# Machine learning of cloud types in satellite observations and climate models

Peter Kuma[1], Frida A.-M. Bender[1], Alex Schuddeboom[2], Adrian J. McDonald[2], and Øyvind Seland[3]

[1]Department of Meteorology (MISU), Stockholm University, Stockholm, Sweden
[2]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, Aotearoa New Zealand
[3]Norwegian Meteorological Institute, Oslo, Norway

**Correspondence:** Peter Kuma (peter.kuma@misu.su.se)

**Abstract.** Uncertainty in cloud feedbacks in climate models is a major limitation in projections of future climate. Therefore, evaluation and improvement of cloud simulation is essential to ensure the accuracy of climate models. We analyse cloud biases and cloud change with respect to global mean near-surface temperature (GMST) in climate models relative to satellite observations, and relate them to equilibrium climate sensitivity, transient climate response and cloud feedback. For this purpose, we develop a supervised deep convolutional artificial neural network for determination of cloud types from low-resolution ($2.5° \times 2.5°$) daily mean top of atmosphere shortwave and longwave radiation fields, corresponding to the World Meteorological Organization (WMO) cloud genera recorded by human observers in the Global Telecommunication System (GTS). We train this network on a top of atmosphere radiation retrieved by the Clouds and the Earth's Radiant Energy System (CERES) and GTS, and apply it on the Climate Model Intercomparison Project phase 5 and 6 (CMIP5 and CMIP6) model output and the ECMWF Reanalysis version 5 (ERA5) and the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2) reanalyses. We compare the cloud types between models and satellite observations. We link biases to climate sensitivity and identify a negative linear relationship between the root mean square error of cloud type occurrence derived from the neural network and model equilibrium climate sensitivity (ECS), transient climate response (TCR) and cloud feedback. This statistical relationship in the model ensemble favours models with higher ECS, TCR and cloud feedback. However, this relationship could be due to the relatively small size of the ensemble used or decoupling between present-day biases and future projected cloud change. Using the abrupt-4xCO2 CMIP5 and CMIP6 experiment, we show that models simulating decreasing stratiform and increasing cumuliform clouds tend to have higher ECS than models simulating increasing stratiform and decreasing cumuliform clouds, and this could also be partially responsible for the difference in ECS observed.

## 1 Introduction

Clouds are a major factor influencing the Earth's climate. They are highly spatially and temporally variable, with the top of atmosphere (TOA) radiation being particularly sensitive to cloud changes due to their high albedo and impact on longwave radiation. Of all climate feedbacks, cloud feedback is the most uncertain feedback in Earth system models (ESMs) (Zelinka et al., 2020; Sherwood et al., 2020). Therefore, it is essential that climate models converge more on a correct representation of future clouds, but also on their representation of present-day clouds, which is a necessary (but not sufficient) condition for

the fidelity of projected cloud change. The estimate of the 'likely' range (66%) of ECS has recently been refined in the 6th
Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC) to 2.5–4 K, from 1.5–4.5 K in AR5.
Evidence for this estimate is only indirectly informed by the Climate Model Intercomparison Project phase 6 (CMIP6) models
(Eyring et al., 2016, 2019), which have a multi-model mean of 3.7 K (Meehl et al., 2020). The combined assessment is based
on paleoclimate and historical evidence, emergent constraints and process understanding. Notably, CMIP6 models predict a
16% higher multi-model mean than CMIP5 (3.2 K) (Meehl et al., 2020; Forster et al., 2020), and this fact has already been
examined in a number of studies (Zelinka et al., 2020; Wyser et al., 2020; Schlund et al., 2020; Dong et al., 2020; Nijsse et al.,
2020; Flynn and Mauritsen, 2020). It is also higher than the combined assessment central value of ECS of 3 K in AR6. The
multi-model spread in CMIP6 is also larger than in CMIP5, with a standard deviation of 1.1 K in CMIP6 vs. 0.7 K in CMIP5.
Modelling groups have prevailingly reported that the higher multi-model mean is due to changes in cloud representation
in the recent generation of models (Meehl et al., 2020, Table 3), supported by the findings of Zelinka et al. (2020). Some
authors have concluded that the high-ECS CMIP6 models are on average overestimating the ECS (Nijsse et al., 2020) and
are not compatible with paleoclimatic records (Zhu et al., 2020, 2021). The high-ECS models are also not supported by the
review studies of Sherwood et al. (2020) and AR6. In contrast, Bjordal et al. (2020) argue that high ECS models might be
plausible because of state-dependent cloud phase feedback in the Southern Ocean. Models which simulate too much ice in the
Southern Ocean clouds, a common bias among CMIP models, are expected to have lower cloud feedbacks globally because of
a spuriously enhanced negative feedback associated with cloud phase changes in that region. Recently, Volodin (2021) reported
that changing cloud parametrisation in the Institute of Numerical Mathematics Coupled Model version 4.8 (INM-CM4-8) from
Smagorinsky type to a prognostic type of Tiedtke (1993) resulted in more than doubling of ECS from 1.8 K to 3.8 K. This
underscores the importance of cloud parametrisation in determining model climate sensitivity.

The effect of clouds on the climate comes primarily from cloud fraction and cloud optical depth, which are determined
by factors and properties such as convection, mass flux, turbulence, atmospheric dynamics, cloud microphysics (cloud phase,
cloud droplet and ice crystal size distribution, number concentration, ice crystal habit), vertical cloud overlap, cloud altitude,
cloud cell structure, cloud lifetime and more. Accurate simulation of cloud within climate models is difficult not only because
of the large number of properties, many of which are subgrid-scale in today's general circulation models, but also because
compensating model biases may produce correct cloud radiative effect (CRE) while simulation of the individual properties
is incorrect. This may be especially true for global radiation budget, as model processes are often tuned to achieve a desired
radiation balance at TOA (Hourdin et al., 2017; Schmidt et al., 2017).

Cloud genera (WMO, 2021a) have been an established way of describing clouds for over a century. They broadly corre-
spond or correlate with the individual cloud properties such as cloud altitude, optical depth, phase, overlap and cell structure.
Therefore, they can be used as a metric for model evaluation which, unlike more synthetically-derived cloud classes, is easy
to understand and has a very long observational record. So far, however, it has not been possible to identify cloud genera in
low-resolution model output, because their identification depends on a high-resolution visual observation, generally from the
ground. Here, we show that it is possible to use a supervised deep convolutional artificial neural network (ANN) to identify
cloud genera in low-resolution model output and satellite observations. Past classifications of cloud types or cloud regimes

derived from satellite datasets have been based on cloud optical depth and cloud top pressure or height by simple partitioning (Rossow and Schiffer, 1991) or by statistical clustering algorithms (Jakob and Tselioudis, 2003; McDonald et al., 2016; Oreopoulos et al., 2016; Cho et al., 2021), and on active radar and lidar sensors (Cesana et al., 2019), which likely only broadly correspond to human-observed cloud genera. More recently, deep ANNs have begun to be used to identify and classify clouds (Zantedeschi et al., 2020). Olsson et al. (2004) developed an ANN for classifying clouds in a numerical weather forecasting model output based on reference satellite data from the Advanced Very High Resolution Radiometer (AVHRR).

We introduce a new method of quantifying cloud types correspoding to the WMO cloud genera in model and satellite data based on an ANN approach. Furthermore, we quantify their global distribution and change with respect to global mean near-surface temperature (GMST) in CMIP5 and CMIP6 models, the Clouds and the Earth's Radiant Energy System (CERES) satellite data and two reanalyses, European Centre for Medium-Range Weather Forecasts's (ECMWF) reanalysis ERA5 (Hersbach et al., 2020) and the Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA-2) (Gelaro et al., 2017). Convolutional artificial neural networks have been used before for cloud detection: Shi et al. (2017), Ye et al. (2017), Wohlfarth et al. (2018), Zhang et al. (2018), Liu and Li (2018) and Zantedeschi et al. (2020) used a convolutional ANN for identification of cloud genera in ground-based cloud images, and Drönner et al. (2018), Shendryk et al. (2019), Guo et al. (2020), Segal-Rozenhaimer et al. (2020) and Liu et al. (2021) developed a convolutional ANN for detecting cloudy pixels in high-resolution satellite imagery. While the determination of cloud types in model and satellite data and application of ANNs to identify cloud types are not new, in contrast to previous methods, we utilise cloud types with a direct correspondence to the established human-observed World Meteorological Organization (WMO) cloud genera to train our ANN. This dataset contains many decades of global cloud observations, recorded several times daily at a large number of stations. For this purpose, we develop an ANN which can be applied to input with low spatial and temporal resolution (2.5°, daily mean). This is because most current climate models provide output with low resolution. This resolution is not sufficient to represent individual clouds, but these can still be inferred statistically from large scale patterns. Likewise, the resolution of some satellite datasets such as CERES is on this spatial scale. We try to answer the question of whether cloud type biases and change with respect to GMST are related to cloud feedback, ECS and TCR in the CMIP models. The ANN and the associated code is made available under an open source license (Kuma et al., 2022).

## 2  Methods

### 2.1  Rationale and methods outline

We trained an ANN on satellite observations from the Clouds and the Earth's Radiant Energy System (CERES) and ground based observations from WMO stations in the Internet Data Distribution (IDD) dataset (Unidata, 2003). Then, we applied the ANN on CERES data, CMIP5 and CMIP6 model output, ERA5 and MERRA-2. A large database of ground-based cloud observations has been compiled in the IDD dataset. This database contains human-observed cloud information at standard synoptic times, encoded as three numbers between 0 and 9 (or missing) for low, mid-level and high clouds specifying the cloud genus and species (WMO, 2011, 2021a). ANNs are typically used for various forms of image labelling, where the ANN

is trained on a set of images either to label whole images (e.g. to identify if a certain object is present in the image), or to perform 'segmentation', where image pixels are classified as belonging to a certain object. Here, we used an ANN capable of quantifying the probability of presence of cloud genera in individual pixels of an image composed of shortwave and longwave channels coming from either satellite or model output. Spatial pattern and magnitude of shortwave (SW) and longwave (LW) radiation provides information about clouds, which can be used to train an ANN to classify clouds. The purpose was to quantify WMO cloud genera on the whole globe (rather than at individual station as available already in the IDD) and in model output. For practical purposes, in our analysis we grouped together multiple cloud genera to a smaller number of 'cloud types'. The ANN training phase consisted of supervised training on daily mean CERES satellite images, where for some pixels we knew the presence or absence of the cloud types based on one or more ground stations located within the pixel. The training of the output is done on these pixels. Because the number of stations and days of observations is relatively large, it was possible to train the ANN to quantify the probability of presence of cloud genera in any pixel of a satellite image or model output, and by extension on the whole globe.

The 'U-Net' ANN (Ronneberger et al., 2015) is a well-established type of ANN used for pixel-wise classification of images. The main feature of this network is its U-shaped sequence of steps of downsampling followed by upsampling, allowing it to learn patterns on different size scales and produce output of the same size as the input. This makes it suitable for our task of quantifying cloud type occurrence probability for each pixel in passive satellite images or an equivalent climate model output.

The schematic in Fig. 1 shows an outline of training and application of this ANN in our analysis. All input spatially-distributed satellite and model data were first resampled to 2.5° spatial resolution to ensure uniformity. In the training phase (Fig. 1a), samples of daily mean TOA SW and LW radiation from CERES were produced. A sample is an image of size $4000{\times}4000$ km and $16{\times}16$ pixels produced by projecting the daily mean TOA SW and LW radiation field in a local geographical projection at a random location on the globe. This step is necessary in order to produce input data which is spatially undistorted (as would be the case with unprojected global fields). Sampling at random locations ensures more robust training of the ANN, which could otherwise more easily be trained to recognise geographical locations as opposed to recognising cloud patterns irrespective of their location. Samples were paired with ground station cloud observations. Supervised training of the ANN was performed using these samples (20 per day, 4582 days in total). The training of cloud type occurrence probability, which is the output of the ANN, was only done for pixels where ground station data were available. In the application phase (Fig. 1b), samples from CERES or a model were produced and supplied to the ANN, which produced samples with quantified cloud type occurrence probability. These were then either merged to reconstruct a geographical distribution, or global means and trends were calculated from all samples. In detail, the ANN input were samples consisting of two channels of SW and LW radiation, and the output were samples consisting of four, 10 or 27 channels (for classifications into four, 10 and 27 cloud types, respectively) of cloud occurrence probability corresponding to the cloud types (Fig. 1c). The classifications of four and 10 cloud types were created by grouping of the full set of 27 WMO cloud genera/species. The advantage of using an ANN for cloud classification over more traditional methods such as partitioning the cloud top pressure–optical depth space (Rossow and Schiffer, 1991) is a true correspondence to human-identified cloud genera, its potential flexibility to identify more specific cloud genera/species, and the ability to extend observations of human-identified cloud genera/species to regions not covered
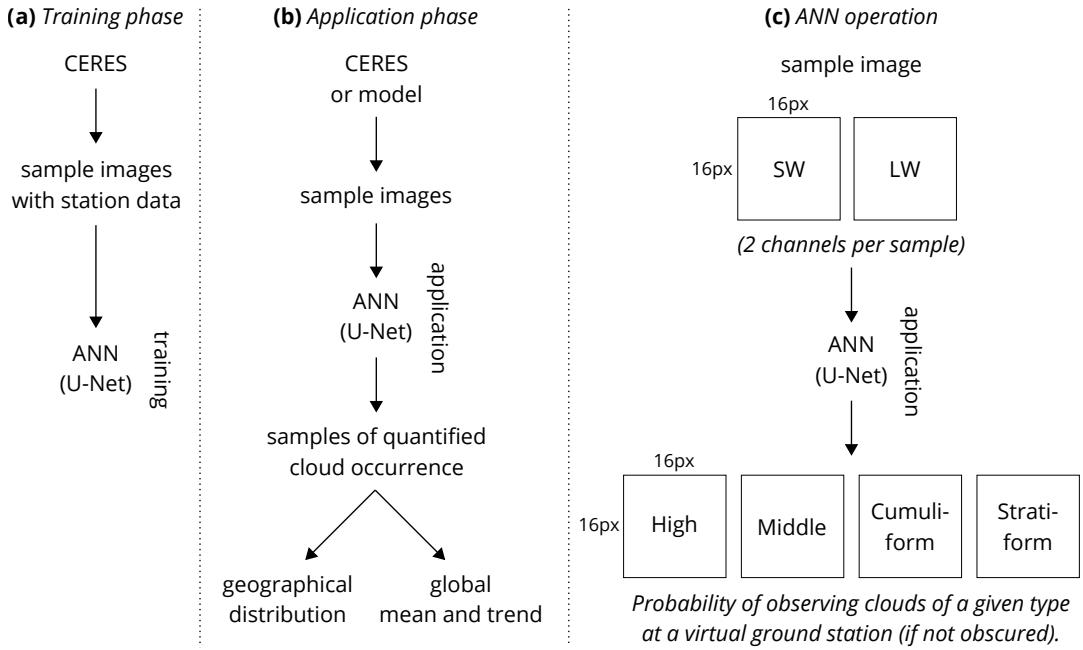
**Figure 1.** Schematic showing the principle of training and applying the ANN. **(a)** Sample images from CERES with reference ground station data (WMO cloud genera) are used in the training phase. **(b)** In the application phase, the ANN quantifies cloud type occurrence in sample images from CERES or a model. **(c)** The ANN takes a sample image consisting of SW and LW channels and produces per-pixel quantification of probability of observing a given cloud type at a virtual ground station located in the pixel. The samples are 16×16 pixels in size and 4000×4000 km spatially.

by ground stations. The method outlined above allowed us to consistently quantify cloud genera/species occurrence in satellite observations and climate models and evaluate model biases.

## 2.2  Data

### 2.2.1  Satellite observations

We used satellite observations from the Clouds and the Earth's Radiant Energy System (CERES) in years 2003–2020 (Wielicki et al., 1996; Doelling et al., 2013; Loeb et al., 2018) as a reference training dataset for the ANN, and in particular, the daily mean adjusted all-sky and clear sky shortwave and longwave fluxes at TOA and shortwave (solar) insolation from SYN1deg 1°-resolution geostationary-enhanced product Terra+Aqua Edition 4.1. For evaluation of cloud top pressure and cloud optical depth, we used satellite-retrieved cloud visible optical depth (from 3.7 μm particle size retrieval) and cloud top pressure from the same product.

### 2.2.2 Climate models

CMIP5 and CMIP6 are the last two iterations of standardised global climate model experiments (Taylor et al., 2012; Eyring et al., 2016). We applied our ANN to the publicly-available model output of the *historical* and *abrupt-4xCO2* CMIP experiments in the daily mean products. An exception were the EC-Earth and NorESM2-LM models, which did not provide the necessary variables. For EC-Earth, we used data from the *hist-1950* experiment (model EC-Earth3P) of the High Resolution Model Intercomparison Project (HighResMIP) (Haarsma et al., 2016, 2020) as a substitute for historical. The model output resolution of EC-Earth3P is the same as EC-Earth. For NorESM2-LM, we obtained the data directly from the model developers. The variables used in our analysis were (exclusively) *rsut* (TOA outgoing shortwave radiation), *rlut* (TOA outgoing longwave radiation), *rsutcs* (TOA outgoing clear-sky shortwave radiation), *rlutcs* (TOA outgoing clear-sky longwave radiation), *rsdt* (TOA incident shortwave radiation) and *tas* (near-surface air temperature). In connection to CMIP models, we used estimates of the model ECS, TCR and cloud feedback from AR6, with missing values supplemented by Meehl et al. (2020), Zelinka et al. (2020), and ECS and TCR calculated with the ESMValTool version 2.4.0 (Righi et al., 2020). Here, we use a definition of cloud feedback adjusted for non-cloud influences as in Zelinka et al. (2020), Soden et al. (2008) and Shell et al. (2008). Table 1 lists CMIP5 and CMIP6 models used in our analysis, their ECS, TCR and cloud feedback. In total, we analysed 4 CMIP5 and 20 CMIP6 models, of which 18 had the necessary data in the historical experiment for comparison with CERES (years 2003–2014), and 22 had the necessary data in the abrupt-4xCO2 experiment. No selection was done on the models, i.e. all CMIP5 and CMIP6 models which provided the required fields in the CMIP archives were analysed here. For some models, ECS, TCR or cloud feedback were not available. For these models, the values were taken from the closest available model (as detailed in Table 1). The model developers of IPSL-CM6A-LR-INCA advised us that its TCR should be the same as IPSL-CM6A-LR (Olivier Boucher, personal communication).

### 2.2.3 Reanalyses

In addition to CMIP, we analysed the output of two reanalyses: ERA5 (Hersbach et al., 2020) and MERRA-2 (Gelaro et al., 2017). From MERRA-2, we used the *M2T1NXRAD* product: daily means of the variables *LWTUP* (upwelling longwave flux at TOA), *LWTUPCLR* (upwelling longwave flux at TOA assuming clear sky), *SWTDN* (TOA incoming shortwave flux), *SWTNT* (TOA net downward shortwave flux) and *SWTNTCLR* (TOA net downward shortwave flux assuming clear sky). From ERA5, we used the ERA5 hourly data on single levels from 1979 to present product variables *tsr* (top net solar radiation), *tsrc* (top net solar radiation, clear sky), *ttr* (top net thermal radiation) and *ttrc* (top net thermal radiation, clear sky). The variables were used in an equivalent way to the CMIP5 and CMIP6 variables (Section 2.2.2).

### 2.2.4 Station observations

In addition to satellite and model data, we used ground-based land and marine station data from the Historical Unidata Internet Data Distribution (IDD) Global Observational Data (Unidata, 2003). This dataset is a collection of the Global Telecommunication System (GTS) (WMO, 2021b) reports, which come from synoptic messages sent by stations to the WMO network.

**Table 1.** Table of CMIP5, CMIP6 models and reanalyses used in our analysis and their CMIP phase, equilibrium climate sensitivity (ECS), transient climate response (TCR), cloud feedback (CLD), if they provided the necessary variables in the *historical* (hist.) (in the case of reanalyses *historical reanalysis*) and *abrupt-4xCO2* experiments ('●' – yes, '○' – no, '-' – not applicable), and model output resolution (Res.) as the number of longitude × latitude bins. Models are sorted by their ECS. ECS, TCR and CLD were sourced from AR6, Zelinka (2021) and Semmler et al. (2021).

| # | Model | Phase | ECS (K) | TCR (K) | CLD (Wm$^{-2}$K$^{-1}$) | hist. | abrupt-4xCO2 | Res. (lon.×lat.) |
|---|---|---|---|---|---|---|---|---|
| 1 | INM-CM4-8 | 6 | 1.83 | 1.33 | -0.09 | ● | ● | 180×120 |
| 2 | INM-CM5-0 | 6 | 1.92 | 1.3 | -0.06 | ● | ● | 180×120 |
| 3 | NorESM2-LM | 6 | 2.54 | 1.48 | 0.44 | ● | ● | 144×96 |
| 4 | MRI-CGCM3 | 5 | 2.60 | 1.60 | 0.28 | - | ● | 320×160 |
| 5 | MPI-ESM-1-2-HAM | 6 | 2.96 | 1.8 | -0.16 | ● | ● | 192×96 |
| 6 | MPI-ESM1-2-HR | 6 | 2.98 | 1.66 | 0.27 | ● | ● | 384×192 |
| 7 | MPI-ESM1-2-LR | 6 | 3.00 | 1.84 | 0.18 | ● | ● | 192×96 |
| 8 | MRI-ESM2-0 | 6 | 3.15 | 1.64 | 0.46 | ● | ● | 320×160 |
| 9 | AWI-ESM-1-1-LR | 6 | 3.29 | 2.11 | 0.29[a] | ● | ○ | 192×96 |
| 10 | MPI-ESM-LR | 5 | 3.63 | 2.00 | 0.44 | - | ● | 192×96 |
| 11 | IPSL-CM5A2-INCA | 6 | 3.79 | 1.9 | 1.05 | ● | ● | 96×96 |
| 12 | GFDL-CM4 | 6 | 3.89 | 2.10 | 0.64 | ○ | ● | 144×90 |
| 13 | IPSL-CM5A-MR | 5 | 4.12 | 2.00 | 1.25 | - | ● | 144×143 |
| 14 | IPSL-CM5A-LR | 5 | 4.13 | 2.00 | 1.18 | - | ● | 96×96 |
| 15 | IPSL-CM6A-LR-INCA | 6 | 4.13 | 2.32[a] | 0.43 | ● | ● | 144×143 |
| 16 | CNRM-CM6-1-HR | 6 | 4.28 | 2.48 | 0.59 | ● | ● | 720×360 |
| 17 | EC-Earth3P | 6 | 4.31[a] | 2.62[a] | 0.37[a] | ● | ○ | 512×256 |
| 18 | IPSL-CM6A-LR | 6 | 4.56 | 2.32 | 0.45 | ● | ● | 144×143 |
| 19 | CNRM-ESM2-1 | 6 | 4.76 | 1.86 | 0.63 | ○ | ● | 256×128 |
| 20 | CNRM-CM6-1 | 6 | 4.83 | 2.14 | 0.61 | ● | ● | 256×128 |
| 21 | UKESM1-0-LL | 6 | 5.34 | 2.79 | 0.87 | ● | ● | 192×144 |
| 22 | HadGEM3-GC31-MM | 6 | 5.42 | 2.58 | 0.91 | ● | ● | 432×324 |
| 23 | HadGEM3-GC31-LL | 6 | 5.55 | 2.55 | 0.84 | ● | ● | 192×144 |
| 24 | CanESM5 | 6 | 5.62 | 2.74 | 0.88 | ● | ● | 128×64 |
| 25 | ERA5 | - | - | - | - | ● | - | 1440×721 |
| 26 | MERRA-5 | - | - | - | - | ● | - | 576×361 |

[a]For some models, ECS, TCR or CLD were not available. For these models, the values were taken from the closest available model (CLD of AWI-ESM-1-1-LR as in AWI-CM-1-1-MR; ECS, TCR and CLD of EC-Earth3P as in EC-Earth3-Veg; and TCR of IPSL-CM6A-LR-INCA as in IPSL-CM6A-LR).

170 They consist of standard synoptic observations. For stations with an observer, clouds are identified visually at three different levels (low, middle and high). For each level, a cloud genus/species category is recorded as a number between 0 and 9 or as not available. Therefore, for each station, up to three numbers are available encoding cloud genera/species, the meaning of which is explained in the WMO Manual on Codes (WMO, 2011) in the code tables 0509, 0513 and 0515. Only one cloud genus/species category can be recorded for each level. Cloud fraction information in the station data was not used in our analysis.

175 The IDD records were available between 2003-05-19 and 2020-12-31 at standard synoptic times (00Z, 03Z, ..., 21Z). We excluded years in which more than three weeks of data were missing: 2006, 2007 and 2008. We used the cloud genus variables of the synoptic (SYNOP) and marine (BUOY) reports: low cloud (IDD variable 'cloudLow') based on the WMO *Code Table 0513*, middle cloud based on *Code Table 0515* (IDD variable 'cloudMiddle') and high cloud based on *Code Table 0509* (IDD variable 'cloudHigh') (WMO, 2011). Furthermore, we grouped the cloud genera/species into four cloud types to simplify our
180 analysis:

1. *high*: cirrus, cirrostratus, cirrocumulus ($C_H$ codes 1–9),

2. *middle*: altostratus, altocumulus ($C_M$ codes 1–9),

3. *cumuliform*: cumulus, cumulonimbus ($C_L$ codes 1–3, 8, 9),

4. *stratiform*: stratocumulus, stratus ($C_L$ codes 4–7).

185 To provide more detail, we also used an extended grouping of 10 cloud types:

1. *Ci*: cirrus ($C_H$ codes 1–6)

2. *Cs*: cirrostratus ($C_H$ codes 7–8)

3. *Cc*: cirrocumulus ($C_H$ code 9)

4. *As*: altostratus ($C_M$ codes 1–2)

190 5. *Ac*: altocumulus ($C_M$ codes 3–9)

6. *Cu*: cumulus ($C_L$ codes 1–3)

7. *Sc*: stratocumulus ($C_L$ codes 4–5)

8. *St*: stratus ($C_L$ codes 6–7)

9. *Cu+Sc*: cumulus and stratocumulus ($C_L$ code 8)

195 10. *Cb*: cumulonimbus ($C_L$ code 9)

As an example of the geographical distribution of stations, Fig. 2a shows the location of SYNOP and BUOY station reports with cloud data available on 1 January 2010. Because the data come from operational weather stations, they are geographically biased to certain locations, especially land, extratropics and the Northern Hemisphere. Undersampled locations are ocean, the Southern Hemisphere and the poles. Cloud type information from stations in the USA and Australia is also not available in the WMO records. Because of partially missing data in 2003, 2006 and 2008, we excluded these years in the ANN training phase. Not all stations in the IDD database provide cloud type information, and such stations were excluded from our analysis. High and middle clouds can be obscured by underlying cloud layers. In such cases, the observation of high or middle clouds is recorded as missing in the IDD data and we exclude such stations from the calculation of statistics for the middle or high cloud types, respectively. This limitation of the dataset means it is less suitable for identifying middle and high clouds than low clouds in multi-layer cloud situations, and a similar, but reverse, limitation exists in spaceborne cloud observations (McErlich et al., 2021).

### 2.2.5 Historical global mean near-surface temperature

Historical GMST was sourced from the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis version 4 (GISTEMP v4) (Lenssen et al., 2019; GISTEMP Team, 2021). This dataset was used in combination with the CERES dataset to determine observed change of cloud type occurrence with respect to GMST.

### 2.3 Artificial neural network

TensorFlow is a machine learning framework for development of artificial neural networks (Abadi et al., 2016), supporting deep and convolutional neural networks. We used the Keras application programming interface (API) of TensorFlow (version 1.14), which provides a simple abstraction layer over TensorFlow, to define, train and apply a deep convolutional ANN on satellite and model data. The ANN was based on a network type called U-Net (Ronneberger et al., 2015), which produces output on the same grid as the input. The input were 2-dimensional arrays of size 16×16 pixels for SW and LW radiation. The output were 2-dimensional arrays of the same size for each cloud type.

The ANN training was performed as follows, demonstrated schematically in Fig. 1 and 2. For each day, we generated 20 samples of 4000×4000 km and 16×16 pixels from CERES data, composed of two channels (shortwave and longwave radiation) projected in a local azimuthal equidistant projection centred at stochastically random locations uniformly distributed on the globe (Fig. 2b). More precisely, the channels were calculated from daily mean TOA all-sky outgoing shortwave and longwave radiation (*rsut* and *rlut*, respectively), clear sky outgoing shortwave and longwave radiation (*rsutcs* and *rlutcs*, respectively) as (1) shortwave CRE normalised to the incoming solar radiation, (2) longwave CRE normalised to clear sky outgoing longwave radiation:

$$\mathrm{CRE_{SW,norm.}} = (\mathrm{rsut} - \mathrm{rsutcs})/\mathrm{rsdt}, \tag{1}$$

$$\mathrm{CRE_{LW,norm.}} = (\mathrm{rlutcs} - \mathrm{rlut})/\mathrm{rlutcs}. \tag{2}$$
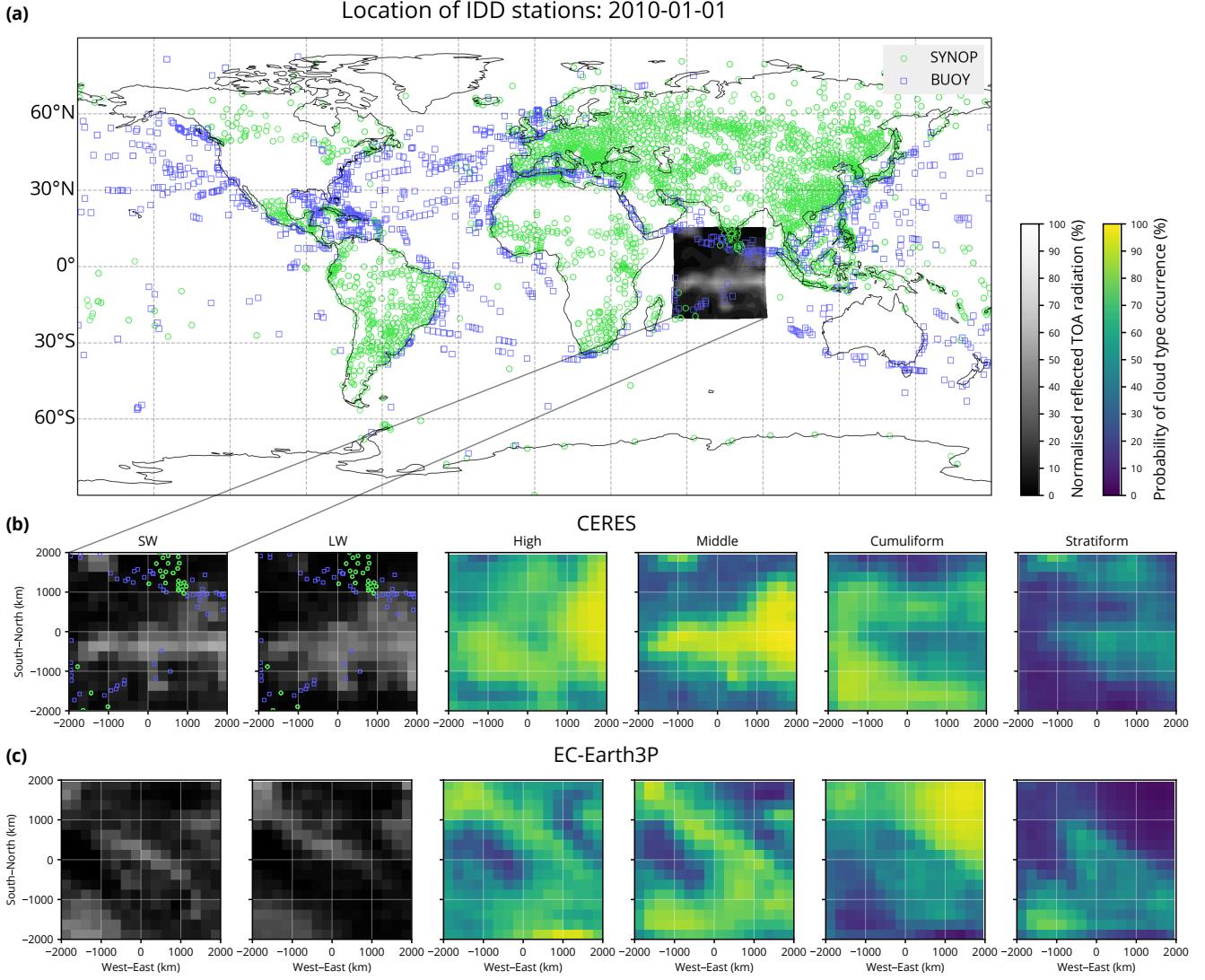
**Figure 2. (a)** A map showing the location of Internet Data Distribution (IDD) station reports containing cloud information on a single day. Shown is a sample of normalised top of atmosphere (TOA) shortwave radiation from CERES. **(b)** The sample as in (a), but shown re-projected in a local azimuthal equidistant projection. Shown is the shortwave (SW) and longwave (LW) radiation, and the probability of cloud type occurrence calculated by the ANN for the classification into four cloud types. **(c)** A similar sample as in (b) but from the climate model EC-Earth3P (historical experiment) taken on the same day at a different location. This sample shows an unrelated cloud scene due to the fact that the model is free running. Coastline data come from the public domain Global Self-consistent, Hierarchical, High-resolution Geography Database (Wessel and Smith, 1996, 2017).

The normalisation was done so that the values were mostly in the [0, 1] interval, which is a more suitable input to the ANN than non-normalised values. In the shortwave radiation, normalisation by incoming shortwave radiation was chosen so that the value represents the fraction of reflected incoming radiation due to clouds. In the longwave radiation, such normalisation is not possible, and normalisation by outgoing clear sky longwave radiation was performed instead.

In order to exclude locations with low solar insolation, where $CRE_{SW,norm.}$ might be ill-defined because of low values of the denominator, we excluded parts of $CRE_{SW,norm.}$ where incoming solar radiation was lower than $50\ \mathrm{Wm^{-2}}$. A downside of this approach is that this may cause bias due to exclusion of wintertime polar regions. If a sample was missing any data points, it was excluded from the analysis. The shortwave and longwave channels were the input to the ANN training phase. The loss function of the ANN training was defined as the negative of log-likelihood of observing cloud types at ground stations. The log-likelihood for each pixel and cloud type was:

$$l = n_{\mathrm{positive}} \ln(p) + (n - n_{\mathrm{positive}}) \ln(1 - p) \tag{3}$$

where $n_{\mathrm{positive}}$ is the number of station records in the pixel which observed a given cloud type, $p$ is the probability of observing the cloud type predicted by the ANN, and $n$ is the total number of station records in the pixel with information about the cloud type. The total log-likelihood to be optimised was the sum of log-likelihood (as defined above) over all pixels in all samples and all cloud types. Station records which reported clear sky were also included. In the optimisation process, in which the internal ANN coefficients were trained to predict the reference observations, the loss function equal to $-l$ was minimised. The optimisation process was run in iterations until the validation set loss function was not improved for three iterations.

In the application phase, 20 random samples per day were generated from CERES and model data. The ANN estimated the probability of cloud type occurrence for every pixel of each sample based on the input consisting of $16{\times}16$ pixel images of SW and LW radiation calculated in the same way as in the training phase. From the samples we reconstructed geographical distributions and calculated global means or the probability of cloud type occurrence, i.e. the probability that a cloud type can be observed at a virtual ground station located in a given pixel.

## 2.4 Cloud top pressure–cloud optical depth evaluation

We calculated cloud top pressure–cloud optical depth histograms corresponding to the cloud types (Section 3.3). The histograms were calculated from cloud top pressure and cloud visible optical depth variables in the CERES SYN1deg daily mean product over the time period of year 2003 to 2020 (inclusive) weighted by the daily cloud type occurrence calculated by the ANN. For each cloud type, the histogram density was calculated by iterating over all samples produced by the ANN and for every pixel incrementing the histogram bin corresponding to the pixel's cloud top pressure and cloud optical depth in the CERES product by the probability of the cloud type occurrence in the pixel calculated by the ANN. The set of samples did not include random samples taken partially over polar night locations as explained in Sect. 2.3. The results shown in the histograms therefore do not include any information about polar regions during polar night.

## 3 Results

### 3.1 Training

We trained the ANN on CERES and IDD data in years 2004, 2005, 2007 and 2009–2017, with years 2007, 2012 and 2017 used as a validation dataset, representing 25% of the total number of years. The training was completed in 32, 40, and 38 iterations (for an ANN of 4, 10 and 27 cloud types, respectively), interrupted automatically once the validation loss function stopped improving for three iterations. The loss function during the training phase is in Fig. S1.

### 3.2 Geographical biases

Fig. 3 and 4 show the geographical distribution of cloud types in CERES, IDD, CMIP models and the reanalyses (ERA5 and MERRA-2) for the four cloud types (the same plots but for 10 and 27 cloud types are available in a supplement). The IDD row represents an observational reference calculated idependently from the ANN (although the ANN is originally trained on this dataset), while rows corresponding to CERES and the CMIP models are calculated by the ANN. The high cloud type is characterised by a peak over the western tropical Pacific and tropical Africa, corresponding to peaks in the IDD. A peak over northern Asia is, however, more muted in the ANN. The middle cloud type peaks over northern Pacific and western tropical Pacific, and has a minimum over central Africa. Northern Pacific is not sampled well in the IDD, but a peak over western tropical Pacific and a minimum over central Africa are also present in the raw dataset. The cumuliform cloud type is strongly concentrated in tropical marine regions over tropical Pacific and Atlantic and Madagascar and minima over polar regions and tropical Africa. This is also present in the IDD. The stratiform cloud type has maxima over polar marine regions and on the west coast of South America, and minima over tropical Pacific, Madagascar and tropical Africa. The maxima in the IDD are co-located but stronger, while the minima are co-located and of similar magnitude.

Model biases are most strongly characterised by a negative bias in the cumuliform cloud type over marine tropical regions and a positive bias in the stratiform cloud type over the same regions. Bias in the high and middle cloud type is more geographically varied. Models with higher ECS tend to have the opposite bias – positive bias in the cumuliform cloud type and negative bias in the stratiform cloud type globally. Notably, models with lower ECS generally have higher biases than models with higher ECS. Of models with ECS below 4 K (9 models), all but two have total RMSE greater than or equal 8%. Of models with ECS above 4 K (9 models), all have RMSE lower than 8%. The two reanalyses (ERA5 and MERRA-2) have relatively low biases compared to the CMIP models. ERA5 has the lowest total RMSE of all models at 3.6%.

Models which are closely related in their code (CNRM-*; ERA5 and EC-Earth3P; HadGEM3-* and UKESM1-0-LL; INM-*; IPSL-*; MPI-*) performed similarly in terms of geographical distribution and magnitude of biases. This means that the ANN method is robust with respect to model resolution, and also that the groups of related models represent clouds very similarly, presumably because this is to a large extent determined by cloud parametrisations in the atmospheric component of the model, without much sensitivity to resolution.
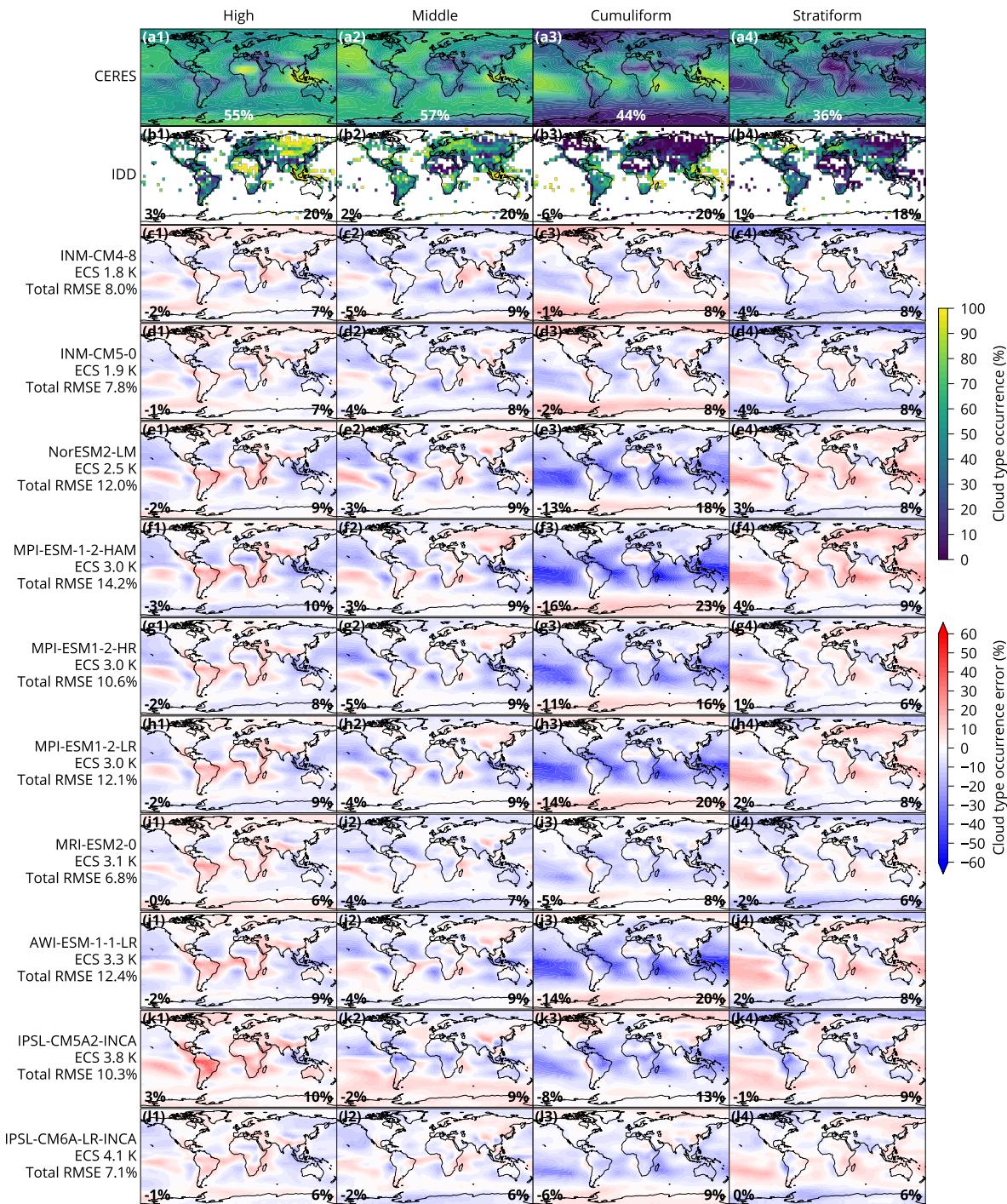
**Figure 3.** Geographical distribution of cloud type occurrence derived by applying the ANN on retrieved CERES satellite data in years 2003–2020 **(a1–a4)**, directly from the IDD (year 2010) **(b1–b4)**, and by applying the ANN on model output of the historical experiment of CMIP6 and reanalyses in years 2003–2014 relative to CERES **(c–k)**. In the lower centre of the CERES plots is the geographical mean occurrence of the cloud type. In the lower left is the mean error and in the lower right is the RMSE. Models are sorted by their ECS from lowest to highest. Coastline data come from the public domain Global Self-consistent, Hierarchical, High-resolution Geography Database (Wessel and Smith, 1996, 2017).
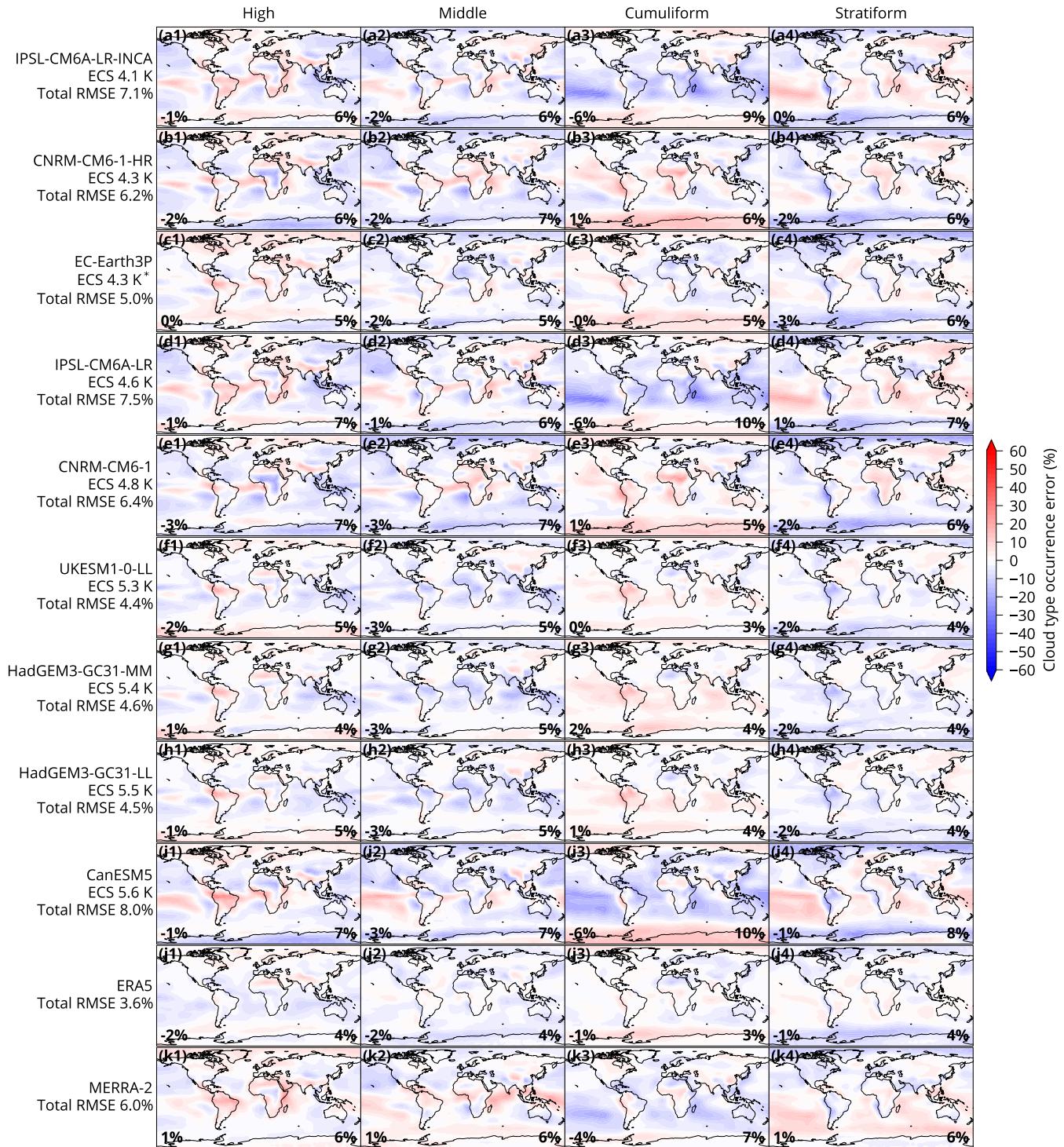
**Figure 4.** Fig. 3 continued. *For some models, ECS was not available and was taken from the closest available model (see Table 1).

### 3.3 Optical properties and vertical distribution

From the ANN-labelled samples we calculated joint histograms of cloud optical depth and cloud top pressure (Fig. 5). This type of histograms relates to previous work on cloud classification of Rossow and Schiffer (1991), Rossow and Schiffer (1999), Hahn et al. (2001), Oreopoulos et al. (2016), Schuddeboom et al. (2018) and others. The diagrams in Fig. 5 show cloud type occurrence binned by cloud optical depth and cloud top pressure for the four types as a difference from the mean of the four types. The high cloud type difference from the mean is characterised by a maximum occurrence at low pressure (300–700 hPa) and low optical depth (below 5) (Fig. 5b). The middle cloud type difference from the mean is characterised by high optical depth (above 2) between 200 and 800 hPa (Fig. 5c). The cumuliform and stratiform types have greater deviation from the mean than the high and middle types. The cumuliform cloud type difference from the mean has a maximum in optically thin clouds (below 4) between 400 and 1000 hPa (Fig. 5d). The stratiform cloud type is almost the inverse of the cumuliform type (Fig. 5e), characterised by mid-to-high optical depth (above 2) and low-to-mid altitude (below 400 hPa, peaking at about 700 hPa). Collectively, the cloud types span discrete regions in four sectors of the diagram, which indicates that they partition the cloud optical depth–cloud top pressure space quite well with little overlap. This means that the classification method distinguishes well between types of clouds in terms of their cloud optical depth and cloud top pressure. This anlaysis also shows that the stratiform and middle cloud types as identified by the ANN are generally more opaque than the cumuliform and high cloud types.

To compare with a more traditional classification by Rossow and Schiffer (1991), we present a diagram corresponding to their International Satellite Cloud Climatology Project (ISCCP) classification (Fig. 5f). Correspondence of the cumuliform and stratiform types between the ANN and the ISCCP classification is quite good, except for deep convection (classified as the cumuliform type in the ISCCP diagram) of highly opaque (optical depth above 20) high clouds (above 400 hPa). The correspondence is less good for the high cloud type, where the the ISCCP class starts at higher altitude (400 hPa vs. 700 hPa) and is also more opaque (optical depth up to 20 vs. 5). The correspondence of the middle cloud type is relatively poor – the ISCCP middle clouds are located at lower altitude (500 hPa vs. 400 hPa) and have lower optical depth (down to 0 vs. down to 2). Therefore, the correspondence between our classification and the ISCCP classification of Rossow and Schiffer (1991) is mixed, with good correspondence of the low cloud types, but disparities in the middle and high cloud types.

In addition to the comparison with the ISCCP classification above, in Appendix B we present a comparison with cloud clusters derived using self-organising maps (SOM) of McDonald and Parsons (2018) and Schuddeboom et al. (2018).

### 3.4 Cloud type global climatology and change with global mean near-surface temperature

We analysed global mean cloud type occurrence in the CMIP historical experiment models and the reanalyses, and change with respect to GMST in the abrupt-4xCO2 experiment. The abrupt-4xCO2 experiment was chosen because it (1) is commonly used for the determination of ECS and cloud feedback, (2) provides a strong forcing by greenhouse gases and therefore a strong signal in cloud change due to increasing GMST, (3) a large number of models provide the necessary data in this experiment in the CMIP5 and CMIP6 archives. As shown in Fig. 6a comparing model global mean cloud type occurrence relative to CERES,
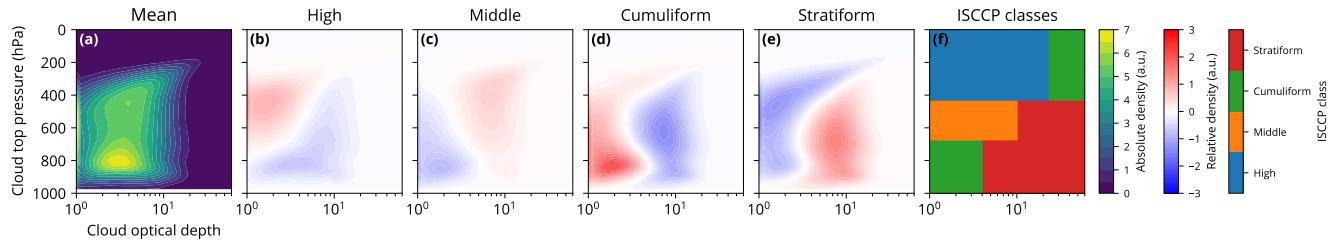
**Figure 5.** Histogram of cloud optical depth and cloud top pressure of the cloud types derived from CERES (2003–2020) by applying the ANN. **(a)** Mean of the four cloud types. **(b–e)** Difference from the mean for the classification of four cloud types. **(f)** ISCCP classification (Rossow and Schiffer, 1991).

the models exhibit a broad range of biases, but with many similarities. Underestimation of the cumuliform cloud by up to 19% is common, as well as smaller underestimation of the middle cloud type (up to 6%) and high cloud type (up to 3%). Overestimation of the stratiform type up to 5% was present in a smaller number of models. Progression from large biases to low biases with increasing model ECS is quite notable, with the exception of INM-* and to a smaller extent IPSL-CM6A-LR and CanESM5. In particular, most models with lower ECS (< 4 K) tend to underestimate the cumuliform type and some also overestimate the stratiform type, while models with higher ECS (> 4 K) to a smaller degree tend to underestimate the stratiform type and some overestimate the cumuliform type. The reanalyses (ERA5 and MERRA-2) have some of the best agreement with CERES compared to the CMIP models.

We also analysed cloud type occurrence change with respect to GMST, defined as the slope of linear regression of cloud type occurrence as a function of GMST in units of $\% \, K^{-1}$, shown in Fig. 6b. It was calculated from the first 100 years for CMIP abrupt-4xCO2, with the exception of two models for which the first 100 years were not available: for MPI-ESM-LR years 1850–1869 and 1970–1989 were used, and for MRI-CGCM3 years 1851–1870 and 1971–1990 were used (1850 is the start of the abrupt-4xCO2 experiment, and the time period is not supposed to correspond to reality). This comparison lacks a reliable observational reference because the CERES record is too short to accurately determine the slope of the regression. The abrupt-4xCO2 experiment is also not directly comparable to reality due to the different $CO_2$ and aerosol forcing. The models exhibit a broad range of values with few common trends. Models with lower ECS (< 4 K) tended to simulate increasing stratiform and middle cloud type and decreasing cumuliform type, while models with higher ECS (> 4 K) tended to simulate decreasing stratiform and middle cloud type and increasing cumuliform type. The corresponding geographical distribution plots are provided in Fig. S2 and S3.

Fig. 7 and 8 show the same as Fig. 6, but for a classification into 10 and 27 cloud types, respectively. The underestimated cumuliform type in models with relatively low ECS (< 4 K) is dominated by Cu and Cu+Sc, and the overestimated stratiform type in the same models is dominated by Sc. Underestimated Ac is a common feature of all models. The increasing stratiform cloud type with GMST in models with low ECS (< 4 K) is dominated by St and the decreasing cumuliform cloud type in the same models is dominated by Cu. The increasing cumuliform type with GMST in high ECS (> 4 K) models is dominated by
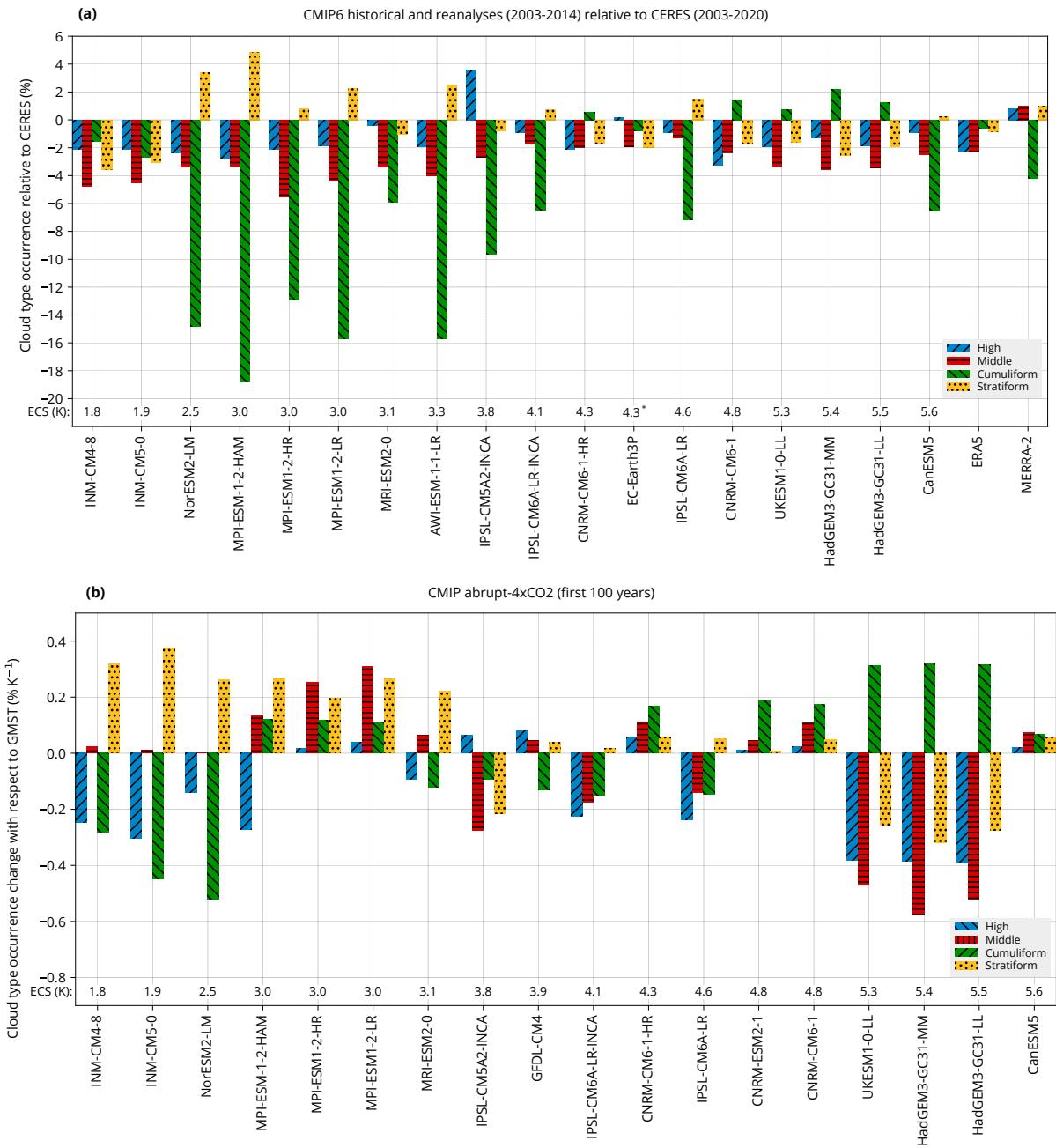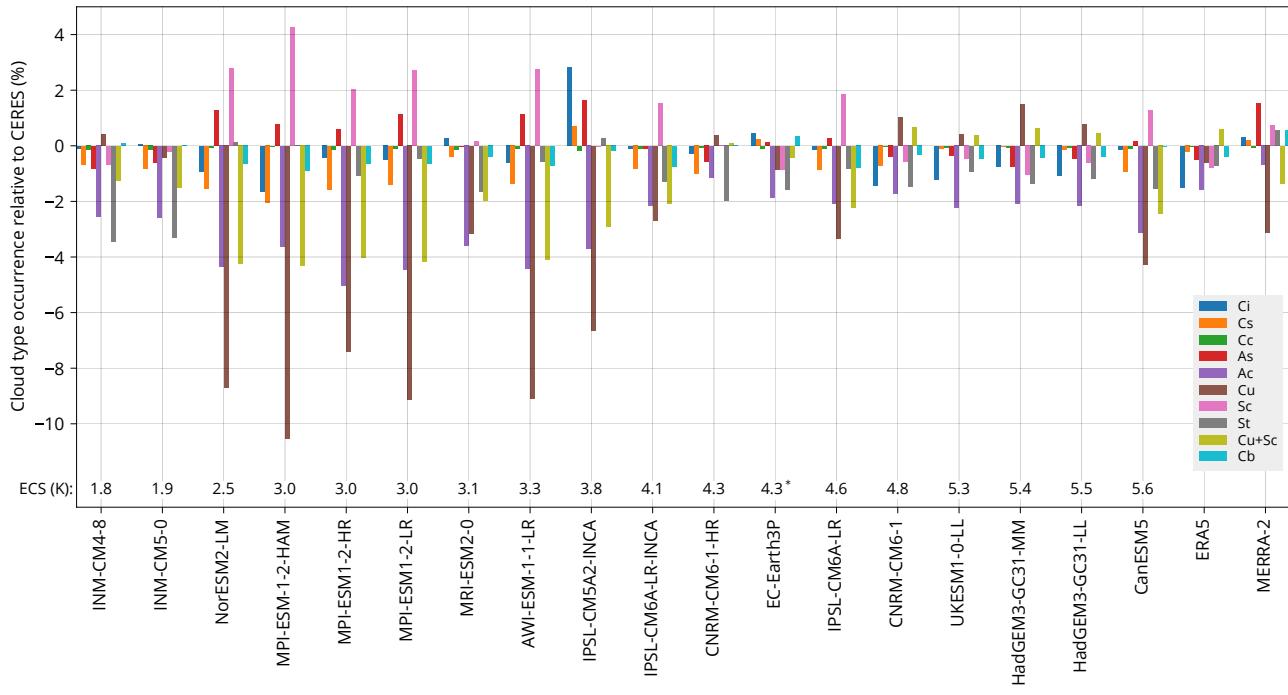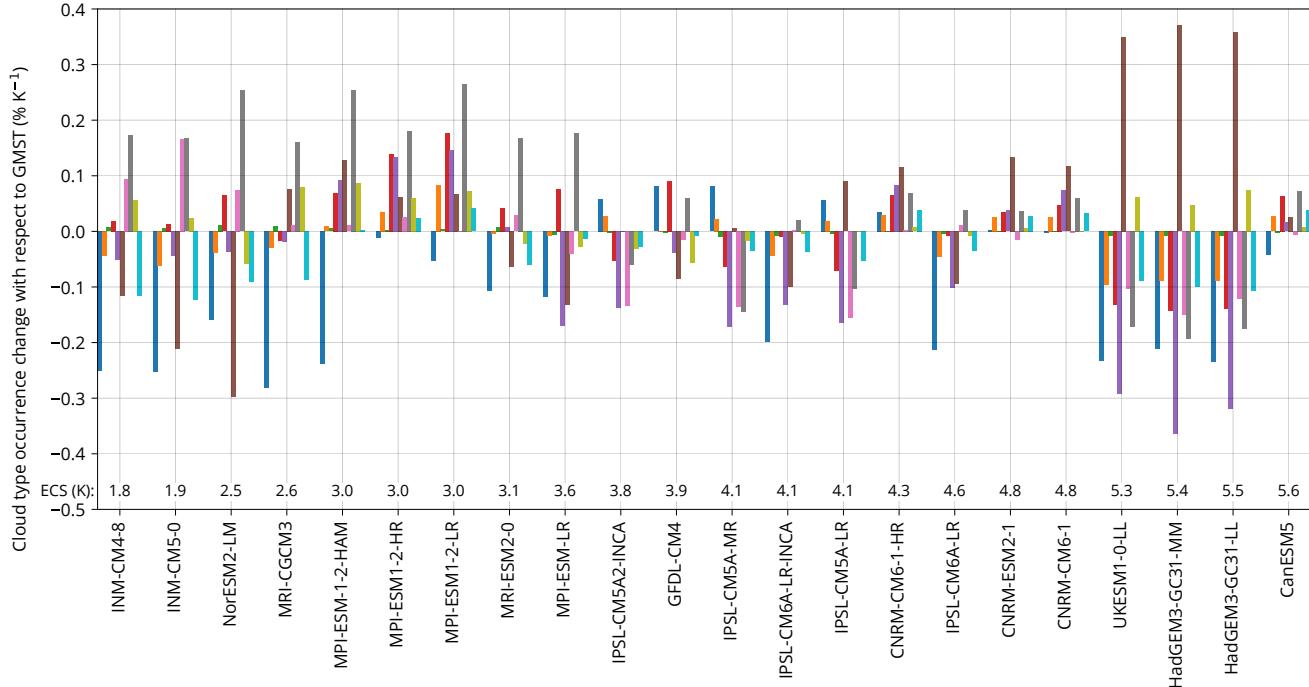
**Figure 6. (a)** Global mean cloud type occurrence in CMIP6 models and reanalyses relative to CERES in the historical experiment. **(b)** Global mean of cloud type occurrence change with respect to global mean near-surface air temperature (GMST) in CMIP5 and CMIP6, calculated by linear regression. Models are sorted by their equilibrium climate sensitivity (ECS). *For some models, ECS was not available and was taken from the closest available model (see Table 1).

**Figure 7.** The same as Fig. 6, but for a classification of 10 cloud genera.
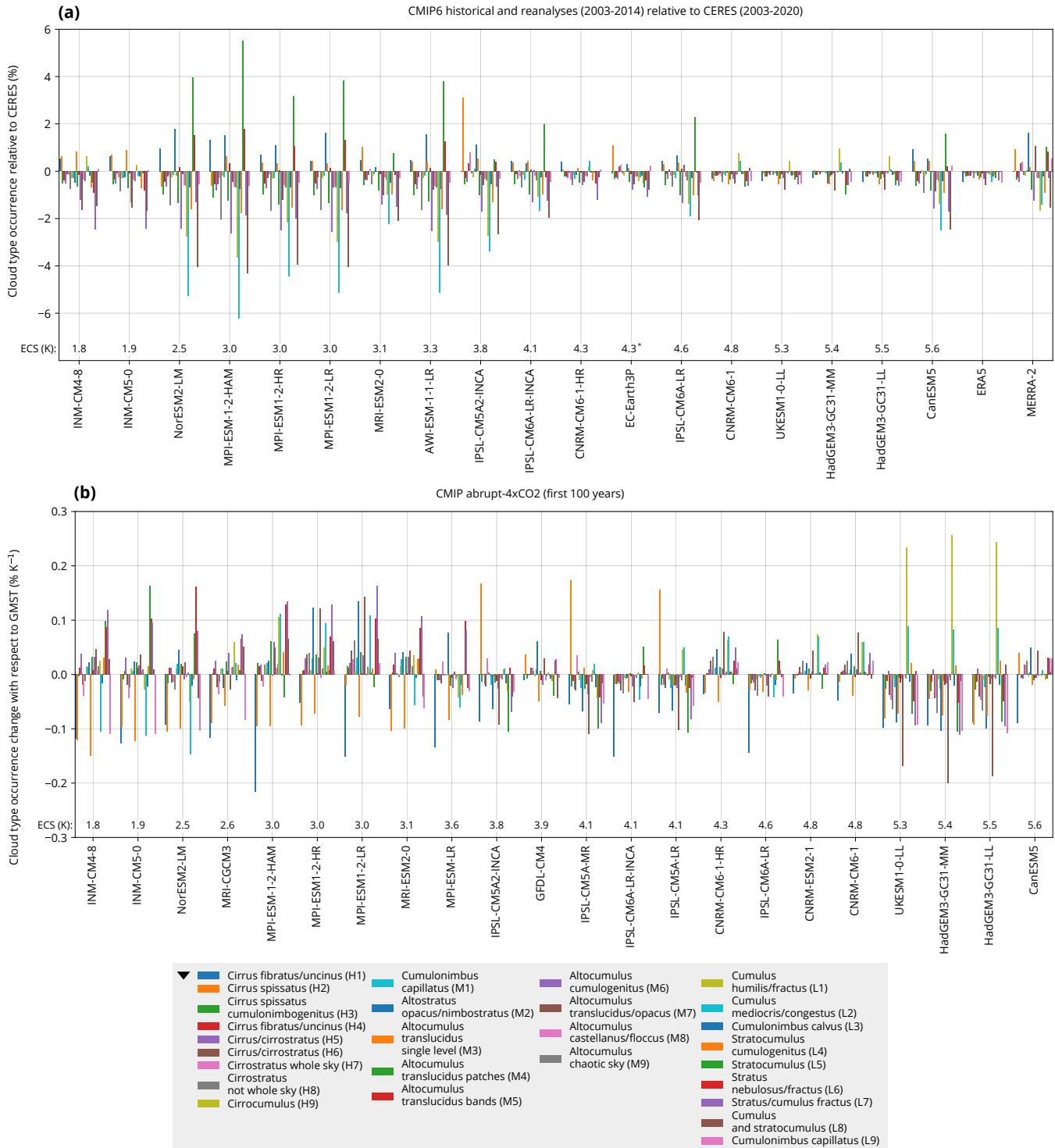
**Figure 8.** The same as Fig. 6, but for a classification of 27 cloud genera/species. Note that the colours are repeating and are ordered from left to right in the plots and top–left to bottom–right in the legend.

Cu the the decreasing stratiform type in the same models is contributed by both St and Sc. In the full classification into 27 types, the largest bias is in cumulus mediocris/congestus (L2) and stratocumulus (L5) and cumulus and stratocumulus (L8). We caution, however, the the classification into 27 cloud types should be considered with caution due to the fact that association between the low-resolution TOA radiation and the cloud type occurrence is inferential, and the individual cloud genera/species in general cannot be directly observed in the radiation fields. The corresponding geographical distribution plots are provided in the supplement.

## 3.5 Climate sensitivity

We analysed how cloud type occurrence change with respect to GMST relates to climate sensitivity. ERA5 and MERRA-2 are excluded from the analysis in this section because climate sensitivity and feedbacks are not estimated for reanalyses. Fig. 9a shows a linear regression of ECS as a function of model's cloud type occurrence change with respect to GMST. The relation of ECS with the stratiform cloud type is the strongest (probability of the null model representing no linear relationship in the data $P(M_0) = 3 \times 10^{-4}$; see Appendix A), with the cumuliform type slightly less strong ($P(M_0) = 1 \times 10^{-3}$), and with the middle cloud type relatively strong ($P(M_0) \approx 0.04$). Relation with the high cloud type was not statistically identifiable. Higher ECS is associated with decreasing stratiform and middle cloud types and increasing cumuliform cloud type with increasing GMST. This may be physically explained by the fact that the cumuliform cloud type has low optical depth compared to the stratiform type (Fig. 5), and therefore if a model simulates a transition from stratiform to cumuliform clouds with increasing GMST, radiative forcing due to cloud feedback is increased.

While the cloud type change with respect to GMST is too uncertain in the observational reference (CERES) to be be useful for quantifying the accuracy of models in the representation of this value, and also the abrupt-4xCO2 experiment assessed here is not directly comparable to reality, we can link present-day cloud biases to climate sensitivity. In Fig. 9b, c, d we show that the total RMSE of cloud type occurrence (calculated for the 27 cloud types) is linearly related to the model ECS ($P(M_0) = 2 \times 10^{-3}$), TCR ($P(M_0) = 9 \times 10^{-3}$) and cloud feedback ($P(M_0) = 1 \times 10^{-2}$). Models with the lowest RMSE tend to have the largest ECS, TCR and cloud feedback, while models with the highest RMSE tend to have the lowest ECS, TCR and cloud feedback. There are, however, several outliers such as INM-* with mid-range RMSE and the lowest ECS and TCR of all models in the ensemble, and CanESM5 with mid-range RMSE and the highest ECS. The relationship could also be artificially strong due to cross-correlation between models of the same origin. If only one model of each model family is retained (10 out of 18), the $P(M_0) = 0.29, 0.29, 0.62$ for ECS, TCR and cloud feedback, respectively (Fig. S4), i.e. the presence of a negative linear relationship is still more likely than not for ECS and TCR, but not cloud feedback, although such a test is relatively weak due to the small number of remaining models.

The above was calculated with the ANN for the full set of 27 cloud types due to its higher statistical strength (as can be expected with a more detailed classification). If done with four and 10 cloud types, $P(M_0) = 5 \times 10^{-3}$, 0.01 and 0.07 (ECS, TCR and cloud feedback, respectively) for four cloud types (Fig. S5) and $P(M_0) = 2 \times 10^{-3}$, 0.01 and 0.05 for 10 cloud types (Fig. S6). This means that the relationship holds well even for the classifications with fewer cloud types, but with lower statistical strength.
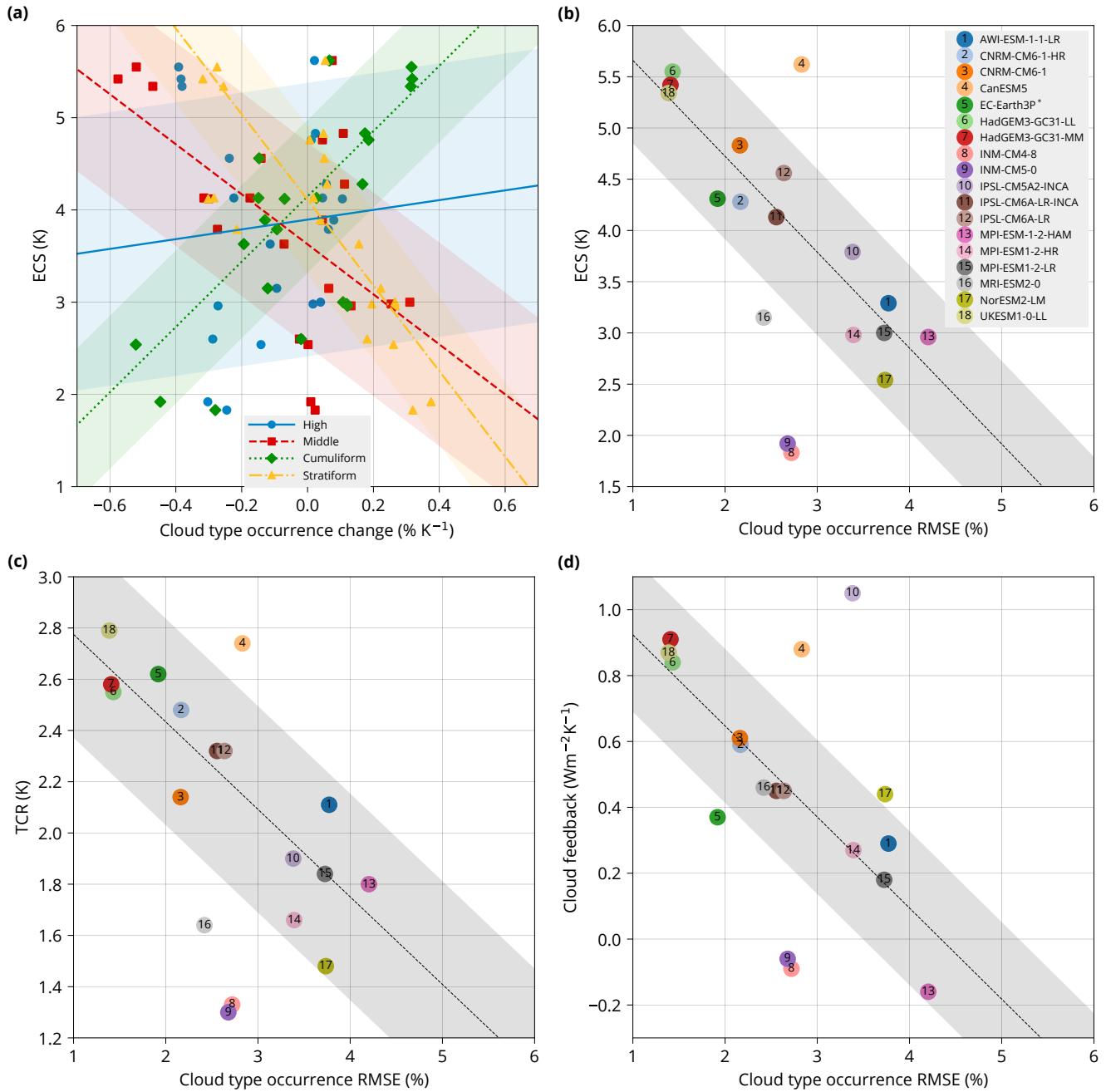
**Figure 9.** **(a)** Dependence of model equilibrium climate sensitivity (ECS) on the cloud type occurrence change with respect to GMST. **(b)** Dependence of equilibrium climate sensitivity (ECS), **(c)** transient climate response (TCR) and **(d)** climate feedback of CMIP6 models on the model total cloud type root mean square error (RMSE) relative to CERES, calculated from the geographical distribution (as in Fig. 3 and 4). **(b–d)** are calculated from the ANN for 27 cloud types. Confidence bands represent 68% range. Linear regression is calculated using Bayesian simulation assuming Cauchy error distribution (Appendix A). *For some models, ECS was not available and was taken from the closest available model (see Table 1).

## 4  Discussion and conclusions

We developed a deep convolutional ANN for the purpose of determining cloud types in retrieved and simulated TOA shortwave and longwave radiation images, trained on global historical records of human observations of WMO cloud genera (IDD). We trained this ANN to identify the probability of occurrence of each cloud type in every pixel of the image for a set of 4, 10 and 27 cloud types. We applied the ANN to satellite observations from CERES, CMIP climate model and reanalysis output to derive geographical distribution, global mean of cloud type occurrence and its change with respect to GMST. This provided a unique quantification of the distribution of WMO cloud genera globally, and enabled us to compare models and observations with this metric. Relative to IDD, the ANN could reproduce the geographical distribution of cloud type occurrence relatively well over regions where reference data were available. CMIP models and reanalyses displayed a variety of biases relative to satellite observations, most notably negative bias in the cumuliform cloud type and smaller negative bias in the middle and high cloud type. Models related in their code base often showed the same pattern and magnitude of biases. Models with lower ECS (< 4 K) had larger biases than models with higher ECS (> 4 K) and reanalyses. Analysis of the abrupt-4xCO2 experiment suggests that low ECS (< 4 K) models tend to simulate decreasing cumuliform and increasing stratiform clouds, while the opposite is true for high ECS (> 4 K) models. By linking the cloud type change with respect to GMST to ECS, we showed that models with decreasing stratiform and middle cloud type and increasing cumuliform cloud type tended to have higher ECS, a physically expected result. We investigated the link between present-day cloud biases and ECS, TCR and cloud feedback. We found that the model cloud biases are correlated with all three quantities. Models with smaller biases had higher ECS, TCR and cloud feedback than models with larger biases.

The method introduced in this study has a number of limitations. The CERES dataset is too short (2003–2020) to reliably detect change with respect to GMST. This means that an observational reference for change with respect to GMST is lacking. While the NOAA and ESA satellite series provide much longer time series, datasets derived from these satellites, the Climate Change Initiative Cloud project (Cloud_cci) (Stengel et al., 2020), the Pathfinder Atmospheres Extended (PATMOS-x) (Foster and Heidinger, 2013) and the 'CM SAF Cloud, Albedo And Surface Radiation dataset from AVHRR data' (CLARA-A2) (Karlsson et al., 2017), appear to be unreliable for determining change with respect to GMST, most likely due to changing orbit and instrument sensors. Potentially, other satellite products which provide TOA radiance information could be used instead of CERES, such as ISCCP, Multi-angle Imaging SpectroRadiometer (MISR), MODIS, CloudSat or the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO). Because an equivalent physical quantity needs to be provided from a model for the ANN to be applicable on both, a satellite simulator such as the Cloud Feedback Model Intercomparison Project (CFMIP) Observation Simulator Package (COSP) could be used to calculate such an equivalent quantity from the model output. Because the ANN was not trained to apply on pixels without SW radiation, polar regions during polar winter were not analysed. The analysed model ensemble was relatively small with several models of the same origin. Therefore, even though relatively strong statistical correlations could be identified, they rest on the assumption of statistical independence. Similar limitations as in past emergent constrains analyses apply, in that a physical explanation would need to accompany a statistical relationship for it to be confirmed.

When compared to the results produced by past clustering approaches based on SOM applied on ISCCP and MODIS (Appendix B), the ANN shows good agreement on the physical properties of clouds, but differences exist potentially due to multi-level cloud situations and low effective spatial resolution of the CERES/ANN dataset (about $5 \times 5°$). The definition of the ANN cloud types is also fundamentally different from previous methods because it is based on visual observations by humans, whereas other methods use a more synthetic approach of partitioning the cloud top pressure–optical depth space either directly (Rossow and Schiffer, 1991) or by machine learning methods such as SOM (McDonald and Parsons, 2018; Schuddeboom et al., 2018). The viewpoint also likely matters. Our method uses a hybrid top–bottom approach, where radiation fields measured from the top by satellites (or simulated by models) are used to derive cloud types corresponding to observations from the ground. Cloud top pressure–optical depth partitioning methods usually rely on radiation fields measured from the top only. Due to obscuration in multi-layer and thick cloud situations, top and bottom approaches can have very different views of reality (McErlich et al., 2021).

An important finding of our analysis is that cloud type occurrence biases in CMIP6 models show that more climate sensitive models are more consistent with observations in this metric. Zelinka et al. (2022) also recently found that the mean-state radiatively-relevant cloud properties in CMIP5 and CMIP6 models are correlated with total cloud feedback, and in particular that better simulating present-day cloud properties is associated with larger cloud feedbacks. They call this an open question for future research, although they note that 'model parameters driving variance in mean-state extratropical cloud-radiative effect across members of the HadGEM3-GA7.05 perturbed physics ensemble differ from those driving the variance in its cloud feedback (Tsushima et al., 2020).' In contrast with Zelinka et al. (2022) and our results, Schuddeboom and McDonald (2021) did not find any relation between mean or compensating cloud errors and ECS in a cloud clustering analysis, although their model ensemble was small (8 models).

During the 20–21$^{st}$ century historical record, the effect of rising greenhouse gas concentrations is obscured by the rapid rise of aerosol concentration since the 1950s (Dittus et al., 2020; Tokarska et al., 2020; Smith and Forster, 2021), masking the effect of greenhouse gases by a direct radiative effect and a highly uncertain indirect radiative effect through changing cloud formation, microphysics and lifetime. Recent understanding of climate sensitivity is represented by diverging results relative to the CMIP6 multi-model mean. de-la Cuesta and Mauritsen (2019) and Tokarska et al. (2020) estimate low ECS based on the historical record, Renoult et al. (2020) estimate low ECS based on paleoclimatic evidence from the last glacial maximum and mid-Pliocene warm period, Bjordal et al. (2020), on the other hand, show that high ECS is possible due to transition to higher Southern Ocean cloud phase change feedback with warming climate. Zhao et al. (2016) showed that it is possible to modify parametrisation of precipitation in convective plumes in the GFDL model and get different Cess sensitivities without increasing CRE error relative to CERES. Zhu et al. (2022) showed that in CESM2, a CMIP6 model with very high ECS of 6.1 K, a physically more consistent cloud microphysics parametrisation reduced the ECS to about 4 K and produced results more consistent with the last glacial maximum.

We suggest that our results about high sensitivity models being more correct in their cloud type representation should be considered with caution due to the novelty of the method, the size and cross-correlations in the model ensemble and need for a physical explanation. However, it could be considered in the context of other factors influencing ECS in a multiple-factor

analysis (Bretherton and Caldwell, 2020; Sherwood et al., 2020), especially if it should be used as an emergent constraint. While our results favour models on the high-end of ECS (in the investigated model ensemble), they are not necessarily in contradiction with Sherwood et al. (2020) or AR6, in that some models which performed well in our analysis would lie on the upper end of the very likely range, because the scope of our study is much smaller utilising only one cloud metric on a limited number of related models. Nevertheless, we think that the strength of the identified relationship and the opposing trends in cumuliform and stratiform clouds in high ($> 4$ K) and low ($> 4$ K) ECS models with increasing GMST warrant a further investigation of links between present-day cloud simulation biases and projected future cloud change, and demonstrates that the ANN method of cloud identification can be a useful tool for climate model evaluation.

## Appendix A:  Linear regression Bayesian model comparison

The linear regression model $M_1$ representing the alternative hypothesis and the null hypothesis model $M_0$ are defined as:

$$M_1 : \quad y = \alpha x + \beta + \epsilon, \tag{A1}$$
$$M_0 : \quad y = \beta + \epsilon, \tag{A2}$$
$$\alpha = \tan(\varphi), \tag{A3}$$
$$\varphi \sim \text{Uniform}(-\pi/2, \pi/2), \tag{A4}$$
$$\beta \sim \text{Uniform}(-100, 100), \tag{A5}$$
$$\epsilon \sim \text{Cauchy}(0, \gamma), \tag{A6}$$

where $x$ is a vector of the independent variables, $y$ is a vector of the dependent variables, $\alpha$ and $\beta$ are the slope and intercept, respectively, $\epsilon$ is a Cauchy-distributed random error, $\gamma$ is the scale parameter of the Cauchy distribution, and $\varphi$ is the angle of the slope. $\varphi$ and $\beta$ come from a continuous uniform prior distribution. The statistical distribution of the free parameters $\varphi$, $\beta$ and the Bayes factor ($P(M_1|x,y)/P(M_0|x,y)$) were determined using the Metropolis algorithm (Metropolis et al., 1953), and simulated with the Python library PyMC3 version 3.11.2 (Salvatier et al., 2016). The prior probability of $M_0$ and $M_1$ was assumed to be equal: $P(M_0) = P(M_1) = 0.5$. Before running the simulation, the variables $x$ and $y$ were normalised by their mean and standard deviation.

## Appendix B:  Comparison with cloud clusters derived using self-organising maps

To understand how the cloud types determined by the ANN relate to cloud clusters constructed by previous studies, we perform a comparison with cloud clusters of Schuddeboom et al. (2018) and McDonald and Parsons (2018), generated by a machine learning method known as self-organising maps (SOM). They use SOM to identify representative cloud clusters using cloud top pressure–cloud optical depth joint histograms from Moderate Resolution Imaging Spectroradiometer (MODIS) (Schuddeboom et al., 2018) and ISCCP (McDonald and Parsons, 2018). They establish characteristics of these clusters by investigating how

they relate to cloud properties. Schuddeboom et al. (2018) uses the clusters to examine model representation of different cloud types while McDonald and Parsons (2018) focuses specifically on how their clusters relate to atmospheric dynamics.

Here we calculate the average CERES/ANN values for each of our four types for every ISCCP/SOM and MODIS/SOM cluster. The CERES/ANN geographical distribution is available on a $5° \times 5°$ global grid (the original grid is $2.5° \times 2.5°$, but the effective resolution is lower), while the ISCCP/SOM data are on a $2.5° \times 2.5°$ grid and MODIS/SOM data are on a $1° \times 1°$ grid. To account for this difference in spatial resolution, all of the ISCCP/SOM and MODIS/SOM grid cells that fall within a corresponding ANN geographical distribution grid cell are considered as having the same occurrence values. This will overestimate the similarity between the clusters, as the small cloud structures that can be identified in the higher resolution dataset will be merged.

In Fig. B1 we present calculated co-occurrence of the CERES/ANN cloud types with the 15 ISCCP/SOM clusters of McDonald and Parsons (2018) and 12 MODIS/SOM clusters of Schuddeboom et al. (2018). The left–right and top–bottom ordering of the ISCCP/SOM and MODIS/SOM grids is the result of the SOM algorithm, from which these clusters were derived. This algorithm results in neighbouring clusters which are closely related, with the most distinct clusters the most distant. For example, in the ISCCP/SOM grid the top row relates to clouds with low cloud top pressure while the bottom row relates to high cloud top pressure. From understanding this relationship, we can see that ordering of the CERES/ANN values suggests good separation into physically distinct cloud types. The values shown suggest small to moderate amounts of every cloud type are present regardless of the ISCCP/MODIS cluster present. This could be at least partially explained by the spatial smoothing effect described above, as well as co-occurrence of different cloud types in a single geographical grid cell.

By considering individual clusters in Fig. B1 and examining their cloud top pressure–cloud optical depth diagrams (McDonald and Parsons, 2018, Fig. 1) and (Schuddeboom et al., 2018, Fig. 2), we can see that they show the expected physical relationship. The CERES/ANN high cloud type is identified as co-occurring most strongly with ISCCP/SOM clusters 1, 6, 7, and 11, which are also the SOM clusters corresponding to high clouds (McDonald and Parsons, 2018, Fig. 1). The CERES/ANN middle cloud type is most strongly associated with ISCCP/SOM clusters 2, 1 and 4 (numbers ordered by the strength of association), which all contain substantial amount of semi-opaque clouds at 180–680 hPa. The CERES/ANN cumuliform cloud type is most strongly associated with ISCCP/SOM clusters 13, 6, 11 and 12. While cluster ISCCP/SOM 13 has a local maximum at low cloud of low-to-mid optical depth, clusters 6, 11 and 12 are less clearly associated with low clouds (they have maximum for high low optical depth clouds), although they still contain substantial amounts of low altitude low optical depth clouds. The CERES/ANN stratiform cloud type is most strongly associated with the ISCCP/SOM clusters 4, 5, 2, 10 and 15. The ISCCP/SOM clusters 10 and 15 are strongly associated with low-altitude mid-to-high optical depth clouds, as expected for stratiform clouds. The ISCCP/SOM clusters 4 and 5 are mostly composed of mid-altitude mid-optical depth clouds. The ISCCP/SOM cluster 2, however, is mostly associated with relatively high clouds above 680 hPa.

The CERES/ANN high cloud type co-occurs most strongly with the MODIS/SOM cluster 1. This cluster has maximum for high-altitude low optical depth clouds (Schuddeboom et al., 2018, Fig. 2) and is identified as tropical. The CERES/ANN middle cloud type is most strongly associated with the MODIS/SOM clusters 9 and 10, identified as mixed level clouds. The MODIS/SOM cluster 9 has the greatest contribution from relatively high clouds above 310 hPa, but substantial amount
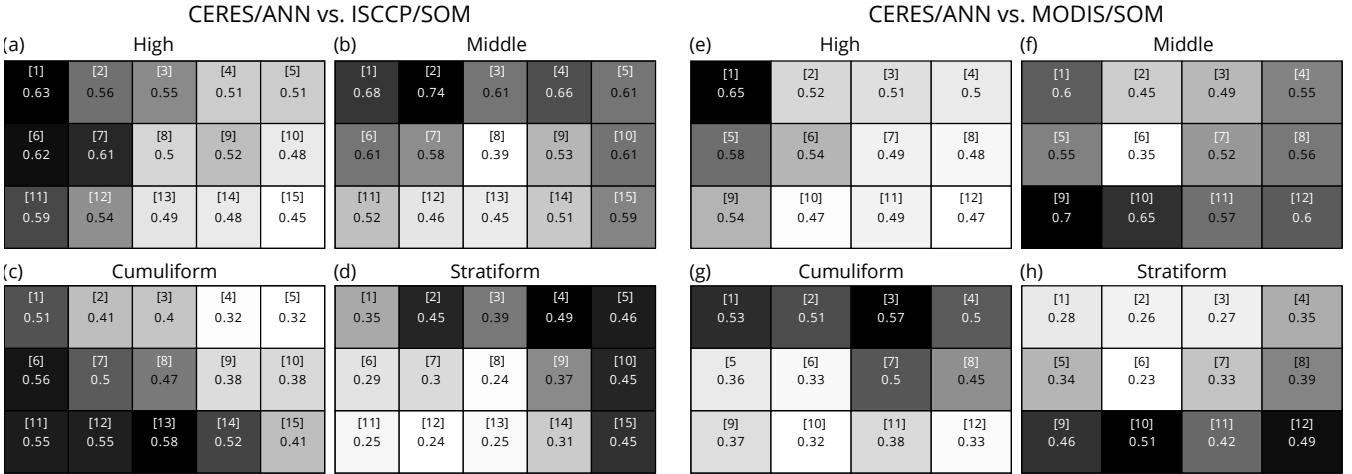
**Figure B1.** Comparison of co-occurrence of our ANN-derived cloud types (CERES/ANN) with SOM-derived cloud clusters (ISCCP/SOM and MODIS/SOM) of McDonald and Parsons (2018) and Schuddeboom et al. (2018), respectively. The grids in **(a–d)** and **(e–h)** represent our four ANN-derived cloud types (high, middle, cumuliform and stratiform). The grid boxes in each subplot correspond to the SOM-derived cloud clusters of the past studies. Numbers in square brackets in the boxes are the ISCCP/SOM and MODIS/SOM cluster numbers as in the original studies. Numbers in the centre of the boxes and box shading are the co-occurrence (scale 0–1) of the ANN-derived cloud type and the SOM-derived cloud cluster. The co-occurrence is calculated from one year (2007) of daily mean values on a global spatial grid. Note that the definition of the ISCCP/SOM and MODIS/SOM clusters is different, and therefore **(a–d)** and **(e–h)** are not expected to be similar.

of clouds at altitudes of 180–800 hPa. The MODIS/SOM cluster 10 has greatest contribution from clouds at relatively low altitude of 680–800 hPa, but also high clouds above 440 hPa. The CERES/ANN cumuliform cloud type is associated with the
515 MODIS/SOM clusters 3, 1, 2, 4, 7 and 8. These are identified as marine or tropical, and have strong contribution from low-altitude low-to-mid optical depth clouds. The CERES/ANN stratiform cloud type is associated with MODIS/SOM clusters 10, 12, 9 and 11. These are identified as mixed level and stratocumulus clouds, and have a strong contribution from low altitude mid-to-high optical depth clouds (clusters 11 and 12) and clouds at various altitudes (clusters 9 and 10).

To summarise, the correspondence between the CERES/ANN cloud types and ISCCP/SOM and MODIS/SOM clusters
520 is relatively good when compared using the cloud top pressure–cloud optical depth diagrams. However, differences exist particularly in cloud types related to mixed level cloud situations.

*Code and data availability.* The datasets used in our analysis are publicly available: CERES (https://ceres.larc.nasa.gov/data/), GISTEMPv4 (https://data.giss.nasa.gov/gistemp/, CMIP5 (https://esgf-node.llnl.gov/search/cmip5/), CMIP6 (https://esgf-node.llnl.gov/projects/cmip6/), MERRA-2 (https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/), ERA (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5),
525 and IDD (https://rda.ucar.edu/datasets/ds336.0/). The code used in our analysis is open source and available on GitHub (https://github.com/peterkuma/ml-clouds-2021) and Zenodo (Kuma et al., 2022).

*Author contributions.* PK participated on conceptualisation and methodology development, developed the artificial neural network model, performed the data analysis and wrote the manuscript; FB participated on conceptualisation and methodology development, review and editing of the manuscript, funding acquisition and project administration. AS and AM performed the comparison with MODIS and ISCCP
530 cloud regimes and reviewed the manuscript. ØS participated on NorESM2-LM data preparation and acquisition, consultation of the analysis and reviewed the manuscript.

*Competing interests.* The authors declare that they have no conflict of interest.

*Acknowledgements.* This work was conducted as part of the FORCeS project: 'Constrained aerosol forcing for improved climate projections' (https://forces-project.eu), funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number
535 821205. We also acknowledge funding from the Swedish e-Science Research Centre (SeRC). AS and AM acknowledge funding from New Zealand's Deep South National Science Challenge 'Cloud and Aerosol Measurements for Improved Climate Projections'. We thank Hossein Azizpour for his consultation and advice on the artificial neural network development. We acknowledge the CERES dataset provided by the NASA Langley Research Center, the IDD dataset provided by Unidata and the University Corporation for Atmospheric Research through the Research Data Archive at the National Center for Atmospheric Research, the GISTEMP dataset provided by the NASA Goddard Institute
540 for Space Studies, the ERA5 dataset provided by the ECMWF through the Copernicus Climate Change Service, and the MERRA-2 dataset provided by the Global Modeling and Assimilation Office, NASA Goddard Space Flight Center Greenbelt. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP5 and CMIP6. We thank the climate modelling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP5, CMIP6 and ESGF. We acknowledge open
545 source software used in our analysis: TensorFlow (Abadi et al., 2016), Python, NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), Matplotlib, cartopy (Met Office, 2010), PyMC3 (Salvatier et al., 2016), parallel (Tange et al., 2011), Pandas (The pandas development team, 2020), pyproj, Cython (Behnel et al., 2011), aria2 and Devuan GNU+Linux.

# References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X.: Tensor-Flow: A System for Large-Scale Machine Learning, in: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16, pp. 265–283, USENIX Association, USA, 2016.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K.: Cython: The Best of Both Worlds, Computing in Science Engineering, 13, 31–39, https://doi.org/10.1109/MCSE.2010.118, 2011.

Bjordal, J., Storelvmo, T., Alterskjær, K., and Carlsen, T.: Equilibrium climate sensitivity above 5 °C plausible due to state-dependent cloud feedback, Nature Geoscience, 13, 718–721, https://doi.org/10.1038/s41561-020-00649-1, 2020.

Bretherton, C. S. and Caldwell, P. M.: Combining Emergent Constraints for Climate Sensitivity, Journal of Climate, 33, 7413–7430, https://doi.org/10.1175/JCLI-D-19-0911.1, 2020.

Cesana, G., Del Genio, A. D., and Chepfer, H.: The Cumulus And Stratocumulus CloudSat-CALIPSO Dataset (CASCCAD), Earth System Science Data, 11, 1745–1764, https://doi.org/10.5194/essd-11-1745-2019, 2019.

Cho, N., Tan, J., and Oreopoulos, L.: Classifying Planetary Cloudiness with an Updated Set of MODIS Cloud Regimes, Journal of Applied Meteorology and Climatology, 60, 981–997, https://doi.org/10.1175/JAMC-D-20-0247.1, 2021.

de-la Cuesta, D. J. and Mauritsen, T.: Emergent constraints on Earth's transient and equilibrium response to doubled CO2 from post-1970s global warming, Nature Geoscience, 12, 902–905, https://doi.org/10.1038/s41561-019-0463-y, 2019.

Dittus, A. J., Hawkins, E., Wilcox, L. J., Sutton, R. T., Smith, C. J., Andrews, M. B., and Forster, P. M.: Sensitivity of Historical Climate Simulations to Uncertain Aerosol Forcing, Geophysical Research Letters, 47, e2019GL085 806, https://doi.org/https://doi.org/10.1029/2019GL085806, e2019GL085806 10.1029/2019GL085806, 2020.

Doelling, D. R., Loeb, N. G., Keyes, D. F., Nordeen, M. L., Morstad, D., Nguyen, C., Wielicki, B. A., Young, D. F., and Sun, M.: Geostationary Enhanced Temporal Interpolation for CERES Flux Products, Journal of Atmospheric and Oceanic Technology, 30, 1072–1090, https://doi.org/10.1175/JTECH-D-12-00136.1, 2013.

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., Zhou, C., and Andrews, T.: Intermodel Spread in the Pattern Effect and Its Contribution to Climate Sensitivity in CMIP5 and CMIP6 Models, Journal of Climate, 33, 7755–7775, https://doi.org/10.1175/JCLI-D-19-1011.1, 2020.

Drönner, J., Korfhage, N., Egli, S., Mühling, M., Thies, B., Bendix, J., Freisleben, B., and Seeger, B.: Fast Cloud Segmentation Using Convolutional Neural Networks, Remote Sensing, 10, https://doi.org/10.3390/rs10111782, 2018.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Flynn, C. M. and Mauritsen, T.: On the climate sensitivity and historical warming evolution in recent coupled model ensembles, Atmospheric Chemistry and Physics, 20, 7829–7842, https://doi.org/10.5194/acp-20-7829-2020, 2020.

Forster, P. M., Maycock, A. C., McKenna, C. M., and Smith, C. J.: Latest climate models confirm need for urgent mitigation, Nature Climate Change, 10, 7–10, https://doi.org/10.1038/s41558-019-0660-0, 2020.

Foster, M. J. and Heidinger, A.: PATMOS-x: Results from a Diurnally Corrected 30-yr Satellite Cloud Climatology, Journal of Climate, 26, 414–425, https://doi.org/10.1175/JCLI-D-11-00666.1, 2013.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), Journal of Climate, 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

GISTEMP Team: GISS Surface Temperature Analysis (GISTEMP), version 4, https://data.giss.nasa.gov/gistemp/, last access: 7 December 2021, 2021.

Guo, Y., Cao, X., Liu, B., and Gao, M.: Cloud Detection for Satellite Imagery Using Attention-Based U-Net Convolutional Neural Network, Symmetry, 12, https://doi.org/10.3390/sym12061056, 2020.

Haarsma, R., Acosta, M., Bakhshi, R., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., Corti, S., Davini, P., Exarchou, E., Fabiano, F., Fladrich, U., Fuentes Franco, R., García-Serrano, J., von Hardenberg, J., Koenigk, T., Levine, X., Meccia, V. L., van Noije, T., van den Oord, G., Palmeiro, F. M., Rodrigo, M., Ruprich-Robert, Y., Le Sager, P., Tourigny, E., Wang, S., van Weele, M., and Wyser, K.: HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – description, model computational performance and basic validation, Geoscientific Model Development, 13, 3507–3527, https://doi.org/10.5194/gmd-13-3507-2020, 2020.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J., and von Storch, J.-S.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, Geoscientific Model Development, 9, 4185–4208, https://doi.org/10.5194/gmd-9-4185-2016, 2016.

Hahn, C. J., Rossow, W. B., and Warren, S. G.: ISCCP Cloud Properties Associated with Standard Cloud Types Identified in Individual Surface Observations, Journal of Climate, 14, 11–28, https://doi.org/10.1175/1520-0442(2001)014<0011:ICPAWS>2.0.CO;2, 2001.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

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

620    Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The Art and Science of Climate Model Tuning, Bulletin of the American Meteorological Society, 98, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1, 2017.

Jakob, C. and Tselioudis, G.: Objective identification of cloud regimes in the Tropical Western Pacific, Geophysical Research Letters, 30, https://doi.org/https://doi.org/10.1029/2003GL018367, 2003.

625    Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Fokke Meirink, J., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., Benas, N., van Zadelhoff, G.-J., Schlundt, C., Stein, D., Finkensieper, S., Håkansson, N., and Hollmann, R.: CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data, Atmospheric Chemistry and Physics, 17, 5809–5828, https://doi.org/10.5194/acp-17-5809-2017, 2017.

Kuma, P., Bender, F. A.-M., Schuddeboom, A., McDonald, A. J., and Seland, Ø.: Code accompanying the manuscript "Machine learning of
630    cloud types shows higher climate sensitivity is associated with lower cloud biases", https://doi.org/10.5281/zenodo.6164982, 2022.

Lenssen, N., Schmidt, G., Hansen, J., Menne, M., Persin, A., Ruedy, R., and Zyss, D.: Improvements in the GISTEMP uncertainty model, J. Geophys. Res. Atmos., 124, 6307–6326, https://doi.org/10.1029/2018JD029522, 2019.

Liu, C., Yang, S., Di, D., Yang, Y., Zhou, C., Hu, X., and Sohn, B.-J.: A Machine Learning-based Cloud Detection Algorithm for the Himawari-8 Spectral Image, Advances in Atmospheric Sciences, https://doi.org/10.1007/s00376-021-0366-x, 2021.

635    Liu, S. and Li, M.: Deep multimodal fusion for ground-based cloud classification in weather station networks, EURASIP Journal on Wireless Communications and Networking, 2018, https://doi.org/10.1186/s13638-018-1062-0, 2018.

Loeb, N., Su, W., Doelling, D., Wong, T., Minnis, P., Thomas, S., and Miller, W.: 5.03 - Earth's Top-of-Atmosphere Radiation Budget, in: Comprehensive Remote Sensing, edited by Liang, S., pp. 67–84, Elsevier, Oxford, https://doi.org/https://doi.org/10.1016/B978-0-12-409548-9.10367-7, 2018.

640    Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., eds.: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom, in press, 2021.

McDonald, A. J. and Parsons, S.: A Comparison of Cloud Classification Methodologies: Differences Between Cloud and Dynamical
645    Regimes, Journal of Geophysical Research: Atmospheres, 123, 11,173–11,193, https://doi.org/https://doi.org/10.1029/2018JD028595, 2018.

McDonald, A. J., Cassano, J. J., Jolly, B., Parsons, S., and Schuddeboom, A.: An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states, Journal of Geophysical Research: Atmospheres, 121, 13 009–13 030, https://doi.org/https://doi.org/10.1002/2016JD025199, 2016.

650    McErlich, C., McDonald, A., Schuddeboom, A., and Silber, I.: Comparing Satellite- and Ground-Based Observations of Cloud Occurrence Over High Southern Latitudes, Journal of Geophysical Research: Atmospheres, 126, e2020JD033 607, https://doi.org/https://doi.org/10.1029/2020JD033607, e2020JD033607 2020JD033607, 2021.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Science Advances, 6, eaba1981,
655    https://doi.org/10.1126/sciadv.aba1981, 2020.

Met Office: Cartopy: a cartographic python library with a Matplotlib interface, Exeter, Devon, https://scitools.org.uk/cartopy, 2010.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equation of State Calculations by Fast Computing Machines, The Journal of Chemical Physics, 21, 1087–1092, https://doi.org/10.1063/1.1699114, 1953.

Nijsse, F. J. M. M., Cox, P. M., and Williamson, M. S.: Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models, Earth System Dynamics, 11, 737–750, https://doi.org/10.5194/esd-11-737-2020, 2020.

Olsson, B., Ynnerman, A., and Lenz, R.: Computing synthetic satellite images from weather prediction data, in: Visualization and Data Analysis 2004, edited by Erbacher, R. F., Chen, P. C., Roberts, J. C., Gröhn, M. T., and Börner, K., vol. 5295, pp. 296–304, International Society for Optics and Photonics, SPIE, https://doi.org/10.1117/12.526829, 2004.

Oreopoulos, L., Cho, N., Lee, D., and Kato, S.: Radiative effects of global MODIS cloud regimes, Journal of Geophysical Research: Atmospheres, 121, 2299–2317, https://doi.org/https://doi.org/10.1002/2015JD024502, 2016.

Renoult, M., Annan, J. D., Hargreaves, J. C., Sagoo, N., Flynn, C., Kapsch, M.-L., Li, Q., Lohmann, G., Mikolajewicz, U., Ohgaito, R., Shi, X., Zhang, Q., and Mauritsen, T.: A Bayesian framework for emergent constraints: case studies of climate sensitivity with PMIP, Climate of the Past, 16, 1715–1735, https://doi.org/10.5194/cp-16-1715-2020, 2020.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/ARXIV.1505.04597, 2015.

Rossow, W. B. and Schiffer, R. A.: ISCCP Cloud Data Products, Bulletin of the American Meteorological Society, 72, 2–20, https://doi.org/10.1175/1520-0477(1991)072<0002:ICDP>2.0.CO;2, 1991.

Rossow, W. B. and Schiffer, R. A.: Advances in Understanding Clouds from ISCCP, Bulletin of the American Meteorological Society, 80, 2261–2288, https://doi.org/10.1175/1520-0477(1999)080<2261:AIUCFI>2.0.CO;2, 1999.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C.: Probabilistic programming in Python using PyMC3, PeerJ Computer Science, 2, e55, https://doi.org/10.7717/peerj-cs.55, 2016.

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., and Eyring, V.: Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6?, Earth System Dynamics, 11, 1233–1258, https://doi.org/10.5194/esd-11-1233-2020, 2020.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers, Geoscientific Model Development, 10, 3207–3223, https://doi.org/10.5194/gmd-10-3207-2017, 2017.

Schuddeboom, A., McDonald, A. J., Morgenstern, O., Harvey, M., and Parsons, S.: Regional Regime-Based Evaluation of Present-Day General Circulation Model Cloud Simulations Using Self-Organizing Maps, Journal of Geophysical Research: Atmospheres, 123, 4259–4272, https://doi.org/https://doi.org/10.1002/2017JD028196, 2018.

Schuddeboom, A. J. and McDonald, A. J.: The Southern Ocean Radiative Bias, Cloud Compensating Errors, and Equilibrium Climate Sensitivity in CMIP6 Models, Journal of Geophysical Research: Atmospheres, 126, 1–16, https://doi.org/10.1029/2021JD035310, 2021.

Segal-Rozenhaimer, M., Li, A., Das, K., and Chirayath, V.: Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN), Remote Sensing of Environment, 237, 111 446, https://doi.org/https://doi.org/10.1016/j.rse.2019.111446, 2020.

Semmler, T., Jungclaus, J., Danek, C., Goessling, H. F., Koldunov, N. V., Rackow, T., and Sidorenko, D.: Ocean Model
695    Formulation Influences Transient Climate Response, Journal of Geophysical Research: Oceans, 126, e2021JC017 633,
https://doi.org/https://doi.org/10.1029/2021JC017633, e2021JC017633 2021JC017633, 2021.

Shell, K. M., Kiehl, J. T., and Shields, C. A.: Using the Radiative Kernel Technique to Calculate Climate Feedbacks in NCAR's Community
Atmospheric Model, Journal of Climate, 21, 2269–2282, https://doi.org/10.1175/2007JCLI2044.1, 2008.

Shendryk, Y., Rist, Y., Ticehurst, C., and Thorburn, P.: Deep learning for multi-modal classification of cloud, shadow and land
700    cover scenes in PlanetScope and Sentinel-2 imagery, ISPRS Journal of Photogrammetry and Remote Sensing, 157, 124–136,
https://doi.org/https://doi.org/10.1016/j.isprsjprs.2019.08.018, 2019.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D.,
Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., von der Heydt, A. S.,
Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.:
705    An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Reviews of Geophysics, 58, e2019RG000 678,
https://doi.org/https://doi.org/10.1029/2019RG000678, e2019RG000678 2019RG000678, 2020.

Shi, C., Wang, C., Wang, Y., and Xiao, B.: Deep Convolutional Activations-Based Features for Ground-Based Cloud Classification, IEEE
Geoscience and Remote Sensing Letters, 14, 816–820, https://doi.org/10.1109/lgrs.2017.2681658, 2017.

Smith, C. J. and Forster, P. M.: Suppressed Late-20th Century Warming in CMIP6 Models Explained by Forcing and Feedbacks, Geophysical
710    Research Letters, 48, e2021GL094 948, https://doi.org/https://doi.org/10.1029/2021GL094948, e2021GL094948 2021GL094948, 2021.

Soden, B. J., Held, I. M., Colman, R., Shell, K. M., Kiehl, J. T., and Shields, C. A.: Quantifying Climate Feedbacks Using Radiative Kernels,
Journal of Climate, 21, 3504–3520, https://doi.org/10.1175/2007JCLI2110.1, 2008.

Stengel, M., Stapelberg, S., Sus, O., Finkensieper, S., Würzler, B., Philipp, D., Hollmann, R., Poulsen, C., Christensen, M., and McGarragh,
G.: Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global
715    cloud and radiation properties, Earth System Science Data, 12, 41–60, https://doi.org/10.5194/essd-12-41-2020, 2020.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., eds.: Climate Change
2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on
Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Tange, O. et al.: Gnu parallel-the command-line power tool, The USENIX Magazine, 36, 42–47, 2011.

720    Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological
Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

The pandas development team: pandas-dev/pandas: Pandas, https://doi.org/10.5281/zenodo.3509134, 2020.

Tiedtke, M.: Representation of Clouds in Large-Scale Models, Monthly Weather Review, 121, 3040–3061, https://doi.org/10.1175/1520-
0493(1993)121<3040:ROCILS>2.0.CO;2, 1993.

725    Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., and Knutti, R.: Past warming trend constrains future
warming in CMIP6 models, Science Advances, 6, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549, 2020.

Unidata, U. C. f. A. R.: Historical Unidata Internet Data Distribution (IDD) Global Observational Data, https://doi.org/10.5065/9235-WJ24,
2003.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J.,
730    van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat,
İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald,

A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and and, Y. V.-B.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Volodin, E.: The Mechanisms of Cloudiness Evolution Responsible for Equilibrium Climate Sensitivity in Climate Model INM-CM4-8, Geophysical Research Letters, 48, e2021GL096 204, https://doi.org/https://doi.org/10.1029/2021GL096204, e2021GL096204 2021GL096204, 2021.

Wessel, P. and Smith, W. H. F.: A global, self-consistent, hierarchical, high-resolution shoreline database, Journal of Geophysical Research: Solid Earth, 101, 8741–8743, https://doi.org/https://doi.org/10.1029/96JB00104, 1996.

Wessel, P. and Smith, W. H. F.: Global Self-consistent, Hierarchical, High-resolution Geography Database Version 2.3.7, https://www.soest.hawaii.edu/pwessel/gshhg/, last access: 14 February 2022, 2017.

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B., Smith, G. L., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment, Bulletin of the American Meteorological Society, 77, 853–868, https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2, 1996.

WMO: Manual on Codes - International Codes, Volume I.1, Annex II to the WMO Technical Regulations: part A - Alphanumeric Codes, World Meteorological Organization (WMO), 2019 edition edn., 2011.

WMO: International Cloud Atlas: Manual on the Observation of Clouds and Other Meteors (WMO-No. 407), https://cloudatlas.wmo.int, 2021a.

WMO: Global Observing System, https://public.wmo.int/en/programmes/global-observing-system, 2021b.

Wohlfarth, K., Schröer, C., Klaß, M., Hakenes, S., Venhaus, M., Kauffmann, S., Wilhelm, T., and Wohler, C.: Dense Cloud Classification on Multispectral Satellite Imagery, in: 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), pp. 1–6, https://doi.org/10.1109/PRRS.2018.8486379, 2018.

Wyser, K., van Noije, T., Yang, S., von Hardenberg, J., O'Donnell, D., and Döscher, R.: On the increased climate sensitivity in the EC-Earth model from CMIP5 to CMIP6, Geoscientific Model Development, 13, 3465–3474, https://doi.org/10.5194/gmd-13-3465-2020, 2020.

Ye, L., Cao, Z., and Xiao, Y.: DeepCloud: Ground-Based Cloud Image Categorization Using Deep Convolutional Features, IEEE Transactions on Geoscience and Remote Sensing, 55, 5729–5740, https://doi.org/10.1109/TGRS.2017.2712809, 2017.

Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., and Watson-Parris, D.: Cumulo: A Dataset for Learning Cloud Classes, 2020.

Zelinka, M. D.: Tables of ECS, Effective Radiative Forcing, and Radiative Feedbacks, https://github.com/mzelinka/cmip56_forcing_feedback_ecs, last access: 26 January 2022, 2021.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophysical Research Letters, 47, e2019GL085 782, https://doi.org/https://doi.org/10.1029/2019GL085782, e2019GL085782 10.1029/2019GL085782, 2020.

Zelinka, M. D., Klein, S. A., Qin, Y., and Myers, T. A.: Evaluating Climate Models' Cloud Feedbacks Against Expert Judgment, Journal of Geophysical Research: Atmospheres, 127, e2021JD035 198, https://doi.org/https://doi.org/10.1029/2021JD035198, e2021JD035198 2021JD035198, 2022.

Zhang, J., Liu, P., Zhang, F., and Song, Q.: CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network, Geophysical Research Letters, 45, 8665–8672, https://doi.org/https://doi.org/10.1029/2018GL077787, 2018.

Zhao, M., Golaz, J.-C., Held, I. M., Ramaswamy, V., Lin, S.-J., Ming, Y., Ginoux, P., Wyman, B., Donner, L. J., Paynter, D., and Guo, H.: Uncertainty in Model Climate Sensitivity Traced to Representations of Cumulus Precipitation Microphysics, Journal of Climate, 29, 543–560, https://doi.org/10.1175/JCLI-D-15-0191.1, 2016.

Zhu, J., Poulsen, C. J., and Otto-Bliesner, B. L.: High climate sensitivity in CMIP6 model not supported by paleoclimate, Nature Climate Change, 10, 378–379, https://doi.org/10.1038/s41558-020-0764-6, 2020.

Zhu, J., Otto-Bliesner, B. L., Brady, E. C., Poulsen, C. J., Tierney, J. E., Lofverstrom, M., and DiNezio, P.: Assessment of Equilibrium Climate Sensitivity of the Community Earth System Model Version 2 Through Simulation of the Last Glacial Maximum, Geophysical Research Letters, 48, e2020GL091 220, https://doi.org/https://doi.org/10.1029/2020GL091220, e2020GL091220 2020GL091220, 2021.

Zhu, J., Otto-Bliesner, B. L., Brady, E. C., Gettelman, A., Bacmeister, J. T., Neale, R. B., Poulsen, C. J., Shaw, J. K., McGraw, Z. S., and Kay, J. E.: LGM Paleoclimate Constraints Inform Cloud Parameterizations and Equilibrium Climate Sensitivity in CESM2, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 776, https://doi.org/https://doi.org/10.1029/2021MS002776, e2021MS002776 2021MS002776, 2022.