}$
$y_{R1,1}$
$y_{R2,1}$
$y_{R1,2}$
$y_{R2,2}$
$y_{R1,N}$
$y_{R2,N}$
y_D
$y_{R2,N+1}$
$h_{\downarrow,1}$
$h_{\downarrow,2}$
$h_{\downarrow,N-1}$
$h_{\downarrow,N}$
$h_{\uparrow,2}$
$h_{\uparrow,3}$
$h_{\uparrow,N-1}$
$h_{\uparrow,N+1}$
$h_{\downarrow 2,1}$
$h_{\downarrow 2,2}$
$h_{\downarrow 2,N-1}$
$h_{\downarrow 2,N}$
...
$\mathrm{Dense}_{\mathrm{sfc}}$
albedo



X_1
X_2
X_N
Dense
Dense
Dense
Dense$_{sfc}$
RNN$_\uparrow$
RNN$_\uparrow$
RNN$_\uparrow$
RNN$_\uparrow$
concat
concat
concat
concat
concat
RNN2$_\downarrow$
RNN2$_\downarrow$
RNN2$_\downarrow$
RNN2$_\downarrow$
Dense
Dense
Dense
Dense
$y_{R1,1}$
$y_{R2,1}$
$y_{R1,2}$
$y_{R2,2}$
$y_{R1,N}$
$y_{R2,N}$
y_D
$y_{R2,N+1}$
$h_{\uparrow,2}$
$h_{\uparrow,3}$
$h_{\uparrow,N-1}$
$h_{\uparrow,N+1}$
$h_{\downarrow 2,1}$
$h_{\downarrow 2,2}$
$h_{\downarrow 2,N-1}$
$h_{\downarrow 2,N}$
$h_{\downarrow,N}$
...
$F_{\downarrow,1}$ $F_{\uparrow,1}$
$F_{\downarrow,2}$ $F_{\uparrow,2}$
$F_{\downarrow,N}$ $F_{\uparrow,N}$
$F_{\downarrow,N+1}$ $F_{\uparrow,N+1}$
albedo

RNNs emulating ecRad, tested in the IFS

Work mainly by Matthew Chantry (ECMWF)

Danmarks Meteorologiske Institut

- RNNs were trained on the inputs and outputs of ecRad (TripleClouds solver) using a hybrid loss incorporating heating rate.
- Training - 2020, Evaluation – 2021
- IFS implementation / online inference by using Infero, a lower-level ML library developed at ECMWF that supports different back-ends

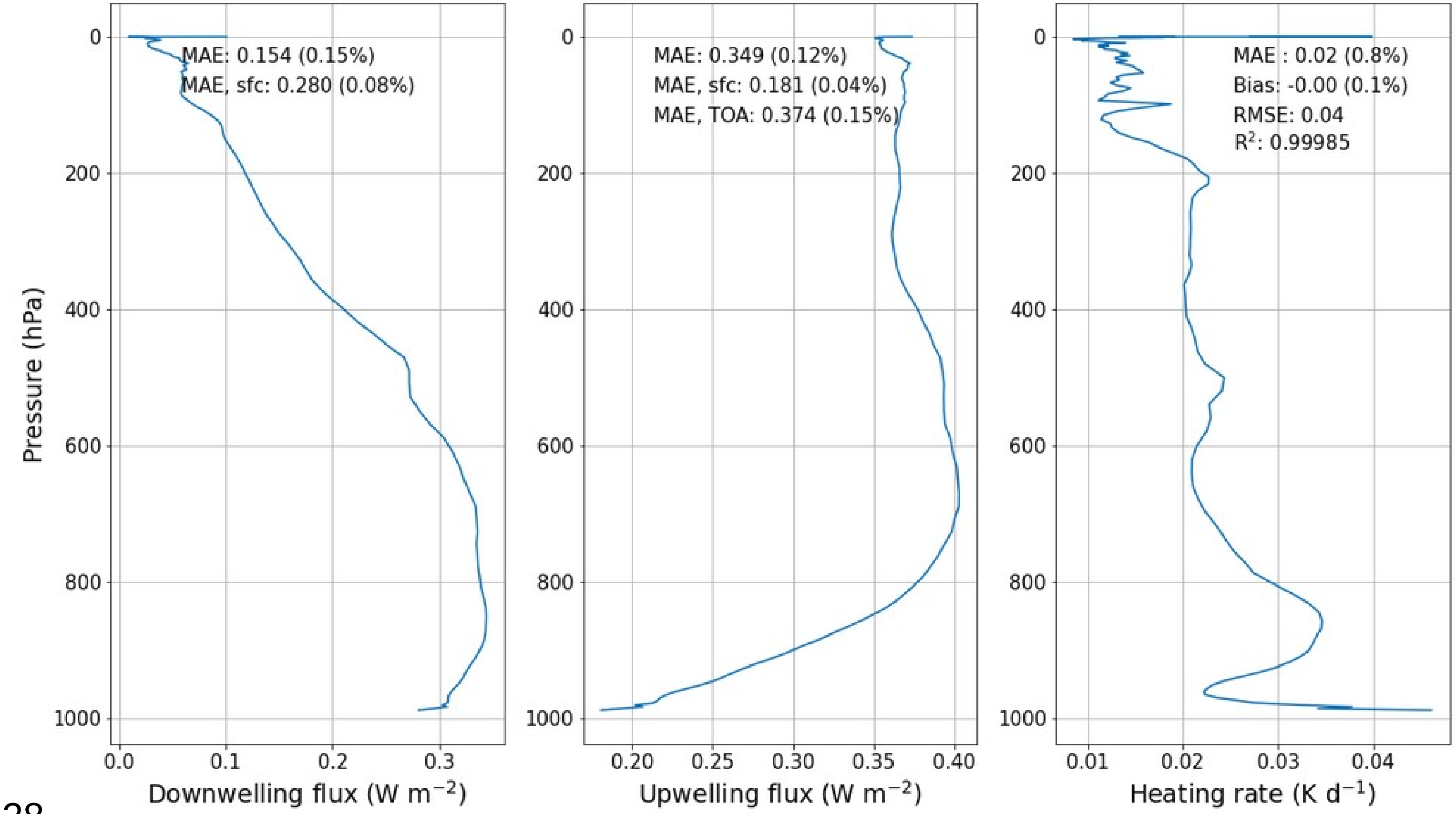
    github.com/ecmwf-projects/infero

RNNs emulating ecRad, tested in the IFS

Work mainly by Matthew Chantry (ECMWF)

Danmarks Meteorologiske Institut

Offline errors

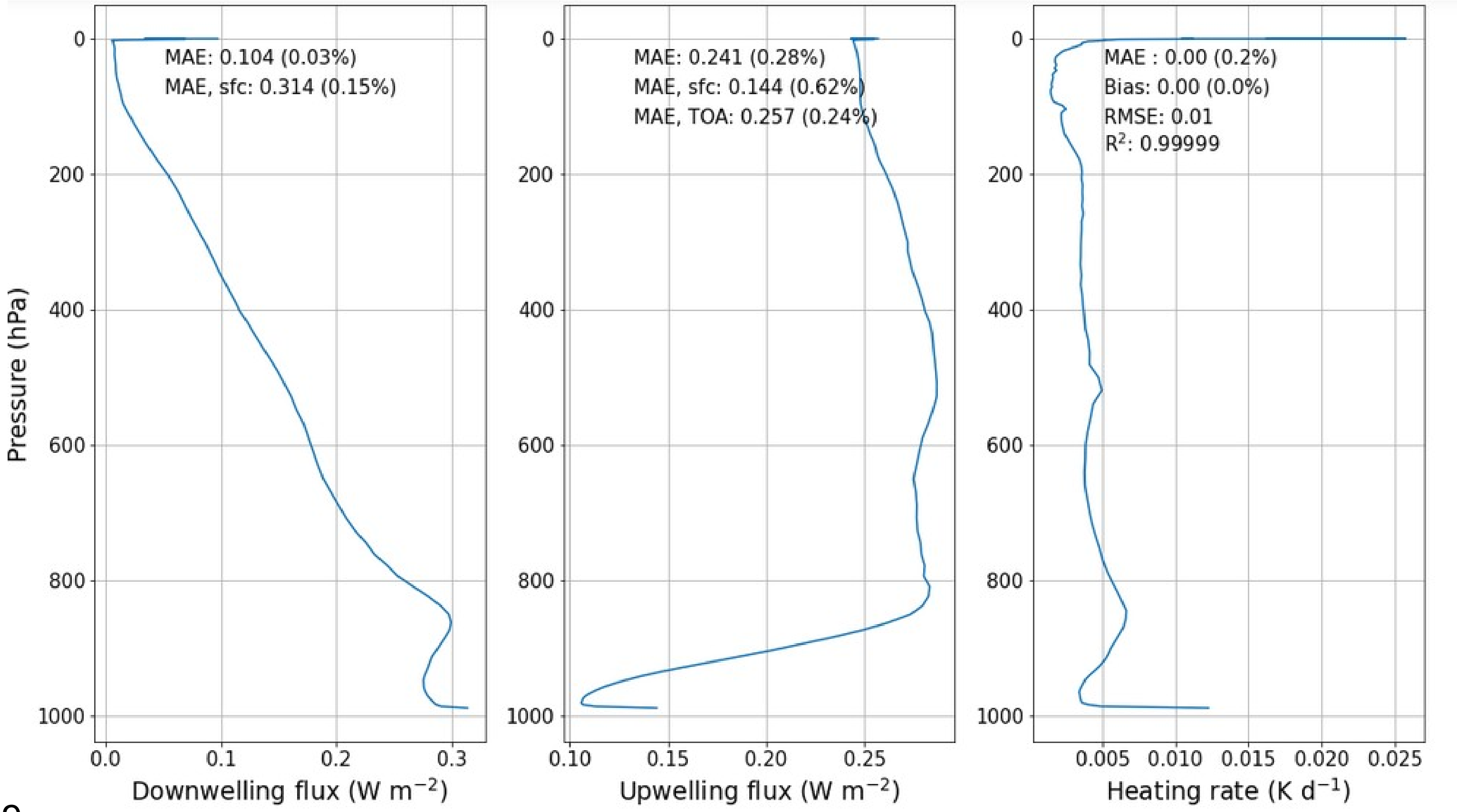


Longwave RNNs (64-neuron LSTMs)

RNNs emulating ecRad, tested in the IFS

Work mainly by Matthew Chantry (ECMWF)

Offline errors

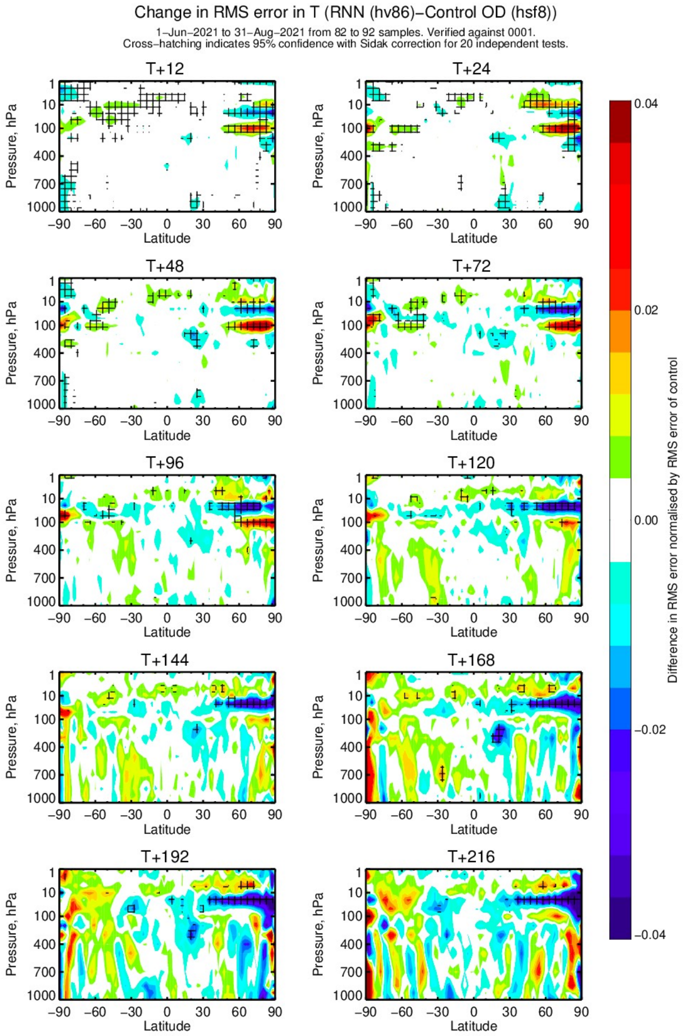


Shortwave RNNs (64-neuron LSTMs)

RNN vs TripleClouds

Plots show the change in RMSE in temperature using a suite of June-July-August IFS experiments at ~30km resolution

Red = degradation

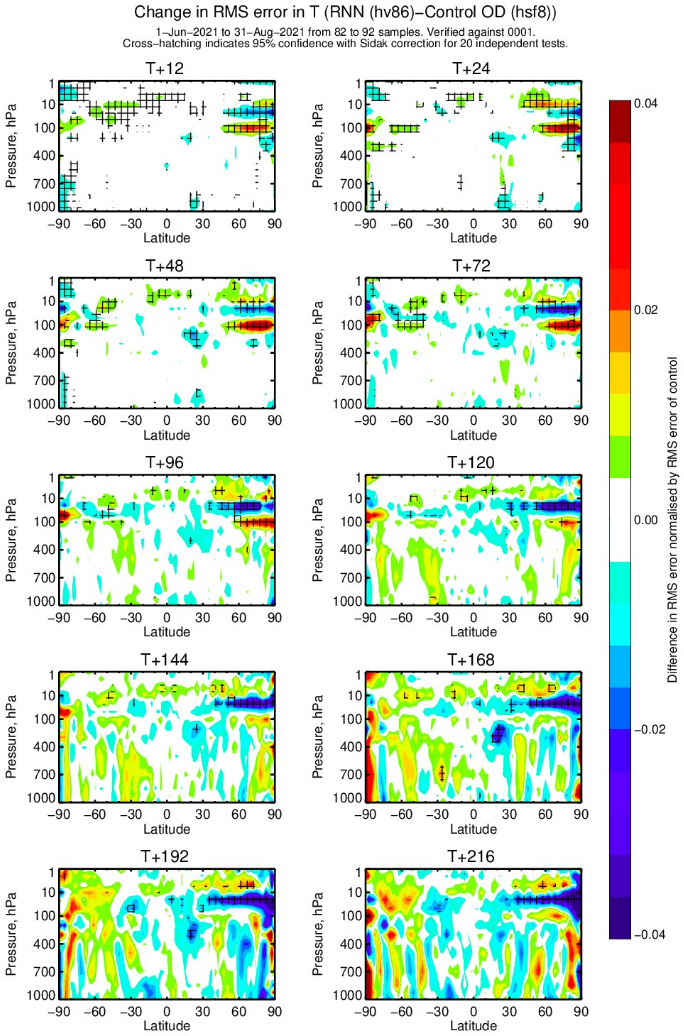


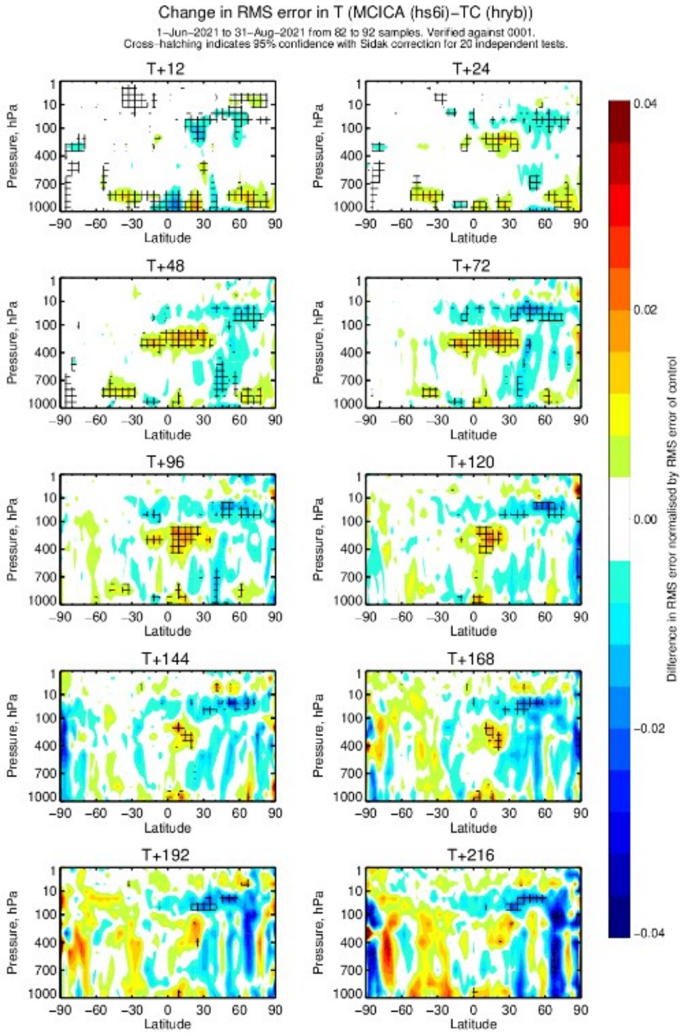
RNN vs TripleClouds

McICA vs TripleClouds

Plots show the change in RMSE in temperature using a suite of June-July-August IFS experiments at ~30km resolution

Red = degradation

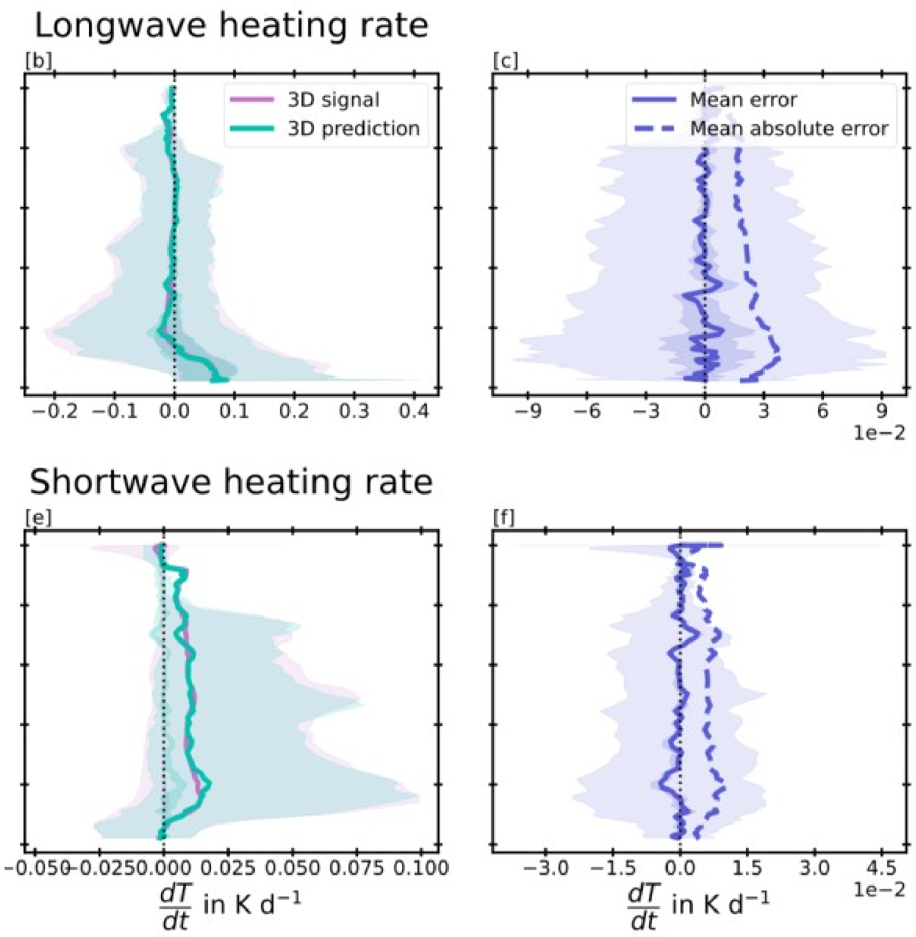




Emulation of 3D cloud radiative effects (Meyer et *al*. 2022)

Danmarks Meteorologiske Institut

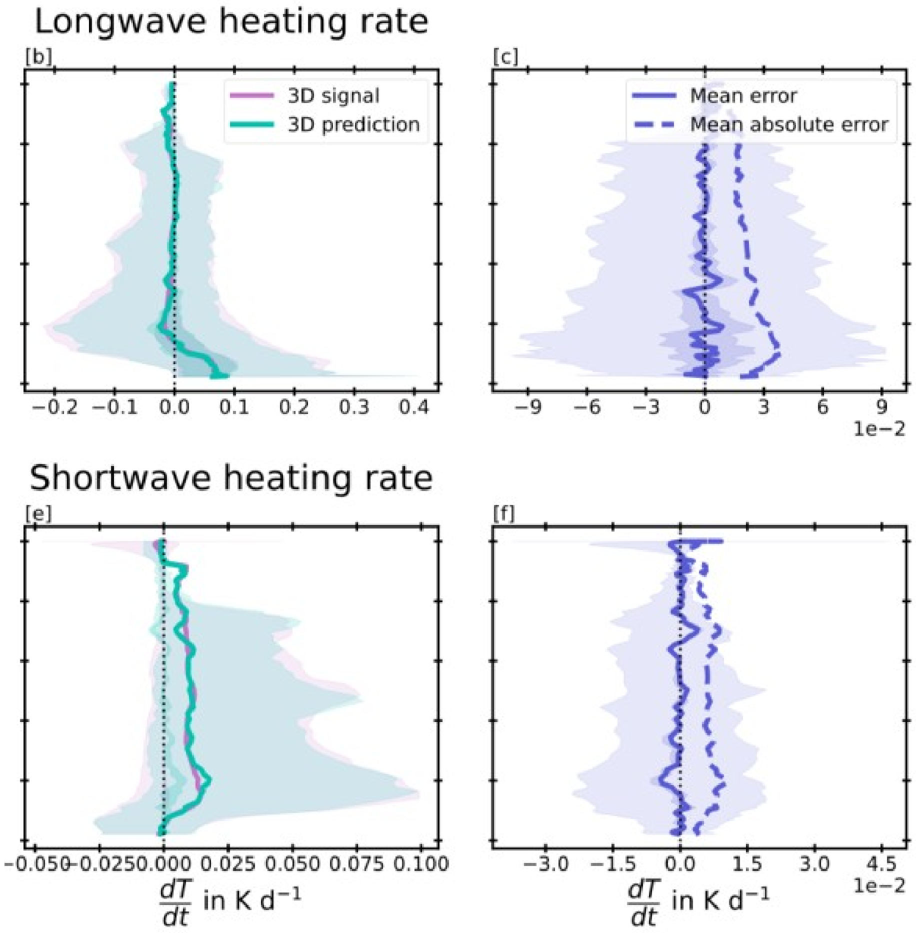
DNNs from paper, ~525k parameters each



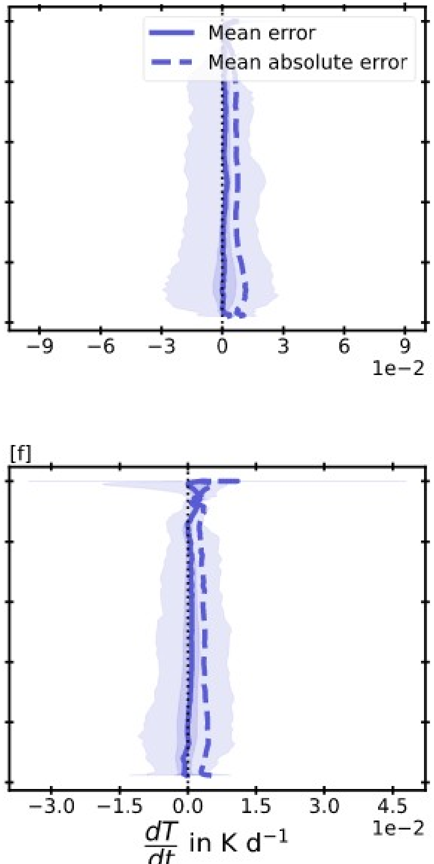
Emulation of 3D cloud radiative effects (Meyer et *al*. 2022)

Danmarks Meteorologiske Institut

DNNs from paper, ~525k parameters each



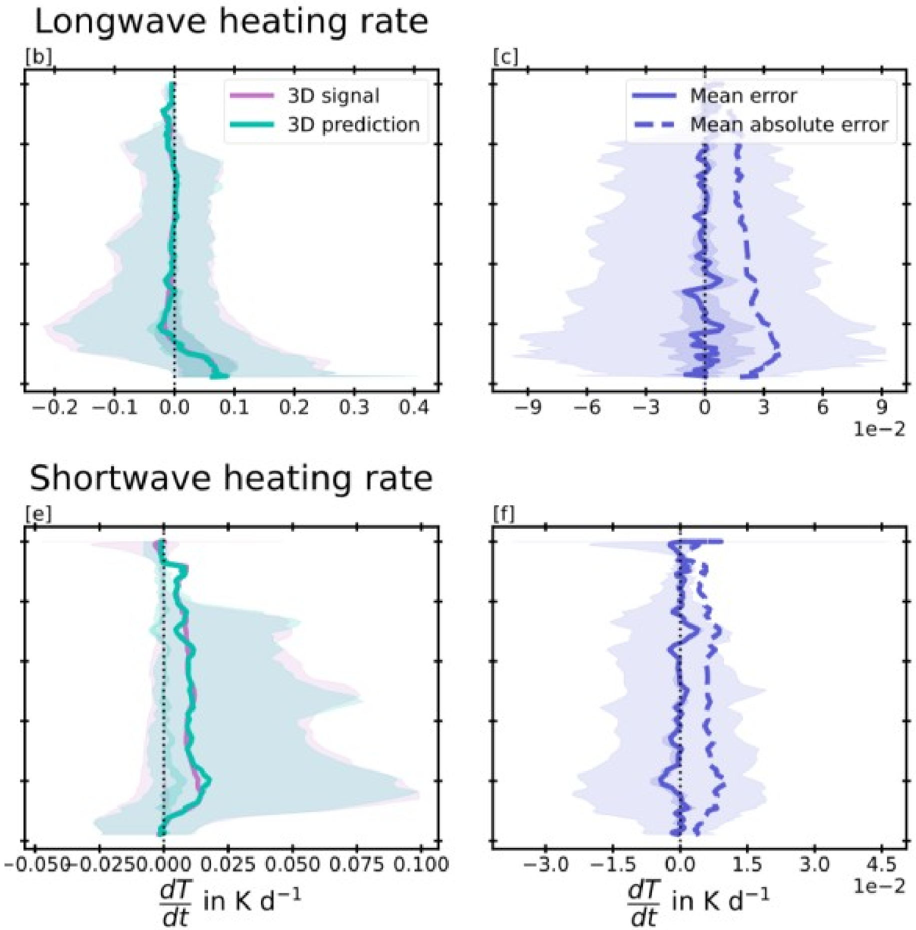
32-neuron BiGRUs, ~17k parameters



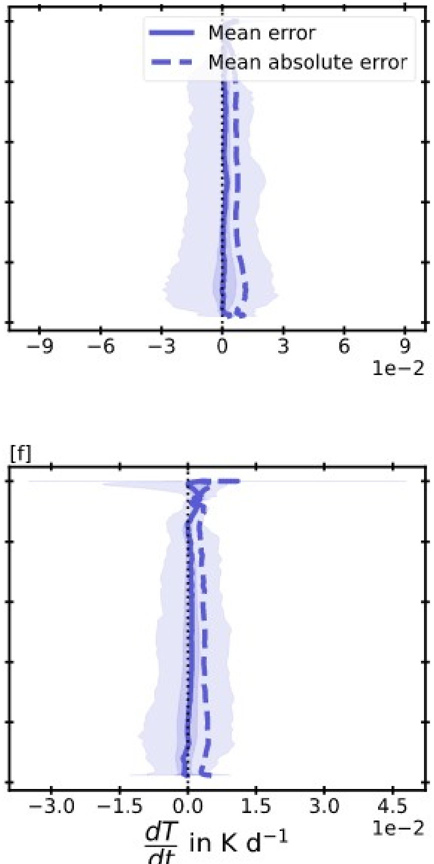
Emulation of 3D cloud radiative effects (Meyer et *al*. 2022)

Danmarks Meteorologiske Institut

DNNs from paper, ~525k parameters each



32-neuron BiGRUs, ~17k parameters

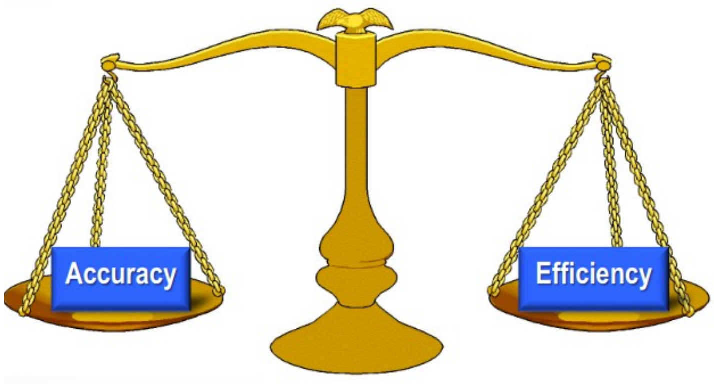


What about speed?

On CPU, RNNs ~4x faster than reference ecRad-SPARTACUS (3D solver that is emulated)..

Recent optimizations and improvements to ecRad change the picture: **now actually ~3x slower;** on the other hand, SPARTACUS is not yet fully stable in single precision, while NNs are

GPUs, half-precision further boost to emulation?



Key question:

Can machine learning actually “improve” the trade-off between accuracy and efficiency for radiation?

Answer: no free lunch with ML. Our results show that recurrent NNs can emulate radiation schemes very closely, but may or may not be faster. Dense nets are fast, but inaccurate.

Improving efficiency

For radiation, the motivation for using ML has been to improve speed, but there are other ways to achieve this

Code optimization is arguably an unexploited potential in improving the efficiency of weather / climate model code

Improving efficiency

For radiation, the motivation for using ML has been to improve speed, but there are other ways to achieve this

Code optimization is arguably an unexploited potential in improving the efficiency of weather / climate model code

Ukkonen and Hogan (*in prep.*): By combining

- Code refactoring to improve e.g. vectorization on modern CPUs, with
- Innovations in algorithms (reducing the spectral dimension)

..we end up with ~10x improvement in speed for TripleClouds and SPARTACUS; SPARTACUS with reduced gas optics actually 2x cheaper than operational ecRad in the IFS!

Summary and outlook

- Past studies emulating radiation and other sub-grid processes have typically used feed-forward NNs, concatenating vertical profiles of several variables into long input/output columns
- However, this approach does not really respect the physics: radiative transfer is **vertically non-local and sequential**. Why introduce spurious connections?
- We can instead **treat the vertical column as a sequence** and feed it to **RNNs**, allowing information to **directly propagate through the vertical column**. Single-level variables can be inserted “where it makes sense”, initializing the RNNs etc.
- **Accuracy is greatly improved**, with relatively simple models being able to **closely emulate radiative transfer**
- The **reduced dimensionality should improve generalization**, a key challenge in using ML for climate and weather model parameterizations
- (Personal take) Improvements to radiation codes have made them difficult to beat using ML, but **other parameterized processes such as clouds and convection are also vertically non-local** (and a **key source of uncertainty** in climate projections) – **could RNNs provide a breakthrough**?

Thanks for listening!

Any questions?