Parameterising cloud base updraft velocity of marine stratocumuli

Jaakko Ahola[1], Tomi Raatikainen[1], Muzaffer Ege Alper[1], Jukka-Pekka Keskinen[1], Harri Kokkola[2],
Antti Kukkurainen[2], Antti Lipponen[2], Jia Liu[3], Kalle Nordling[1], Antti-Ilari Partanen[1],
Sami Romakkaniemi[2], Petri Räisänen[1], Juha Tonttila[2], and Hannele Korhonen[1]

[1]Finnish Meteorological Institute, Helsinki, Finland
[2]Finnish Meteorological Institute, Kuopio, Finland
[3]Research School of Earth Sciences, College of Science and Research School of Finance, Actuarial Studies and Statistics,
Australian National University, Canberra, Australia

Correspondence: Tomi Raatikainen (tomi.raatikainen@fmi.fi)

Abstract. The number of cloud droplets formed at the cloud base depends both on the properties of aerosol particles and the updraft velocity of an air parcel at the cloud base. As the spatial scale of updrafts is too small to be resolved in global atmospheric models, the updraft velocity is commonly parameterised based on the available turbulent kinetic energy. Here we present alternative methods through parameterising updraft velocity based on high-resolution large eddy simulation (LES) runs in the case of marine stratocumulus clouds. First we use our simulations to assess the accuracy of a simple linear parametrisation where the updraft velocity depends only on cloud top radiative cooling. In addition, we present two different machine learning methods (Gaussian process emulation and random forest) that account for different boundary layer conditions and cloud properties. We conclude that both machine learning parameterisations reproduce the LES-based updraft velocities at about the same accuracy, while the simple approach employing radiative cooling only produce on average lower coefficient of determination and higher root mean square error values. Finally, we apply these machine learning methods to find the key parameters affecting cloud base updraft velocities.

1 Introduction

Clouds are important for the global climate due to their ability to reflect solar radiation (shortwave radiation) and trap outgoing longwave infrared radiation but there are still several unknowns and uncertainties related to their dynamics including aerosol-cloud interactions (e.g., Wood, 2012; Seinfeld et al., 2016; Schneider et al., 2017; Rosenfeld et al., 2019). Cloud formation requires supersaturation with respect to water vapour and aerosol particles that can act as cloud condensation nuclei (CCN). Although several processes can lead to the formation of supersaturation within an air parcel, the most important one is the adiabatic cooling caused by updrafts. Malavelle et al. (2014) state that updraft velocities strongly control the activation of aerosol particles. Together with aerosol properties and concentrations, the strength of updraft determines the number of droplets formed. Cloud droplet number concentration (CDNC) directly impacts both the precipitation formation and the radiative properties of clouds.

The relative importance of aerosol properties and updraft speed on droplet concentration and aerosol indirect effect has been widely studied and found to be dependent on cloud type and aerosol loading (e.g., Lance et al., 2004; McFiggans et al., 2006;

Reutter et al., 2009; Bougiatioti et al., 2020). Sullivan et al. (2016) state that input updraft velocity fluctuations can explain as much as 61 % in droplet number variability in GEOS-5. Kacarab et al. (2020) found that the characteristic vertical velocity (w^*) plays a very important role in driving droplet formation in a rather polluted marine boundary layer (MBL) regime, in which even a small shift in w^* may make a significant difference in droplet concentrations. Regayre et al. (2018) stated that vertical velocity standard deviation perturbations have the greatest influence on cloud droplet concentrations in regions of relatively high aerosol concentrations because in such environments droplet activation is updraft-limited rather than aerosol-limited. On a global scale, Yoshioka et al. (2019) found the uncertainty in the updraft velocity to be the second most important cause for uncertainty in the aerosol radiative forcing.

Globally, shallow marine clouds have significant climate impacts, because they cover major parts of oceans (Bennartz, 2007). At the same time, these clouds are particularly difficult to model, because they require high model resolution to capture the turbulence (Honnert et al., 2020). Such resolution is available for example, by using Large Eddy Simulation (LES). These models solve large turbulent eddies explicitly and smallest eddies are parameterised (e.g., Schneider et al., 2017). LES models are currently widely used and offer a good dynamical core and basis for cloud model development.

The resolution of cloud processes in a global model can be increased through the use of LES or Cloud Resolving Models (CRM) in each GCM column, which is known as superparameterisation, (e.g., Khairoutdinov and Randall, 2001; Khairoutdinov et al., 2005). However, this is a computationally very demanding solution, so parameterisations are commonly used.

As the resolution of global atmospheric models is too coarse to resolve the cloud-scale updrafts, different parameterisations are employed either to estimate the probability density function of updrafts or a single characteristic updraft for the whole cloud. Most global models employ information on resolved turbulent kinetic energy or eddy diffusivity to parameterise the subgrid turbulence, which can be related to updraft velocity (Golaz et al., 2011). Other suggested methods, given in remote sensing studies (Zheng and Rosenfeld, 2015; Zheng et al., 2016), rely on parameterising the updraft based on the boundary layer properties and the forcing causing the turbulence. Zheng et al. (2016) stated that the parameterisation of vertical velocity has long been recognised as a core issue in numerical weather prediction. For example, it has been suggested that updraft depends on the cloud top radiative cooling in case of stratocumulus (Zheng et al., 2016) or cloud base height for cumulus clouds (Zheng and Rosenfeld, 2015). Recently, machine learning approaches have gained attention, because they can be used as parameterisations with higher accuracy than traditional parameterisations and with much smaller computational costs compared to explicit simulations (e.g., Rasp et al., 2018; Silva et al., 2021; Kashinath et al., 2021).

Here, we present three LES-based boundary layer cloud updraft velocity parameterisations that could be used within a global climate model. The parameterisations are based on detailed LES runs. The three methods are based on 1) a Gaussian process emulator, 2) a linear parameterisation based on cloud radiative cooling (Zheng et al., 2016), and 3) machine learning to correct the approximation error of the linear fit obtained with method 2 (Lipponen et al., 2013, 2018). We also examine which parameters are most important for predicting updraft velocities.

2 Methods

The parameterisations are specifically designed for marine clouds seen in ECHAM. This means that each LES run is initialised using an ECHAM cloud state as an input and then the LES predicts the updraft velocity related to this cloud state. The process for creating a LES-based updraft velocity parameterisation for marine boundary layer clouds is shown as a pipeline in Fig. 1. Here the aim is to create a parameterisation that would represent detailed LES runs with low computational cost. The pipeline has three main stages. First, we need source data that describes boundary layer conditions in a large number of marine boundary layer cloud cases (Sect. 2.2.1). Second, a representative subset of this data is selected (Sect. 2.2.2) and the corresponding simulations are run with the LES model (Sect. 2.3). Third, specific LES model outputs are selected from these simulations and corresponding parameterisations are created (Sect. 2.4).

2.1 LES

The parameterisations developed in this study are based on simulations with UCLALES-SALSA (Tonttila et al., 2017; Ahola et al., 2020), which models atmospheric dynamics with a LES and includes a four-stream radiative transfer solver (Fu and Liou, 1993). We used two different cloud microphysical modules with the LES. First, the default UCLALES (Stevens et al., 1999, 2005) includes Seifert & Beheng microphysics (SB) with diagnostic clouds (saturation adjustment for cloud water, constant cloud droplet number concentration as an input) and a two-moment bulk scheme with parameterised autoconversion for rain (Seifert and Beheng, 2006). The second set-up employs SALSA, which is a sectional scheme where aerosol, cloud droplet and raindrop size distributions and chemical composition are described with several size bins (Kokkola et al., 2008; Tonttila et al., 2017). Cloud activation is calculated by solving equations for condensation of water on aerosol particles and then counting the number of droplets reaching the critical droplet size (prognostic supersaturation scheme). Rain formation uses the same autoconversion parameterisation as used in the SB microphysics. This scheme is employed because UCLALES-SALSA cloud and aerosol bins are based on the dry size. Tracking the dry size means that the wet size may become inaccurate when the droplets grow larger. For this reason, we have implemented rain bins that are based on wet size. Both rain and cloud droplets grow by condensation and collision-coalescence. The autoconversion scheme moves the largest cloud droplets to rain bins with a minimum size of 50 μm, where their size-dependent dynamics can be modelled accurately. Due to the high number of prognostics variables, SALSA simulations are about ten times computationally heavier compared to the SB simulations.

Both SALSA and SB simulations are initialised with temperature and humidity profiles, and solar zenith angle is an additional input for daytime simulations. In addition, SALSA requires aerosol composition and size distributions while cloud droplet number concentration is the only cloud input for SB. Remaining model parameters and settings are the same for all simulations. These inputs are obtained from a global climate model as described in the next section.

2.2 Sampling representative initial cloud states

2.2.1 Source data from ECHAM

Source data describes typical boundary layer cloud conditions so that these can be used to initialise LES runs. In principle, source data could be collected from different satellite products or re-analysis data sets, but we decided to use global aerosol-

chemistry-climate model ECHAM-HAMMOZ (echam6.3-ham2.3-moz1.0) outputs because it provides all the needed variables
including details about aerosols (Tegen et al., 2019). With the given resolution ECHAM-HAMMOZ provided a large number
of shallow clouds that could not be clearly labelled as stratocumuli. Because we wanted to develop a parameterisation that
could be used for other than ideal stratocumuli, the selection criteria were relaxed.

The source data was collected from standard model outputs of a one-year ECHAM-HAMMOZ AMIP (Atmospheric Model
Intercomparison Project) type run (Fig. 1 point A). Filtered source data was sampled from open ocean columns that represent
single-layer low clouds (Fig. 1 point 1.). In practise, first, continental or sea ice covered columns were excluded. Next, columns
without a single-layer cloud above the sea surface and below 700 hPa (about 3000 m) level were screened out. Stratocumulus
clouds rarely extend beyond 1500 m, but ECHAM produced such liquid clouds at higher altitudes. The current 3000 m cut-off
limit is based on the computational requirements of the LES model, as with this limit we can still undertake high resolution
simulations with reasonable computational cost. The threshold cloud water content for a cloudy grid cell was 0.01 g kg^{-1},
which is the limit used in UCLALES-SALSA. Liquid (LWP) and ice (IWP) water paths were calculated for the low cloud and
for the whole column. The column was accepted if the column IWP was less than 10 % of the low cloud's LWP and the low
cloud's LWP was more than half of the total cloud water path (LWP+IWP). These ~~conditions~~ 10% and 50% thresholds should
ensure that the selected columns contain mostly liquid single-layer clouds below 700 hPa level ~~(about 3000)~~and eliminate
radiative impacts from other clouds above the stratocumulus deck.

The previously described calculations produced two of the source data variables, namely LWP for the low cloud and plane-
tary boundary layer height (H_{PBL}), which was defined as the difference between sea level and cloud top pressures. These and
the rest of the source data variables are shown in Table 1. The first five meteorological variables in the table describe liquid
water potential temperature profiles and liquid water content for a well-mixed boundary layer. The last variable is the cosine
of the solar zenith angle, which determines the solar radiative flux at the top of atmosphere.

SB microphysics requires cloud droplet number concentration as input. On the other hand, SALSA needs aerosol size
distributions and chemical composition or hygroscopicity. In order to reduce the number of variables, log-normal particle
size modes are described by using effective dry radii (r_{eff}). Hygroscopicity parameter κ (Petters and Kreidenweis, 2007)
in ECHAM-HAMMOZ depends on aerosol chemical composition while the aerosol in SALSA is assumed to be sulphuric
acid for which $\kappa_{H_2SO_4} = 1.008$ is constant. Sulphuric acid was chosen as the only chemical component of aerosols to limit
computational expenses and it resembles aged sea salt aerosols. When critical supersaturation for cloud activation is the same
in ECHAM-HAMMOZ and SALSA, then $\kappa r^3 = \kappa_{H_2SO_4} r_{eff}^3$, where r is the dry radius in ECHAM. The effective dry radius
of the accumulation mode is considered as a variable, because this mode can be important for cloud activation. For Aitken
and coarse modes we fixed the effective dry radii to 0.0105 and 0.4813 μm, respectively, based on their time averages in
the source data. Aitken, accumulation and coarse modes have fixed standard deviations 1.59, 1.59 and 2.0, respectively, in
ECHAM-HAMMOZ and these are used also in the LES simulations.

2.2.2 Sampling the source data

The next part of the pipeline is sampling representative subsets from the filtered source data (Point 2. in Fig. 1). Here, these subsets are called designs (Point C in Fig. 1). These are used to initialise the LES simulations and further on as a part of inputs for the updraft parameterisations after filtering (Table 2).

Creating a design is a crucial part of the parameterisation development (Diggle and Lophaven, 2006). For instance, a Gaussian Process emulator in general is accurate near the points it is trained with but loses accuracy as one tries to predict conditions far away from them (Rasmussen and Williams, 2006). Therefore, determining the distribution of the points (a good design) plays a vital role in the prediction. It is of key importance to identify regions where a higher accuracy is required, either due to their relevance to the application or simply because predictions are more often done there. On the other hand, one should not focus too heavily on any single region because of the diminishing return from each additional nearby point. As the designs are model-based, the aforementioned aspects imply two competing requirements of a design: i) the design should cover the entire domain of interest, and ii) it should focus on sub-regions of interest where a higher accuracy is desired (see also Liu and Vanhatalo (2020)).

In this work, we used a simple stratified sampling method based on binary space partitioning (BSP) trees (e.g., Fuchs et al., 1980; Tóth, 2005). The goal of the algorithm is to create a partitioning that represents the high level distributional properties of the collection (Sect. 2.2.1) and then uniformly sample a point in each partition. By enforcing a single sample per each partition, we ensure a space-covering property while still letting the distribution of the points in the collection guide which areas to focus.

The BSP method was used to generate separate day and night designs for both SB and SALSA microphysics. The designs are separate, because they contain different variables. For example, cosine of the solar zenith angle has an impact on updraft velocity only during daytime. The number of samples taken from the source data for each design depends on the number of variables as well as the computational resources available for the LES simulations. Both day and night SB designs have 500 samples while SALSA designs have 135 and 150 samples in total for night and day, respectively. Ten samples per variable can be considered as the minimum at least for the Gaussian process emulator (Loeppky et al., 2009).

The probability density function (PDF) of each design variable is given in Fig. 2. The numbers show the full range of values in the designs from min to max. The highest cloud droplet and mode number concentration values are excluded from the figures for clarity. The method used to calculate distributions for the design variables is the Kernel Density Estimation (Scott, 1992), which smooths the PDF and was used due to the low number of samples. It should be noted that the spikes of H_{PBL} values in Filtered ECHAM data in Fig. 2e are related to the vertical discretisation of the model. We added noise to the H_{PBL} values to get a smoother, i.e. more realistic, planetary boundary layer height distribution.

2.3 Setting up the LES runs

After creating the four designs (SB and SALSA microphysics, day and night), we generate the LES inputs and run the simulations (point 3. in Fig. 1). Initial temperature and humidity profiles for LES are reconstructed from the design variables (Table 1) while assuming a well-mixed cloud-topped marine boundary layer. With this assumption, boundary layer total water mixing

ratio (q_t) was solved from LWP, and boundary layer height was converted from the pressure difference to meters above sea surface.

Figure 3 shows an example of the initial temperature and humidity profiles that were generated using the five meteorological input variables. Here, inversion layer thickness is based on an assumed liquid water potential temperature gradient of 0.3 K m^{-1}. Above the inversion layer, θ_L increases at a rate of 3 K km^{-1} and q_t decreases linearly so that it would reach zero 2 km above the top of the inversion layer. Note that the hypothetical total water zero point would be at least 1 km above the top of the LES domain.

The remaining design variables are related to cloud microphysics and solar radiation. Daytime simulations have a fixed solar zenith angle, i.e. it is not changing with time. We did not include diurnal cycle as that would disconnect the end state of the simulations from the input values, which would violate the underlying assumptions of the machine-learning-based approach presented in this study. For SB microphysics, the only cloud micro-physical parameter is cloud droplet number concentration. The SALSA simulations are initialised with the specified tri-modal aerosol size distribution in each grid cell (Sect. 2.2.1).

The model also has options related to initialisation, surface interactions, radiative transfer, large scale forcing, and microphysics. For most of these, default values were used in the simulations, but the following parameters were adjusted. Initial horizontal wind profiles were set to 10 m s^{-1} in the East-West direction for all altitudes. Surface sensible and latent heat fluxes were set to zero, which is in line with the idealised initial temperature and humidity profiles. The radiative solver needs the surface skin temperature, which was set to be the same as the air temperature in the first model layer above the surface. Large-scale divergence was set to 1.5×10^{-6}s^{-1}.

In the simulations, horizontal grids cover 10 km in each direction with a 50 m resolution. Vertical grid is case-dependent so that it extends from the sea-surface up to a height that is 1.333 times the planetary boundary layer height (H_{PBL}). The vertical resolution is 10–20 m depending on the boundary layer height (maximum about 3000 m). The adaptive time step of the model was set to be 2 s minimum and the statistics sampling time period was set to 300 s. Sub-grid fluxes are calculated with Smagorinsky model. The advection of momentum is based on fourth-order difference equations with leapfrog method as a time stepping scheme. The scalar advection uses a second-order flux-limited scheme. The time integration is executed with a simple Eulerian forward time stepping method.

Coagulation, cloud-to-rain autoconversion and sedimentation processes were switched off during the first 1.5 h spin-up, which allows enough time for turbulence to develop. The simulation time was limited to 3.5 hours to prevent the model from drifting too far from the state representing the design variables. For the same reason, we nudged the liquid water potential temperature and SALSA aerosol concentrations towards their initial values using a relaxation time $\tau = 3600$ s. This choice for the relaxation parameter is a compromise value which aims to eliminate too large changes in boundary layer height while still allowing radiative cooling to create mixing. Aerosol was nudged only in the first grid layer above sea surface.

Once the simulations were ready (point D in Fig. 1), we collected standard model outputs cloud top radiative cooling values and updraft velocities from the last simulation hour and calculated averages over the domain (point 4. in Fig. 1). These simulation statistics are used in creating the updraft parameterisations (point E in Fig. 1).

Although we show here only the few selected parameters, some additional details about the cloud evolution is provided in the Supplementary material. This analysis includes discussion about the possible decoupling of clouds from the surface, cloud top and base changes, water path development, surface precipitation and how the turbulence is developed.

2.4 Updraft parameterisations

In this study we created three different updraft parameterisations (point 5. in Fig. 1). The first parameterisation is a simple linear fit (LF) based on cloud top radiative cooling (Zheng et al., 2016) (Sect. 2.4.1), the second is the same linear fit with a random forest (LFRF) improvement (Sect. 2.4.2) and the third is a Gaussian Process Emulator (GPE) (Sect 2.4.3). Parameterisations are trained based on LES run outputs, where parameterisation input values can include design variables (=LES run input variables, Table 1) and/or LES run outputs (i.e. cloud radiative cooling). The objective of the parameterisations is to reproduce the simulated domain mean positive updraft velocities at cloud base (LES output) and predict those according to selected input variables. Since training input values include both LES inputs and outputs, it is vital that the simulations do not drift too far away from the initial state.

We used only one non-machine learning method (LF, Zheng et al. (2016)) as it has only one input variable and hence is possibly the simplest parameterisation available. Yet, as we will show below, it still performs quite well. Here, we focus on introducing machine learning based parameterisations for which the LF parameterisation serves as a baseline.

The main statistical measures for parameterisation intercomparisons are coefficient of determination (R^2) and Root Mean Square Error (RMSE). R^2 is a measure that tells how well the regression model fits the data. R^2 equals the proportion of variance in the dependent variable (y-axis values) that can be explained by the independent variable (x-axis value). However, R^2 does not indicate whether there are enough data points to make a solid conclusion. RMSE is frequently used to measure the differences between predicted values (here parameterisations) and the values regarded as a ground truth (LES runs). RMSE is dependent on the scale of the numbers used, i.e. has the same physical unit as in the original data, and is sensitive to outliers.

2.4.1 Linear fit (LF)

The linear fit approach is inspired by the observational study of Zheng et al. (2016) who found a strong correlation between cloud top radiative cooling (CTRC) and updraft velocity at the cloud base (W_b). Based on that, they proposed a parameterisation

$$W_b = -0.44 \times CTRC + 22.30 \pm 13, \tag{1}$$

where W_b and CTRC have units of $\mathrm{cm\ s^{-1}}$ and $\mathrm{W\ m^{-2}}$, respectively. Here, negative CTRC values indicate cooling. This parameterisation is based on both day and night radar observations and radiative transfer model simulations on non-precipitating marine stratocumulus clouds.

Here, we make a linear fit, W_b as a function of CTRC, where data is obtained from the LES runs, and compare it with the parameterisation of Zheng et al. (2016). We did not exclude precipitating cases, because this would have limited the

applicability of the method in a global model. The parameterisation coefficients were derived separately for the SB and SALSA microphysics and for day and night.

2.4.2 Linear Fit improved with Random Forest (LFRF)

The second parameterisation builds on the linear updraft velocity model (Sect. 2.4.1). However, instead of directly predicting the updraft velocity, this parameterisation predicts the *approximation error*, i.e. the difference between the LF and the LES output for given design variable values (Sect. 2.3). This error prediction is then used to correct the predictions of the linear fit. The approach used in this study is similar to the approximation error correction method introduced in Lipponen et al. (2013) and Lipponen et al. (2018). In Lipponen et al. (2018), it was shown that predicting and correcting the approximation error of the output of an approximative model often leads to more accurate results than directly predicting the model output.

In this study, we trained a Random Forest (Breiman, 2001) regression model to correct for the approximation error in the linear updraft velocity model. Random Forest regressor consists of an ensemble of binary regression trees. Random Forests can learn non-linear functions, and they are relatively tolerant against overfitting. As a result, Random Forests have provided highly accurate results in many applications. For more information on Random Forests, and training and evaluation of the models, see Breiman (2001). In this study, the Random Forest regressor model was trained using the Scikit-learn machine learning package (Pedregosa et al., 2011). The Random Forest model consisted of 200 regression trees, otherwise, the default training parameters of the Scikit-learn package were used. Inputs of the approximation error correction model during training include the design variables and cloud top radiative cooling from the LES simulations. The software used to produce the results is available in GitHub (Ahola et al., 2022b).

2.4.3 Gaussian process emulator (GPE)

The third method we used is a Gaussian process emulator (O'Hagan, 1978; O'Hagan, 2006) to directly predict the updraft velocity as a function of design variables (Table 1). Our implementation is based on the Gaussian Process Regressor given in the Scikit-learn machine learning package (Pedregosa et al., 2011) that uses Rasmussen and Williams (2006) as the theoretical foundation. Gaussian process regression models are based on multivariate normal distributions and covariance functions. Given an input data, a Gaussian process regression model trained with a training dataset can predict the probability distribution of the outputs corresponding to the input data. For more information on training and evaluation of the Gaussian process regression models see, for example, Rasmussen and Williams (2006). The software used produce the results is available in GitHub (Ahola et al., 2021a, 2022b).

2.5 Post-processing the LES runs

In this study we focus on the updraft velocity at the cloud base where the majority of the cloud droplets form. From the different definitions of cloud base updraft velocity (Romakkaniemi et al., 2009), we found that the following gives the best agreement

with the Zheng et al. (2016) observations:

$$W_b = \frac{\sum_i^c W_i^2}{\sum_i^c W_i}, \tag{2}$$

where W_i is positive updraft velocity at the lowest cloudy grid cell (liquid water content LWC > 0.01 g kg^{-1}) in a column and c is the number of cloudy columns. Negative W_i values (downdrafts) are set to zero, i.e. ignored. The sum includes 3D LES model outputs from the last simulation hour.

Linear Fit (LF) and Linear Fit improved with Random Forest (LFRF) are based on cloud top radiative cooling (CTRC), which is calculated from the LES outputs as

$$\mathrm{CTRC} = \frac{\sum_i^c (\mathrm{R}_{i,base} - \mathrm{R}_{i,top})}{c}, \tag{3}$$

where $\mathrm{R}_{i,base}$ and $\mathrm{R}_{i,top}$ are the net (shortwave + longwave) radiative fluxes at the cloud base and top, respectively. Because net fluxes are defined to be positive upward, negative CTRC means cooling. The sum includes 3D (SB microphysics) or horizontally averaged 1D (SALSA microphysics) model outputs that are averaged over the last simulation hour. For SALSA microphysics, 3D radiation fields were not saved in order to reduce the high number of outputs. In the SB simulations, differences between CTRC values calculated from the 3D and 1D radiation outputs were found to be negligible.

Training data for the parameterisations includes the design variables and the corresponding LES updraft velocity (Eq. 2) and CTRC (Eq. 3) outputs. Some simulations that diverged significantly from the initial conditions produced outliers, which reduced the accuracy of the parameterisations in representing the rest of the cases. Therefore, in all following results, before creating a parameterisation, we filtered out simulations where cloud fraction was smaller than 0.61 or cloud top rose more than 10 % (see Table 2). ~~These cases were spotted in~~ In an initial analysis conducted when calculating the LF, these cases were found to produce most of the outliers. All filtering parameter values were the last retrievable values from the simulations. The cases where the cloud top rose more than 10 % were mostly related to weak temperature inversions at the cloud top (Sect. 2.2.1). For example, our temperature inversions start from 0.78 K (Fig. 2b) while Feingold et al. (2016) excluded values lower than 6 K. Weak temperature inversions are less effective in reducing entrainment mixing, which causes the deepening of the boundary layer (Wood, 2012). Nudging the model fields towards the initial conditions was used to suppress the issue, but it was not entirely eliminated. Using more restrictive initial conditions would have produced a more idealised stratocumulus sample, but at the same time we would have lost the ability to predict updraft velocities for less ideal cases commonly present in ECHAM simulation. In short, the limits for the filter (Table 2) were chosen so that they eliminate clear outliers from the linear fit. Changing these thresholds would have an impact for the accuracy of the Linear Fit parametrisation simply by changing the number of outliers, but not so for the two machine learning parameterisations that can represent complex dependencies. In other words, the LF is sensitive on the thresholds while the two machine learning methods are less sensitive.

3   Results and discussion

3.1   Parameterisation intercomparison

In the comparisons of the three parameterisations, LES runs are regarded as the ground truth, LF is regarded as a baseline simple parameterisation, LFRF improves the LF with a machine learning method, and GPE is a standalone machine learning method. The presented parameterisations are compared to each other and the LES runs.

The Linear Fit (LF) to simulated updraft and CTRC values overlaps quite well with the parameterisation from the Zheng et al. (2016) study (marked with grey line and shading) especially in daytime simulations (Fig. 4). In nighttime simulations, the cloud radiative cooling explains much less of the variance of updraft velocity than during daytime, and the Linear Fit nighttime R^2 values for SB and SALSA are 0.36 (Fig.4a) and 0.29 (Fig.4c), respectively, compared to 0.67 (Fig. 4b) and 0.56 (Fig. 4d) during daytime. The daytime simulations make a better fit, because they have a wider distribution of radiative cooling values due to the large variability in solar radiation. In the absence of solar radiation, CTRC values in SB microphysics nighttime simulations are clustered around the value of -70 Wm^{-2} and the exceptions are related to high boundary layer heights (increased CTRC and W_b), thin clouds (reduced CTRC and W_b), and rising clouds (reduced W_b). A cooling value of -70 Wm^{-2} is a typical difference in net longwave radiation between cloud top and ~~and~~ cloud base for shallow boundary layer clouds.

Figure 5 shows the evaluation of the three different parameterisations, Linear Fit (LF), Linear Fit improved with Random Forest (LFRF) and Gaussian Process Emulator (GPE), based on a cross-validation approach. To get the predicted points, training data is shuffled randomly and then a 10-fold cross-validation is used. This means that for each fold, 90 % of the training data is used to train the parameterisation and predicted values are retrieved for the remaining 10 % of the points. When comparing parameterisations and simulation sets, R^2 and RMSE values are in line with each other, meaning that as R^2 increases (= better fit), RMSE decreases (=smaller error).

Figure 5 shows that LFRF performs better than the other methods in all simulation sets based on R^2 and RMSE values. GPE is close second in all but the daytime SALSA simulations. Both LFRF and GPE work generally well. The probable reason why the LFRF method performs so well, outperforming even GPE, is that it is supported with an embedded dependency, the linear model LF. This is in line with Lipponen et al. (2013, 2018) where they showed that including a dependency (= correcting a rough empirical model with random forest) improved results compared to a pure random forest learning method. LF is the least accurate method in all simulation sets, except in SALSA day where it outperforms the GPE only slightly. Because LF shows worse results during night than during day (see Fig. 4), there is more room for improvement for the machine learning methods during nighttime.

Parameterisations perform slightly better for simulation sets with SB than the sets with SALSA microphysics. One reason is that accounting for aerosol-cloud interactions increases the variability of model predictions ~~, which~~ as prognostic supersaturation scheme increases degrees of freedom. Hence, the variability can be difficult to capture in a parameterisation based on a relatively small set of training data. This additional variability can be seen as lower R^2 values in Fig. 4. The

other main reason is that the number of computationally heavy SALSA simulations had to be limited to the lowest possible. Limited learning data set will have an impact on the accuracy of the predictions.

3.2 Permutation feature importance

We calculated permutation feature importances to reveal the significance of individual input parameters for the LFRF and GPE updraft velocity parameterisations. The permutation feature importance is defined to be the decrease in a model score ($=R^2$) when a single feature (=input variable) value is randomly shuffled (Breiman, 2001; Pedregosa et al., 2011). This method breaks the relationship between the feature and the target, thus the decrease in the model score shows how much the model depends on the feature. This technique does not depend on the used model and it can be applied many times with different permutations of the feature. A high feature importance value means that the variable is important for the model. Some variables had small negative feature importance values, meaning that the variable actually decreased the model score. These values were set to zero as small negative values can be regarded as numerical artefacts. The legend of Fig. 6 shows the overall order of significance based on the mean of fractional (absolute values divided by the total in each subplot of Fig. 6) feature importances.

It should be noted that, as GPE is a standalone method, those variables that physically contribute to updraft velocity are the most important features also in Fig. 6. On the other hand, when improving LF with the learning method (=LFRF), an important physical aspect (=CTRC) is already incorporated, and variables that contribute to the error of the LF are shown as important features. Hence, GPE is easier to interpret physically while the results for LFRF tell which variables, beyond radiative cooling, impact the updraft velocity in the selected setup. LFRF can also be considered as a piecewise-defined constant function, when within a certain subset of inputs, some features can be more important than outside the subset. As the learning method with LFRF is random, it enables that these kind of subsets can be very small and sharp-edged, meaning that training points outside a specific subset do not affect the outputs of the subset. In contrast, GPE makes predictions based on the whole training point domain. These distinct characteristics of GPE and LFRF indicate that important features for GPE are probably relevant in the whole training domain, while for LFRF, the features may be important only in certain subdomains.

Figure 6 shows that the most important parameters are LWP, $\cos\mu$, θ_L, Δq_t, and N_{Ait}. Of these variables, N_{Ait} is a micro-physical variable and specific only to the SALSA microphysics scheme. LWP, θ_L and Δq_t are meteorological variables and common to all simulation sets. $\cos\mu$ is used only in the daytime simulation sets.

Unravelling the meaning of these variables is simplest to start with $\cos\mu$ as it affects updraft velocity quite straightforwardly. For the GPE parameterisation during daytime, $\cos\mu$ is the most important feature, because it strongly influences cloud top radiative cooling, which indicates that the learning method grasps on significant physical aspects that affect updraft velocity. For LFRF, $\cos\mu$ is not that significant since the linear fit for the updraft velocity ($w_{lin.fit}$) already incorporates solar radiation dependency through cloud top radiative cooling.

LWP is important for both LFRF and GPE. First, radiative cooling requires sufficient LWP, so it is important for GPE. On the other hand, high LWP increases precipitation probability. Precipitation causes deviations from the linear trend of W_b vs. CTRC by stabilising the boundary layer and reducing updraft velocities, so LWP is important for the LFRF as well. Zheng et al. (2016) used a radar to measure vertical velocities, so they had to exclude precipitating cases. Our parameterisations are

345 aimed for GCMs where precipitation cannot be excluded as updraft velocity at the cloud base is solved before calculating the cloud droplet number concentration and precipitation formation.

Boundary layer liquid-water potential temperature (θ_L) and boundary layer height (H_{PBL}) determine the cloud top temperature, so they have a direct impact on radiative cooling rates. This can be seen with GPE. With LFRF radiative cooling is already incorporated with $w_{lin.fit}$, but LF does not work well with high and low θ_L values, hence θ_L is an important feature

350 with LFRF.

A strong humidity jump (Δq_t) means a dry free troposphere, which allows strong radiative cooling. Radiative cooling is an important factor for turbulence, and therefore, Δq_t is an important feature for GPE. LF accounts for the radiative cooling, so Δq_t is not an important feature with LFRF.

Regarding aerosol parameters, CDNC is not shown as an important feature with GPE and SB microphysics. This is likely

355 because the bulk model uses saturation adjustment to calculate cloud water mixing ratios. However, with SALSA, the droplet concentration change affects condensation/evaporation rate, buoyancy and thus updrafts, which consequently is seen as high importances with aerosol parameters. This phenomenon (= high aerosol related feature importances) is better shown with LFRF (Figs. 6e and 6g) as this is most likely a significant characteristic only within a subset of simulations with a low aerosol concentration and GPE does not pick it that well as it is not a common feature. Interestingly with LFRF, the relative role of θ_L

360 seems to be smaller when SALSA is employed instead of SB.

4 Conclusions

In this study we ~~developed~~ present three cloud base updraft velocity parameterisations which are based on detailed cloud simulations that can be used in global atmospheric models. The parameterisations represent the predictions of the large-eddy simulation model UCLALES-SALSA (Tonttila et al., 2017) for a wide range of marine boundary layer clouds described by

365 the global climate model ECHAM. One parameterisation is a linear fit (LF) depending on cloud top radiative cooling only. The LF was first presented in Zheng et al. (2016) and is based on cloud observations. Another is based on the linear fit which is improved with a random forest model (LFRF). The random forest model was trained to predict the error of the linear fit as function of parameters describing marine boundary layer clouds. The third is a stand-alone Gaussian Process Emulator (GPE) for predicting updraft velocities based on the cloud parameters.

370 As can be expected, the simple LF works well for cases where radiative cooling is the main driver for turbulence. The other machine learning techniques perform better, because they account for additional variables such as cloud thickness and inversion strength, which have an additional influence on turbulence via processes like cloud top entrainment mixing and evaporative cooling, and drizzle formation (Wood, 2012). Overall, LFRF performs slightly better than GPE.

When choosing between LFRF and GPE, there are some points that should be considered. GPE is purely based on machine

375 learning without any underlying information about physical processes while LFRF includes the effect of cloud top radiative cooling. Therefore, when extrapolating outside of the range of the training inputs, the GPE prediction is the mean of the training outputs that may not be a good prediction. LFRF, on the other hand, reduces to the radiative cooling described by the linear

fit when outside of the range of training outputs. The downside of the LFRF method is that it requires the value for the cloud top radiative cooling (CTRC) from the host model, which must be taken from the previous model time step. As a result, any uncertainties in the input CTRC will influence the predictions while CTRC is not needed for the GPE. On the other hand, using CTRC as an input provides a possibility to extend LFRF also to cases with thin overlaying cloud layers, whereas extending GPE to such cases would require including these cases in the training set.

We developed updraft parameterisations for two different cloud microphysics modules implemented in UCLALES-SALSA. One microphysics module (SALSA) accounts for aerosol-cloud interactions and the other (SB) uses prescribed values for CDNC. Full prediction of the effect of aerosols on clouds and updraft velocity is complicated by the fact that meteorological parameters such as the inversion strength and the liquid water path dominate over aerosol effects. The increased computational cost of SALSA also limited the size of the training data set, which reduced the accuracy of the predictions. Nevertheless, we were able to show that aerosol number concentration has an impact on updraft velocities. In addition, SALSA could be used to predict properties that are more dependent on aerosol such as rain formation.

To conclude, machine learning techniques are becoming commonly used methods for parameterising complex dependencies (e.g. Rasp et al., 2018). Our work shows that the dependency of updraft velocity on boundary layer state can be predicted with a reasonable accuracy and machine learning methods are able to capture also the effect of variables that are important only in a limited subset of all possible conditions, like is the case with aerosol number concentrations. Our results also show that machine learning is effective when it is supported with an underlying physical dependency, which is in line with previous studies (Lipponen et al., 2013, 2018; Silva et al., 2021; Kashinath et al., 2021). The parameterisations introduced are valid only for marine stratocumulus, but extension of the training set to cover the effects of surface heat fluxes and wind shear would improve the physical foundation of updraft parameterisations, and increase the applicability of the method to continental stratiform boundary layer clouds also (e.g., Matheou and Teixeira, 2019).

Code and data availability. Filter source data (ECHAM): https://github.com/ECLAIRscripts/FilterSourceData (Nordling, 2021).

BSP algorithm: https://github.com/ECLAIRscripts/StateSpaceDesign (Alper and Liu, 2021).

Designs: https://github.com/JaakkoAhola/DESIGN/

SB microphysics during NIGHT time (Alper, 2021d),

SB microphysics during DAY time (Alper, 2021c),

SALSA microphysics during NIGHT time (Alper, 2021b),

SALSA microphysics during DAY time (Alper, 2021a).

Creating inputs for training simulations based on the design: https://github.com/JaakkoAhola/LES-emulator-01prepros (Ahola and Raatikainen, 2021).

LES runs with UCLALES-SALSA, ECLAIR branch: https://github.com/UCLALES-SALSA/UCLALES-SALSA/tree/ECLAIR (Tonttila et al., 2021b).

Postprocessing training simulations: https://github.com/JaakkoAhola/LES-02postpros (Tonttila et al., 2021a).

Configuration files for result and figure scripts https://github.com/JaakkoAhola/LES-emulator-04configFiles/

(Ahola, 2021).

Creating parameterisations: https://github.com/JaakkoAhola/LES-emulator-02postpros (Ahola et al., 2022b).

Gaussian Process emulator script: https://github.com/JaakkoAhola/GPEmulatorPython (Ahola et al., 2021a).

415 Library for python scripts: https://github.com/JaakkoAhola/LES-03plotting (Ahola, 2022a).

Plotting the results: https://github.com/JaakkoAhola/LES-emulator-03plotting (Ahola, 2022b).

Data of LES runs: https://fmi.b2share.csc.fi/records/296483f247b1412ebd27f0b82dd1bb76 (Ahola et al., 2022a).

Results available at https://fmi.b2share.csc.fi/records/477af35be02f4a158e2f7e852022ec60 (Ahola et al., 2021b).

Author contributions. KN provided the ECHAM data with the help from HarK, AIP and PR. The design of LES runs was made by MEA. JA, TR, JT, HarK and SR have contributed in developing the UCLALES-SALSA model and JA run the simulations. JA made the LF results with help from TR. JA made the LFRF results with development help from AL and AK. JA made the GPE results with development help from MEA, AK, JL, JPK and AL. JA plotted and analysed the data with help from TR, SR and AL. JA wrote the paper with TR and SR with comments from all co-authors. TR, HarK, AIP, SR, PR and HanK provided senior scientific guidance for the project. HanK came up with emulation concept and supervised the project.

Competing interests. Authors declare no competing interests.

Acknowledgements. This study has been funded by the European Research Council (Consolidator Grant 646857). This research has been supported by the Academy of Finland (grant nos. 322532 and 283031,) and Horizon 2020 Research and Innovation Programme (grant no. 821205). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

References

Ahola, J.: LES-emulator-04configFiles: v1.0.1, https://doi.org/10.5281/zenodo.5383581, 2021.

Ahola, J.: LES-03plotting: v2.0.1, https://doi.org/10.5281/zenodo.6075748, 2022a.

Ahola, J.: LES-emulator-03plotting: v1.1, https://doi.org/10.5281/zenodo.6075779, 2022b.

Ahola, J. and Raatikainen, T.: LES-emulator-01prepros: v1.0, https://doi.org/10.5281/zenodo.5336989, 2021.

Ahola, J., Korhonen, H., Tonttila, J., Romakkaniemi, S., Kokkola, H., and Raatikainen, T.: Modelling mixed-phase clouds with the large-eddy model UCLALES–SALSA, Atmospheric Chemistry and Physics, 20, 11 639–11 654, https://doi.org/10.5194/acp-20-11639-2020, 2020.

Ahola, J., Kukkurainen, A., Alper, M. E., Liu, J., and Lipponen, A.: GPEmulatorPython: v1.0, https://doi.org/10.5281/zenodo.5347718, 2021a.

Ahola, J., Raatikainen, T., Alper, M. E., Keskinen, J.-P., Kokkola, H., Kukkurainen, A., Lipponen, A., Liu, J., Nordling, K., Partanen, A.-I., Romakkaniemi, S., Räisänen, P., Tonttila, J., and Korhonen, H.: Updraft velocity parameterisation data and figures of "Parameterising cloud base updraft velocity of marine stratocumuli" -manuscript, https://doi.org/10.23728/FMI-B2SHARE.477AF35BE02F4A158E2F7E852022EC60, 2021b.

Ahola, J., Raatikainen, T., Alper, M. E., Keskinen, J.-P., Kokkola, H., Nordling, K., Partanen, A.-I., Romakkaniemi, S., Räisänen, P., Tonttila, J., and Korhonen, H.: LES simulations of "Parameterising cloud base updraft velocity of marine stratocumuli" -manuscript, https://doi.org/10.23728/FMI-B2SHARE.179721B8F65643718FF4A5FECF230F7C, 2022a.

Ahola, J., Raatikainen, T., Kukkurainen, A., Alper, M. E., Liu, J., Keskinen, J.-P., and Lipponen, A.: LES-emulator-02postpros: v2.1, https://doi.org/10.5281/zenodo.6075756, 2022b.

Alper, M. E.: DESIGN: SALSA daytime 150 simulations, https://doi.org/10.5281/zenodo.5346794, 2021a.

Alper, M. E.: DESIGN: SALSA nighttime 135 simulations, https://doi.org/10.5281/zenodo.5346789, 2021b.

Alper, M. E.: DESIGN: SB 500 daytime simulations, https://doi.org/10.5281/zenodo.5346768, 2021c.

Alper, M. E.: DESIGN: SB 500 nighttime simulations, https://doi.org/10.5281/zenodo.5346765, 2021d.

Alper, M. E. and Liu, J.: ECLAIRscripts/StateSpaceDesign, https://doi.org/10.5281/zenodo.5343366, 2021.

Bennartz, R.: Global assessment of marine boundary layer cloud droplet number concentration from satellite, J. Geophys. Res. Atmos., 112, https://doi.org/10.1029/2006JD007547, 2007.

Bougiatioti, A., Nenes, A., Lin, J. J., Brock, C. A., de Gouw, J. A., Liao, J., Middlebrook, A. M., and Welti, A.: Drivers of cloud droplet number variability in the summertime in the southeastern United States, Atmospheric Chemistry and Physics, 20, 12 163–12 176, https://doi.org/10.5194/acp-20-12163-2020, 2020.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Diggle, P. and Lophaven, S.: Bayesian Geostatistical Design, Scandinavian Journal of Statistics, 33, 53–64, https://doi.org/10.1111/j.1467-9469.2005.00469.x, 2006.

Feingold, G., McComiskey, A., Yamaguchi, T., Johnson, J. S., Carslaw, K. S., and Schmidt, K. S.: New approaches to quantifying aerosol influence on the cloud radiative effect, Proceedings of the National Academy of Sciences, 113, 5812–5819, https://doi.org/10.1073/pnas.1514035112, 2016.

Fu, Q. and Liou, K. N.: Parameterization of the Radiative Properties of Cirrus Clouds, J. Atmospheric Sci., 50, 2008–2025, https://doi.org/10.1175/1520-0469(1993)050<2008:POTRPO>2.0.CO;2, 1993.

465 Fuchs, H., Kedem, Z. M., and Naylor, B. F.: On Visible Surface Generation by a Priori Tree Structures, in: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, vol. 14 of *SIGGRAPH '80*, p. 124–133, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/965105.807481, 1980.

Golaz, J.-C., Salzmann, M., Donner, L. J., Horowitz, L. W., Ming, Y., and Zhao, M.: Sensitivity of the Aerosol Indirect Effect to Sub-grid Variability in the Cloud Parameterization of the GFDL Atmosphere General Circulation Model AM3, J. Clim., 24, 3145–3160,
470 https://doi.org/10.1175/2010JCLI3945.1, 2011.

Honnert, R., Efstathiou, G. A., Beare, R. J., Ito, J., Lock, A., Neggers, R., Plant, R. S., Shin, H. H., Tomassini, L., and Zhou, B.: The Atmospheric Boundary Layer and the "Gray Zone" of Turbulence: A Critical Review, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD030317, 2020.

Kacarab, M., Thornhill, K. L., Dobracki, A., Howell, S. G., O'Brien, J. R., Freitag, S., Poellot, M. R., Wood, R., Zuidema, P., Redemann, J.,
475 and Nenes, A.: Biomass burning aerosol as a modulator of the droplet number in the southeast Atlantic region, Atmospheric Chemistry and Physics, 20, 3029–3040, https://doi.org/10.5194/acp-20-3029-2020, 2020.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal
480 Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 093, https://doi.org/10.1098/rsta.2020.0093, 2021.

Khairoutdinov, M., Randall, D., and DeMott, C.: Simulations of the Atmospheric General Circulation Using a Cloud-Resolving Model as a Superparameterization of Physical Processes, Journal of the Atmospheric Sciences, 62, 2136 – 2154, https://doi.org/10.1175/JAS3453.1, 2005.

Khairoutdinov, M. F. and Randall, D. A.: A cloud resolving model as a cloud parameterization in the NCAR Community Climate System
485 Model: Preliminary results, Geophysical Research Letters, 28, 3617–3620, https://doi.org/10.1029/2001GL013552, 2001.

Kokkola, H., Korhonen, H., Lehtinen, K. E. J., Makkonen, R., Asmi, A., Järvenoja, S., Anttila, T., Partanen, A.-I. I., Kulmala, M., Järvinen, H., Laaksonen, A., and Kerminen, V.-M. M.: SALSA - a sectional aerosol module for large scale applications, Atmos. Chem. Phys., 8, 2469–2483, https://doi.org/10.5194/acp-8-2469-2008, 2008.

Lance, S., Nenes, A., and Rissman, T. A.: Chemical and dynamical effects on cloud droplet number: Implications for estimates of the aerosol
490 indirect effect, J. Geophys. Res. Atmos., 109, https://doi.org/10.1029/2004JD004596, 2004.

Lipponen, A., Kolehmainen, V., Romakkaniemi, S., and Kokkola, H.: Correction of approximation errors with Random Forests applied to modelling of cloud droplet formation, Geoscientific Model Development, 6, 2087–2098, 2013.

Lipponen, A., Huttunen, J. M. J., Romakkaniemi, S., Kokkola, H., and Kolehmainen, V.: Correction of Model Reduction Errors in Simulations, SIAM J. Sci. Comput., 40, B305–B327, https://doi.org/10.1137/15M1052421, 2018.

495 Liu, J. and Vanhatalo, J.: Bayesian model based spatiotemporal survey designs and partially observed log Gaussian Cox process, Spatial Statistics, 35, 100 392, https://doi.org/10.1016/j.spasta.2019.100392, 2020.

Loeppky, J. L., Sacks, J., and Welch, W. J.: Choosing the Sample Size of a Computer Experiment: A Practical Guide, Technometrics, 51, 366–376, https://doi.org/10.1198/tech.2009.08040, 2009.

Malavelle, F. F., Haywood, J. M., Field, P. R., Hill, A. A., Abel, S. J., Lock, A. P., Shipway, B. J., and McBeath, K.: A method to represent
500 subgrid-scale updraft velocity in kilometer-scale models: Implication for aerosol activation, J. Geophys. Res. Atmos., 119, 4149–4173, https://doi.org/10.1002/2013JD021218, 2014.

Matheou, G. and Teixeira, J.: Sensitivity to Physical and Numerical Aspects of Large-Eddy Simulation of Stratocumulus, Mon. Weather Rev., 147, 2621–2639, https://doi.org/10.1175/MWR-D-18-0294.1, 2019.

McFiggans, G., Artaxo, P., Baltensperger, U., Coe, H., Facchini, M. C., Feingold, G., Fuzzi, S., Gysel, M., Laaksonen, A., Lohmann, U., Mentel, T. F., Murphy, D. M., O'Dowd, C. D., Snider, J. R., and Weingartner, E.: The effect of physical and chemical aerosol properties on warm cloud droplet activation, Atmos. Chem. Phys., 6, 2593–2649, https://doi.org/10.5194/acp-6-2593-2006, 2006.

Nordling, K.: ECLAIRscripts/FilterSourceData, https://doi.org/10.5281/zenodo.5343428, 2021.

O'Hagan, A.: Curve Fitting and Optimal Design for Prediction, Journal of the Royal Statistical Society: Series B (Methodological), 40, 1–24, https://doi.org/10.1111/j.2517-6161.1978.tb01643.x, 1978.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering & System Safety, 91, 1290–1300, https://doi.org/10.1016/j.ress.2005.11.025, the Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004), 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Petters, M. D. and Kreidenweis, S. M.: A single parameter representation of hygroscopic growth and cloud condensation nucleus activity, Atmos. Chem. Phys., 7, 1961–1971, https://doi.org/10.5194/acp-7-1961-2007, 2007.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, the MIT Press, 2006.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proc. Natl. Acad. Sci., 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Regayre, L. A., Johnson, J. S., Yoshioka, M., Pringle, K. J., Sexton, D. M. H., Booth, B. B. B., Lee, L. A., Bellouin, N., and Carslaw, K. S.: Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF, Atmospheric Chemistry and Physics, 18, 9975–10 006, https://doi.org/10.5194/acp-18-9975-2018, 2018.

Reutter, P., Su, H., Trentmann, J., Simmel, M., Rose, D., Gunthe, S. S., Wernli, H., Andreae, M. O., and Pöschl, U.: Aerosol- and updraft-limited regimes of cloud droplet formation: influence of particle number, size and hygroscopicity on the activation of cloud condensation nuclei (CCN), Atmospheric Chemistry and Physics, 9, 7067–7080, https://doi.org/10.5194/acp-9-7067-2009, 2009.

Romakkaniemi, S., McFiggans, G., Bower, K. N., Brown, P., Coe, H., and Choularton, T. W.: A comparison between trajectory ensemble and adiabatic parcel modeled cloud properties and evaluation against airborne measurements, J. Geophys. Res., 114, D06 214, https://doi.org/10.1029/2008JD011286, 2009.

Rosenfeld, D., Zhu, Y., Wang, M., Zheng, Y., Goren, T., and Yu, S.: Aerosol-driven droplet concentrations dominate coverage and water of oceanic low-level clouds, Science (80-.)., 363, eaav0566, https://doi.org/10.1126/science.aav0566, 2019.

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., and Siebesma, A. P.: Climate goals and computing the future of clouds, Nat. Clim. Chang., 7, 3–5, https://doi.org/10.1038/nclimate3190, 2017.

Scott, D. W.: Kernel Density Estimators, chap. 6, pp. 125–193, John Wiley & Sons, Ltd, https://doi.org/10.1002/9780470316849.ch6, 1992.

Seifert, A. and Beheng, K. D.: A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description, Meteorology and Atmospheric Physics, 92, 45–66, https://doi.org/10.1007/s00703-005-0112-4, 2006.

Seinfeld, J. H., Bretherton, C., Carslaw, K. S., Coe, H., DeMott, P. J., Dunlea, E. J., Feingold, G., Ghan, S., Guenther, A. B., Kahn, R., Kraucunas, I., Kreidenweis, S. M., Molina, M. J., Nenes, A., Penner, J. E., Prather, K. A., Ramanathan, V., Ramaswamy, V., Rasch, P. J., Ravishankara, A. R., Rosenfeld, D., Stephens, G., and Wood, R.: Improving our fundamental understanding of the role of aerosol-cloud

540   interactions in the climate system., Proceedings of the National Academy of Sciences of the United States of America, 113, 5781–90, https://doi.org/10.1073/pnas.1514043113, 2016.

Silva, S. J., Ma, P.-L., Hardin, J. C., and Rothenberg, D.: Physically regularized machine learning emulators of aerosol activation, Geoscientific Model Development, 14, 3067–3077, https://doi.org/10.5194/gmd-14-3067-2021, 2021.

Stevens, B., Moeng, C.-H., and Sullivan, P. P.: Large-Eddy Simulations of Radiatively Driven Convection: Sensitivities to
545   the Representation of Small Scales, Journal of the Atmospheric Sciences, 56, 3963–3984, https://doi.org/10.1175/1520-0469(1999)056<3963:LESORD>2.0.CO;2, 1999.

Stevens, B., Moeng, C.-H., Ackerman, A. S., Bretherton, C. S., Chlond, A., de Roode, S., Edwards, J., Golaz, J.-C., Jiang, H., Khairoutdinov, M., Kirkpatrick, M. P., Lewellen, D. C., Lock, A., Müller, F., Stevens, D. E., Whelan, E., and Zhu, P.: Evaluation of Large-Eddy Simulations via Observations of Nocturnal Marine Stratocumulus, Monthly Weather Review, 133, 1443–1462,
550   https://doi.org/10.1175/MWR2930.1, 2005.

Sullivan, S. C., Lee, D., Oreopoulos, L., and Nenes, A.: Role of updraft velocity in temporal variability of global cloud hydrometeor number, Proceedings of the National Academy of Sciences, 113, 5791–5796, https://doi.org/10.1073/pnas.1514039113, 2016.

Tegen, I., Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Bey, I., Schutgens, N., Stier, P., Watson-Parris, D., Stanelle, T., Schmidt, H., Rast, S., Kokkola, H., Schultz, M., Schroeder, S., Daskalakis, N., Barthel, S., Heinold, B., and Lohmann, U.: The
555   global aerosol–climate model ECHAM6.3–HAM2.3 – Part 1: Aerosol evaluation, Geoscientific Model Development, 12, 1643–1677, https://doi.org/10.5194/gmd-12-1643-2019, 2019.

Tonttila, J., Maalick, Z., Raatikainen, T., Kokkola, H., Kühn, T., and Romakkaniemi, S.: UCLALES–SALSA v1.0: a large-eddy model with interactive sectional microphysics for aerosol, clouds and precipitation, Geosci. Model Dev., 10, 169–188, https://doi.org/10.5194/gmd-10-169-2017, 2017.

560   Tonttila, J., Ahola, J., and Raatikainen, T.: LES-02postpros, https://doi.org/10.5281/zenodo.5347269, 2021a.

Tonttila, J., Raatikainen, T., Ahola, J., Kokkola, H., Ruuskanen, A., and Romakkaniemi, S.: UCLALES-SALSA/UCLALES-SALSA: Ahola et al. 2021, https://doi.org/10.5281/zenodo.5289397, 2021b.

Tóth, C. D.: Binary space partitions: recent developments, in: Combinatorial and Computational Geometry, edited by Goodman, J. E., P. J. and Welzl, E., vol. 52 of MSRI Publications, chap. 29, pp. 529–556, Cambridge University Press, Cambridge, 2005.

565   Wood, R.: Stratocumulus Clouds, Mon. Weather Rev., 140, 2373–2423, https://doi.org/10.1175/MWR-D-11-00121.1, 2012.

Yoshioka, M., Regayre, L. A., Pringle, K. J., Johnson, J. S., Mann, G. W., Partridge, D. G., Sexton, D. M. H., Lister, G. M. S., Schutgens, N., Stier, P., Kipling, Z., Bellouin, N., Browse, J., Booth, B. B. B., Johnson, C. E., Johnson, B., Mollard, J. D. P., Lee, L., and Carslaw, K. S.: Ensembles of Global Climate Model Variants Designed for the Quantification and Constraint of Uncertainty in Aerosols and Their Radiative Forcing, Journal of Advances in Modeling Earth Systems, 11, 3728–3754, https://doi.org/10.1029/2019MS001628, 2019.

570   Zheng, Y. and Rosenfeld, D.: Linear relation between convective cloud base height and updrafts and application to satellite retrievals, Geophys. Res. Lett., 42, 6485–6491, https://doi.org/10.1002/2015GL064809, 2015.

Zheng, Y., Rosenfeld, D., and Li, Z.: Quantifying cloud base updraft speeds of marine stratocumulus from cloud top radiative cooling, Geophys. Res. Lett., 43, 11,407–11,413, https://doi.org/10.1002/2016GL071185, 2016.
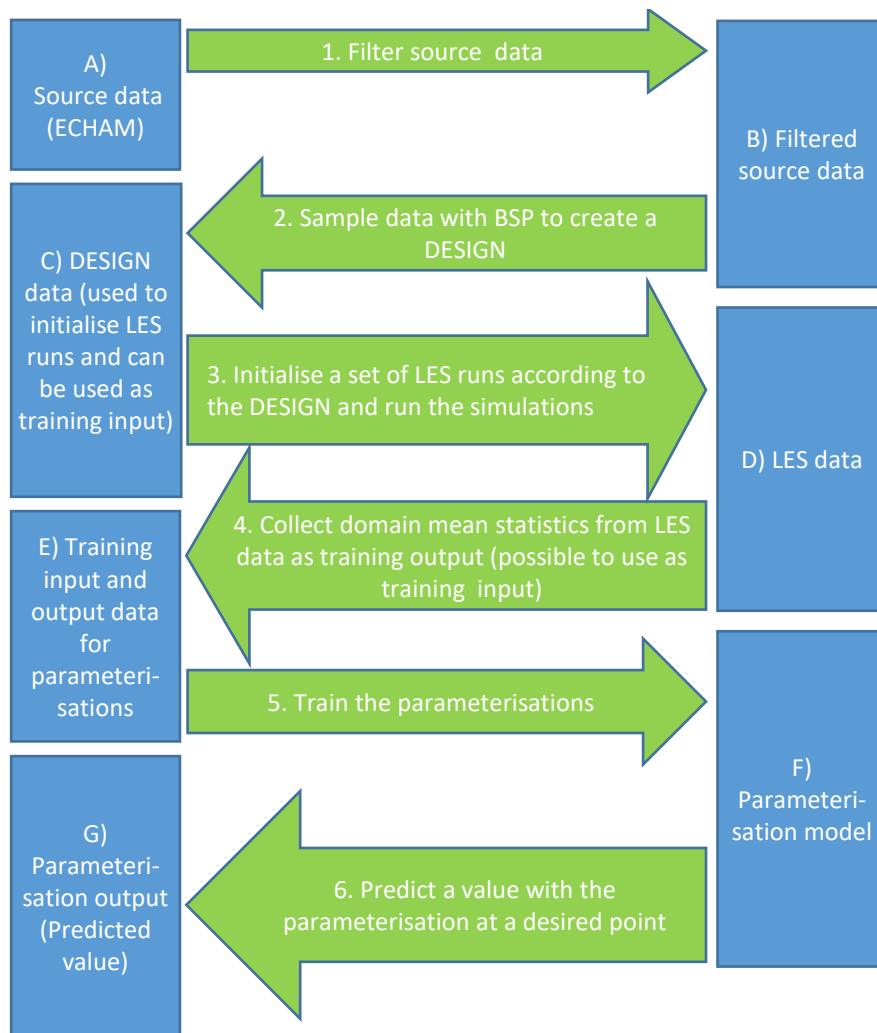
Figure 1. Pipeline of parameterisation development.
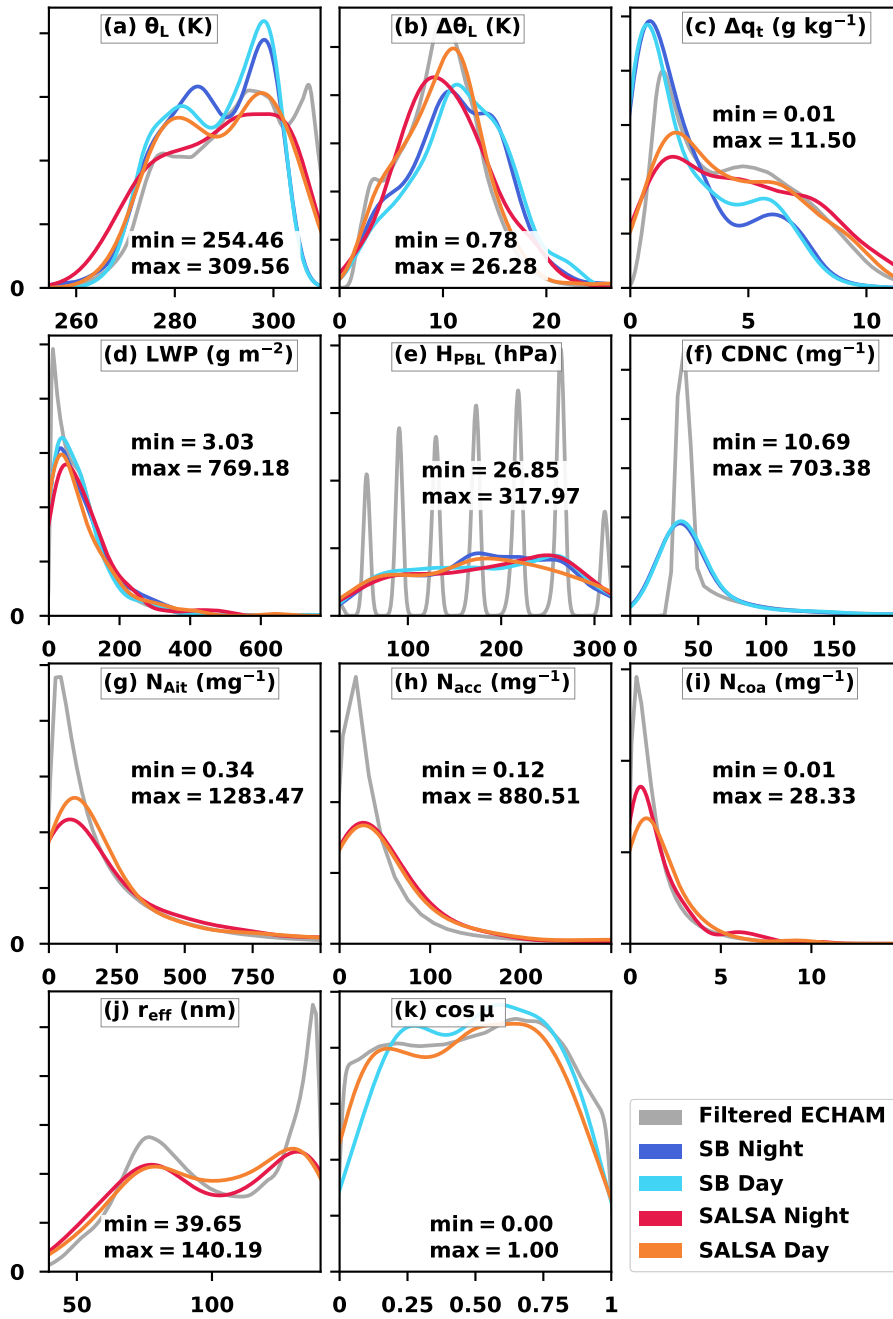
Figure 2. Filtered ECHAM data and Kernel Density Estimated (KDE) LES input variable probability density functions (PDFs). The highest values in panels f,g,h,i are not shown as the density is close to zero. Min and max values are extreme values of all LES simulations sets (i.e. not ECHAM values).
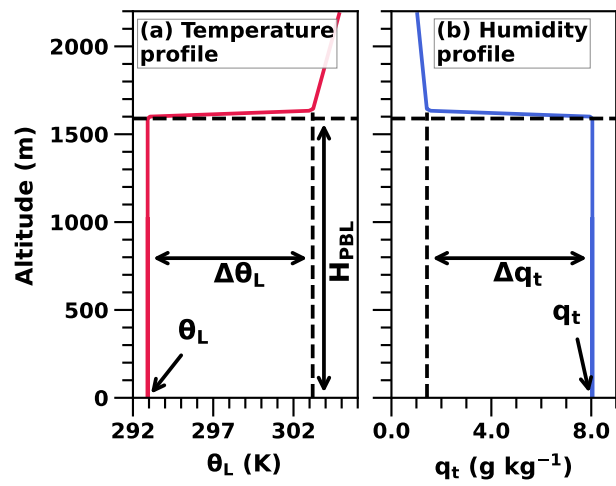
Figure 3. An example of the initial atmospheric profiles generated from the meteorological input parameters.
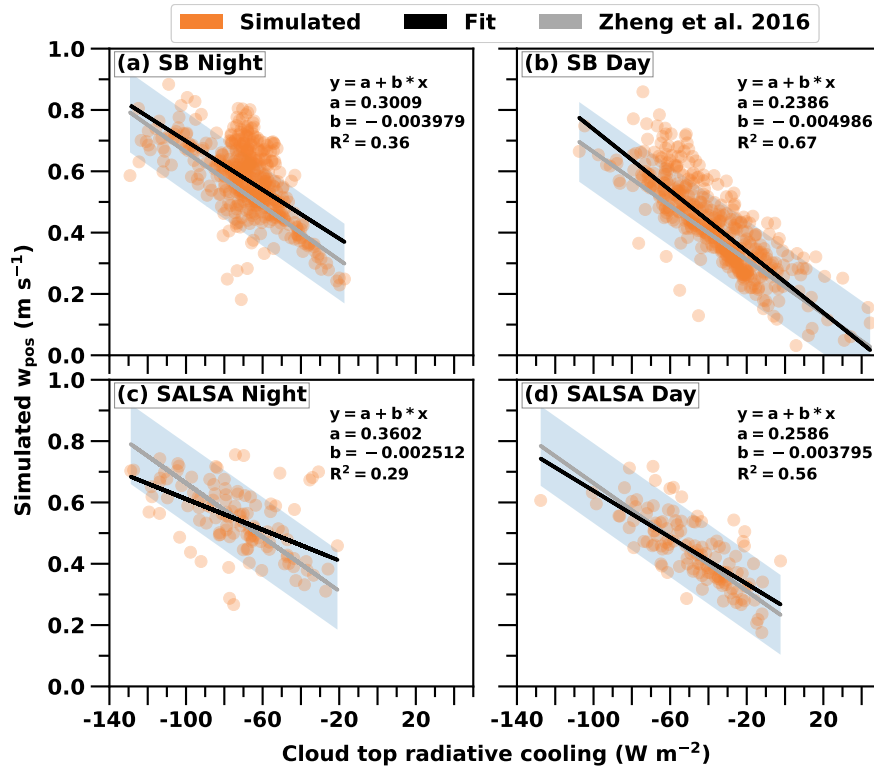
Figure 4. Simulated updraft velocities as a function of simulated cloud top radiative cooling values. Also shown are the linear fits to the data. R^2 is coefficient of determination. SB and SALSA refer to Seifert-Beheng bulk microphysics and sectional microphysics, respectively. In daytime simulations, the solar zenith angle was used as one of the input parameters for the simulations. The gray lines and shadings represent the Zheng et al. (2016) parameterisation with its uncertainty range.
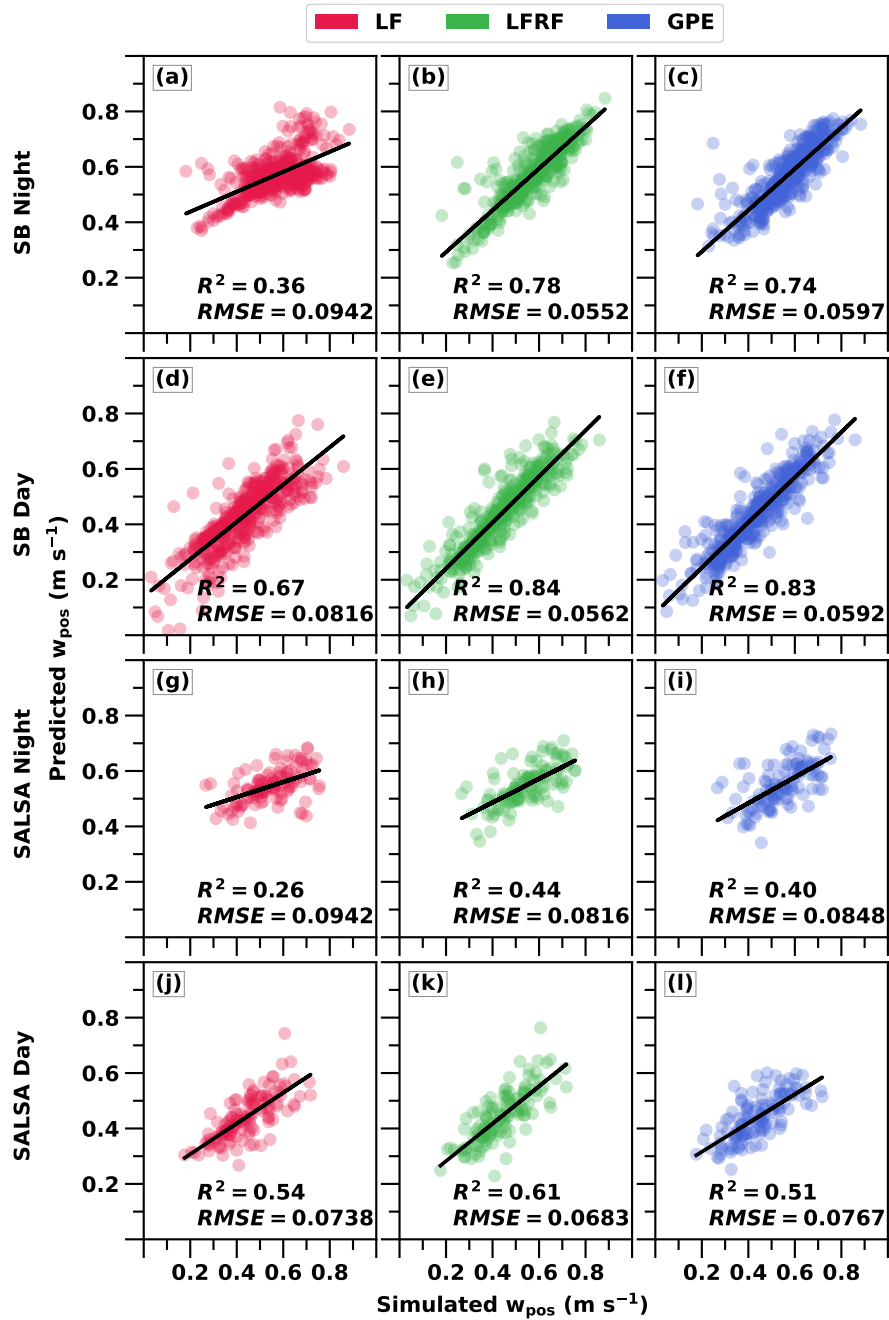
Figure 5. Predicted updraft velocities from different parameterisations vs. simulated updraft velocities. 1st column: Linear fit (LF), 2nd column: Linear Fit improved with Random Forest (LFRF), 3rd column: Gaussian Process Emulator (GPE). Coefficient of determination (R^2) and Root Mean Square Error (RMSE, unit in ms^{-1}) values are also given.
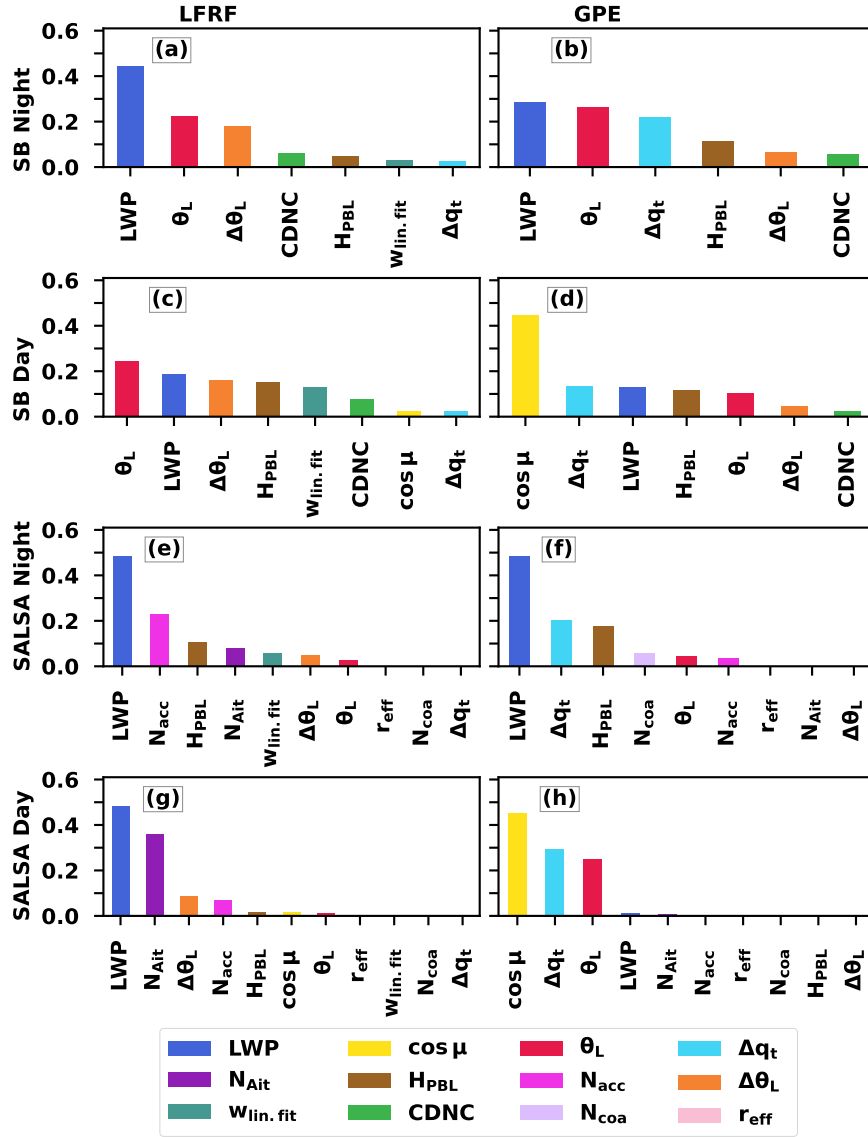
Figure 6. Permutation feature importance of (left) Linear Fit improved with Random Forest (LFRF) and (right) Gaussian Process Emulator (GPE).

Table 1. Variables retrieved from ECHAM.

Variable name	Unit	Variable explanation	Way of retrieving from ECHAM
Δq_t	g kg^{-1}	jump in total water mass mixing ratio at the boundary layer top	difference of max and min values of the total water within two levels from the cloud top
θ_L	K	liquid water potential temperature in the boundary layer	minimum value of potential temperature (the same levels as for Δq_t)
$\Delta\theta_L$	K	inversion strength of liquid water potential temperature	difference of max and min values of potential temperature (the same levels as for Δq_t)
LWP	g m^{-2}	liquid water path for the cloud	integrated from the surface up to the cloud top
H_{PBL}	hPa	planetary boundary layer height described as a pressure difference from the surface	pressure difference from surface up to the cloud top
CDNC	mg^{-1}	cloud droplet number concentration (SB only)	averaged over the cloud
r_{eff}	nm	effective dry radius of accumulation mode (SALSA only)	calculated based on values from the lowest level
N_{Ait}	mg^{-1}	aerosol number concentration in the Aitken mode (SALSA only)	concentration from the lowest level
N_{acc}	mg^{-1}	aerosol number concentration in the accumulation mode (SALSA only)	concentration from the lowest level
N_{coa}	mg^{-1}	aerosol number concentration in the coarse mode (SALSA only)	concentration from the lowest level
$\cos\mu$	-	cosine of solar zenith angle	as is

Table 2. Filtering of the LES simulations.

Dataset	Number of simulations	All filtered cases	Percentage excluded
SB Night	500	472	5.6 %
SB Day	500	461	7.2 %
SALSA Night	135	117	13.3 %
SALSA Day	150	130	12.8%