9-1-2012

Continuous Single-Column Model Evaluation at a Permanent Meteorological Supersite

Roel A.J. Neggers
Royal Netherlands Meteorological Institute, roel.neggers@knmi.nl

A. Pier Siebesma
Royal Netherlands Meteorological Institute

Thijs Heus
Cleveland State University, t.heus@csuohio.edu

CONTINUOUS SINGLE-COLUMN MODEL EVALUATION AT A PERMANENT METEOROLOGICAL SUPERSITE

BY R. A. J. NEGGERS, A. P. SIEBESMA, AND T. HEUS

A facility in the Netherlands brings together simulations and observations, helping scientists improve efficiency and statistical significance of process-level evaluations of numerical weather and climate prediction models.

Uncertainties in numerical predictions of global weather and climate can often be linked to the representation of fast diabatic processes that act on such small scales that they remain unresolved by the general circulation model (GCM). Such processes include turbulence, convection, clouds, and radiative transfer (e.g., Bony and Dufresne 2005). The functional relationships included in a GCM to statistically represent the impact of these subgrid processes on the larger-scale circulation, as a deterministic function of the resolved model state, are often referred to as "parameterizations." The necessity to evaluate and improve these parameterization schemes has motivated intense scientific research in the last few decades, and has in fact created its own active branch within the atmospheric sciences that is dedicated to this purpose. Good examples are international research projects such as the Global Energy and Water Cycle Experiment (GEWEX) Cloud System Study (GCSS; Browning et al. 1993) and various working groups within the Atmospheric System Research (ASR) program of the U.S. Department of Energy (e.g., Stokes and Schwartz 1994; Ackerman and Stokes 2003).

Two research tools have often been applied in the evaluation and development of parameterizations for GCMs. The first is the numerical simulation of turbulence, convection, and clouds in a three-dimensional domain at high resolutions; this technique is known as cloud-resolving modeling (CRM) or large-eddy simulation (LES; e.g., Deardorff 1972; Sommeria 1976). The capacity of CRM and LES to resolve turbulence and convective clouds at high resolutions allows its application as a virtual laboratory, in which small-scale behavior can be studied and understood,

AFFILIATIONS: NEGGERS—Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands; SIEBESMA—Royal Netherlands Meteorological Institute (KNMI), De Bilt, and Delft University of Technology, Delft, Netherlands; HEUS—Max Planck Institut für Meteorologie, Hamburg, Germany
CORRESPONDING AUTHOR: R. A. J. Neggers, Royal Netherlands Meteorological Institute (KNMI), P.O. Box 201, 3730 AE De Bilt, Netherlands
E-mail: roel.neggers@knmi.nl

The abstract for this article can be found in this issue, following the table of contents.
DOI:10.1175/BAMS-D-11-00162.1

In final form 21 November 2011

and against which parameterizations can thus be evaluated. This capacity is still unmatched by meteorological instrumentation. The second research tool is single-column model (SCM) simulation, which stands for the time integration of the standalone code of the suite of subgrid physics in a GCM, using prescribed forcings and boundary conditions (e.g., Tiedtke 1977; Betts and Miller 1986; Randall et al. 1996). A key advantage of the SCM technique is the high model transparency, due to i) the constrained mode of the simulation (i.e., the absence of interaction with the larger-scale circulation) and ii) the easy access (compared to a GCM) to output on all possible model parameters. Combined with the high computational efficiency of SCM simulation, which facilitates sensitivity studies, these benefits act together to increase insight at the process level.

In practice, both methods have typically been applied in combination: first idealized cases are constructed based on observational datasets and simulated with CRM/LES, the results of which then serve as a reference for subsequent SCM simulations. This approach has led to demonstrable improvement of parameterization schemes in operational GCMs. However, with the growing experience with this approach in the research community some shortcomings have been identified. First, idealized cases might not represent actual climate. As a result, parameterizations might get tuned to rare situations. Second, there is no guarantee that such cases, often chosen because they are considered typical for a certain weather regime, also represent those situations that are most troublesome in GCMs. Third, although in recent years a wealth of observational datasets has become available for model evaluation, for various reasons the use of observational data has been disappointingly limited in most SCM and LES case studies (Jakob 2010). Typically, cases have been constructed based on only one or two observational datasets, whereas ideally one would like to simultaneously confront all relevant parameters in a subgrid scheme with their equivalent measurements; only then can one identify compensating errors between parametric components. To summarize, these arguments motivate a move toward a more comprehensive approach in model evaluation, in combination with a more efficient use of available observational datasets.

The recently initiated project described in this paper, named the Royal Netherlands Meteorological Institute (KNMI) Parametrization Testbed (KPT), should be seen as part of a general move toward more statistically significant process-level evaluation. With an emphasis on the representation of atmospheric boundary layer processes, KPT has two main goals that are designed to address the shortcomings of single idealized case studies as mentioned above:

- To reproduce with both the SCM and LES the same statistical level at which a GCM climate is typically evaluated, by generating continuous series of daily simulations that cover long (i.e., multiyear) periods of time, and
- To evaluate the complete parameterized system at multiple time scales against as many independent observational datasets as possible, for example, as available at permanent meteorological sites.

The remainder of this paper is dedicated to motivating these targets and illustrating their potential.

INFRASTRUCTURE. The KPT basically consists of two main components: i) an archive of data streams and ii) an interactive graphical user interface (GUI) for the visualization and intercomparison of the data streams. The various types of data streams include both observational datasets and model output. All data streams are stored at their original resolutions as files in a single, easily accessible data archive. These files have a network common data form (NetCDF) and follow the same unit conventions. The interface resides on a server that is directly coupled to this data archive; its role is to allow quick visualization and intercomparison of all types of data streams, at a range of different time scales. The latter is achieved by means of interactive time averaging during the visualization process, yielding, for example, monthly means, quarterly means, and yearly means. The option to study both long-term composites as well as daily data at its original high resolution is one of the essential aspects of the strategy behind KPT, as will be discussed in more detail in the next section. Plot types include time series, scatterplots, profiles, and contour plots. Observational data quality can be assessed, as well as model performance, by means of simple statistical metrics. Figure 1 shows a snapshot of the interface and an example plot of a monthly-mean evaluation. [A beta version of the KPT interface is accessible on the internet at www.knmi.nl/~neggers/KPT.]

The model data streams currently available in KPT include three types: GCM, SCM, and LES. Some model codes are installed and simulated locally at KNMI, whereas others are simulated at external locations, the results of which are uploaded to the KPT archive through file transfer protocol (ftp). Model simulations can be generated in two modes, either

in automated a priori mode, usually in the form of short-range forecasts, or in manual a posteriori mode, covering periods in the past. Currently participating SCM codes represent various major operational European circulation models. These include the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF; Simmons et al. 1989), the ECHAM5 climate model of the Max Planck Institute for Meteorology in Hamburg (Roeckner et al. 2003), the Hirlam Aladdin Research for Mesoscale Operational NWP in Europe (HARMONIE) mesoscale weather prediction model (http://hirlam.org/), and the Weather Research and Forecasting Model (WRF; Skamarock et al. 2005). The Dutch Atmospheric LES model (DALES; Heus et al. 2010) provides the LES datasets, and can be run on either a central processing unit (CPU) or a graphics processing unit (GPU). The latter option, as recently developed at the Delft University of Technology, significantly enhances the computational speed of the LES, in that it enables a modern standalone computer to obtain the same processor throughput as a single supercomputer node. In practice, this allows the automated daily simulation of weather at Cabauw at high (i.e., cumulus cloud resolving) resolutions at speeds 30 times faster than real time. For the full details of this approach and its illustration, please see Schalkwijk et al. (2012).

Observational data streams from various continuously operational meteorological sites are available in KPT, including most European CloudNet sites (Illingworth et al. 2007) and the Southern Great Plains (SGP) site of the Atmospheric Radiation Measurement Program (ARM) of the U.S. Department of Energy. Currently, the observational data archive is most extensive for the Cabauw Experimental Site for Atmospheric Research (CESAR; see www.cesar-observatory.nl/), the site for which KPT was originally developed. Situated in a flat grassland area in the vicinity of the small village of Cabauw in the Netherlands, the site has been operated by the Royal Netherlands Meteorological Institute since 1973. Its main asset is the 213-m tower (see Fig. 2) equipped at regular intervals with sensors for the purpose of

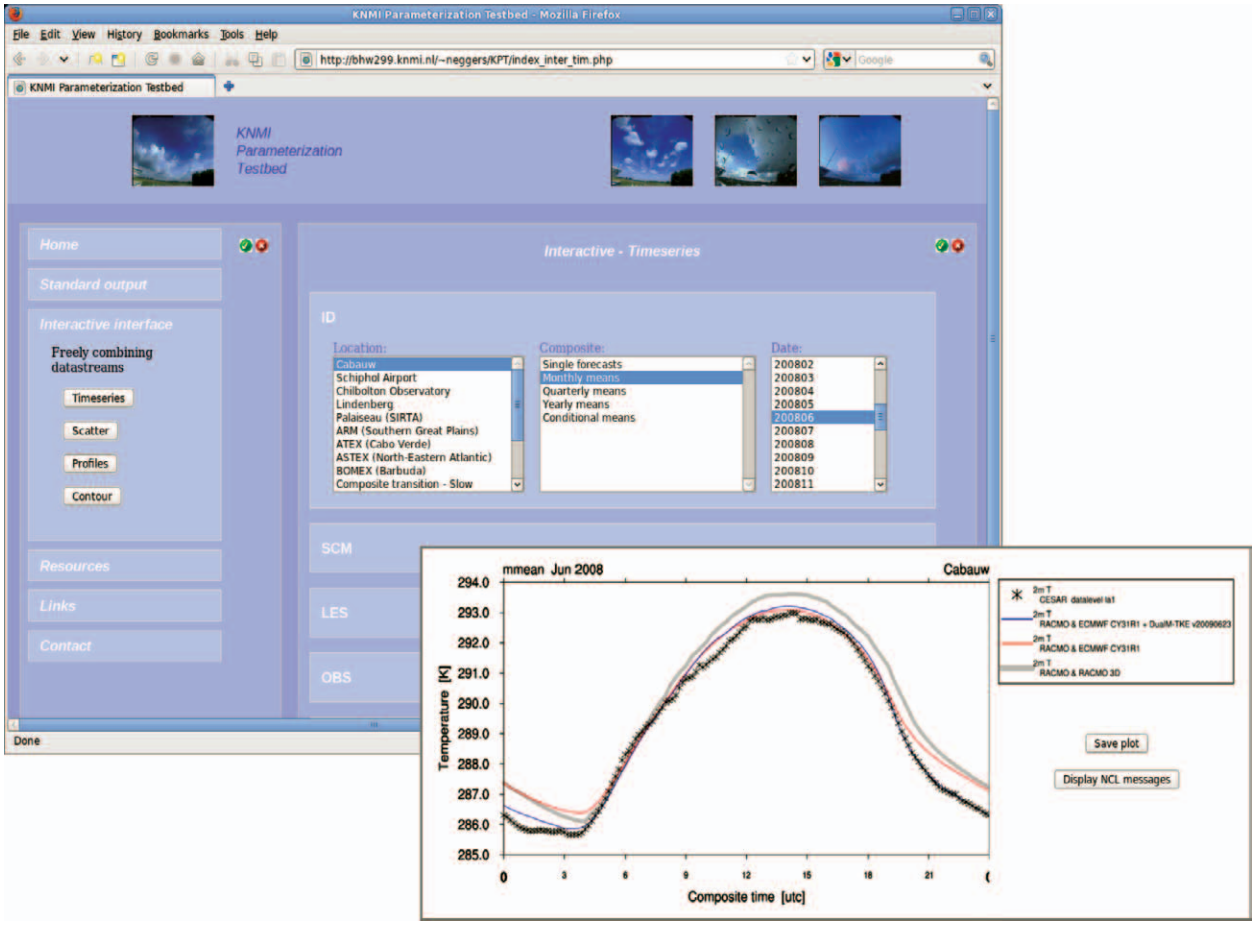

Fig. 1. A snapshot of the KPT interface, including the main selection menu (background) and an example plot (foreground) that evaluates monthly-mean model data (solid lines) against Cabauw measurements (asterisks).

atmospheric boundary layer research, air pollution studies, and climate monitoring (e.g., Driedonks et al. 1978; Van Ulden and Wieringa 1996). In addition, an array of continuously operational instruments is installed at the site, including both in situ and remote sensing equipment [described in detail by Russchenberg et al. (2005)]. The data streams from Cabauw basically come at two data levels, either near-real time or quality checked; both are accessible in the test bed. The Cabauw site participates in the CloudNet project (Illingworth et al. 2007), and as a result all CloudNet products are available in KPT for model evaluation.

STRATEGY. With an infrastructure for the generation, storage, and visualization of all types of data streams in place, we advocate the application of the following evaluation strategy that allows SCM evaluation to become more statistically significant while still maintaining the benefits of single-case studies.

Model hierarchy. A model hierarchy is maintained in KPT to generate the model data streams, as illustrated in Fig. 3. At the top of the hierarchy stands the larger-scale model. These so-called "host models" can provide the large-scale forcings at point locations required to perform the SCM and LES runs. Lower in the model hierarchy stand the SCM and LES models as these are partially, but not completely, "slaved" to the larger-scale flow. As illustrated in Fig. 3b, prescribed advective forcing can be combined with continuous nudging in order to prevent excessive model drift in time. This nudging can be directed toward either the host model state, through relaxation, or an observed state, through assimilation. The tightness of the applied nudging depends on the problem of interest; for example, choosing a synoptic time scale of 6 h is long enough to give fast PBL physics enough freedom to establish their own unique state, but is short enough to make the simulation follow slow large-scale disturbances such as weather fronts. In this setup the LES can be interpreted as a "downscaling" of the host model state at high spatial and temporal resolutions. The LES also serves as a virtual laboratory, providing additional information on 3D variability that the instrumentation at observational sites cannot currently provide. Good examples are the vertical structures of the turbulent variances, covariances, and clouds throughout the boundary layer. The fact that the LES and SCM are forced in exactly the same way ensures that their intercomparison remains meaningful.

Fig. 2. The 213-m tower at the Cabauw site in the Netherlands, with its base partially obscured by morning fog. The 35-GHz cloud radar can be seen on the right. (Figure courtesy of Jacques Warmer.)

Building composites. Following this model hierarchy, the first main target of KPT is the generation of long (multiyear) continuous series of SCM and LES simulations, at integration time steps much shorter than the diurnal time scale (typically less than an hour). These series can consist of many short simulations (covering single days) but also of a smaller number of longer simulations (each covering months or years). Covering long and continuous time periods with both SCM and LES is a relatively recent technique. In the case of SCMs, as already mentioned in the introduction, this is due to its previously preferred application to idealized case studies, lasting a few days at most. In the case of LES, this is due to the significant computational load involved; covering time periods much longer than a few days has only recently become possible due to the significant increase in the computing power of GPUs. Accordingly, the application of GPU-based LES for long-term model evaluation as proposed here is yet unprecedented.

The main purpose of such long time coverage is that it allows calculating long-term averages, or composites. These composites can be simply monthly means, quarterly means, or yearly means but can also be conditional means representing certain weather regimes (e.g., Baas et al. 2010). The model evaluation through long-term composites brings a number of benefits. First, it allows a fair comparison of SCM results to GCM results, at the same long-term statistical level at which the latter are typically evaluated. Second, simulating all individual days in a composite

at subdiurnal integration time steps implies that the composite–internal variability is resolved. This allows selecting those days for detailed "classical" single-case process study, for example, to determine which contribute most to a significant bias in the long-term composite. An attractive aspect of SCM simulation in this respect is its low computational cost, which makes the (re)generation of long-term model composites very time efficient (compared to a GCM). This greatly facilitates sensitivity studies, which in turn can speed up the process of understanding model behavior at the process level, both on short (fast physics) time scales and on long (composite) time scales.

One aspect of the KPT infrastructure that is key to the success of this approach is the capability to interactively calculate and visualize the longer-term composites while still having access to the high-frequency original simulation data. This way, the interface provides the flexibility to the user to choose the time scale of evaluation, depending on the problem of interest.

Multiple independent measurements. The second main target of KPT is to cover as many atmospheric processes and states as possible with high-frequency measurements, similarly covering long continuous periods of time. This approach is motivated by one of the longstanding structural problems in the parameterization of a system of interacting subgrid processes, which is the risk of introducing so-called compensating errors in parameterization schemes. These are situations in which one parameterization erroneously compensates the bias introduced by another, with the net effect that the bias is absent—an undesirable situation, because it is not guaranteed that in a shifting future climate this erroneous cancellation will still occur. By covering as many relevant parameters as possible with independent measurements, assessment of the representation of each individual component in a system of interacting parameterizations is enabled. An example of such an interacting system of fast-acting physics that is relevant for numerical climate prediction is the cloud–radiation–surface interaction; boundary layer clouds are efficient in reflecting the downwelling shortwave radiation, which reduces the surface energy budget, which in turn affects the boundary layer thermodynamic state, which finally affects the low-level clouds again (e.g., Betts et al. 1996). Fully covering this interacting system would require measurement of, among others, i) boundary layer cloud properties, ii) the surface radiative fluxes, iii) the surface energy budget, and iv) the thermodynamic state of the boundary layer. The broad observational coverage of relevant parameters for long continuous periods of time that is required for this approach can currently only be provided by a few permanent atmospheric "supersites" in the world.

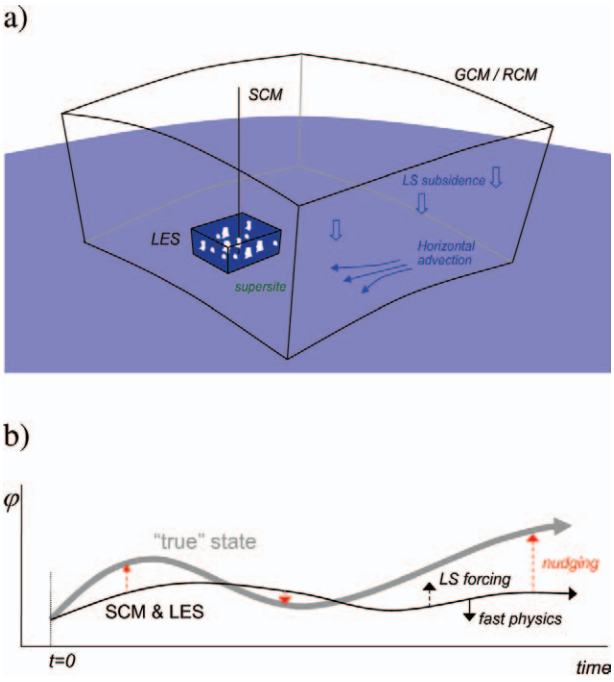

Fig. 3. Schematic illustration of the hierarchy of atmospheric models that is used in KPT. (a) Overview of the various models and domains employed in KPT. (b) Overview of the setup of an SCM or LES simulation in KPT. Various processes acting on a state variable φ are represented by the vertical arrows, such as the prescribed large-scale forcing (dashed black), the continuous nudging (dashed red), and the fast physics (solid black). The directions of the black arrows in this illustration are arbitrary. The "true" or "background" state can be either a GCM state, a purely observed state, or a blending of both.

Guidance by the GCM. The combination of i) the availability of long continuous series of both observational and model data in one framework, ii) the broad range of observed parameters, and iii) the capacity to interactively evaluate composites at a range of different time scales allows the application of the following strategy that lets GCM statistics guide the SCM evaluation. This strategy in principle follows the proposal of Jakob (2003, 2010) but has some essential additions concerning the SCM activity, as schematically illustrated in Fig. 4. Suppose a bias is diagnosed in a long-term mean of a GCM variable relative to observations at a meteorological site. The next step is then to exactly reproduce the same long-term composite

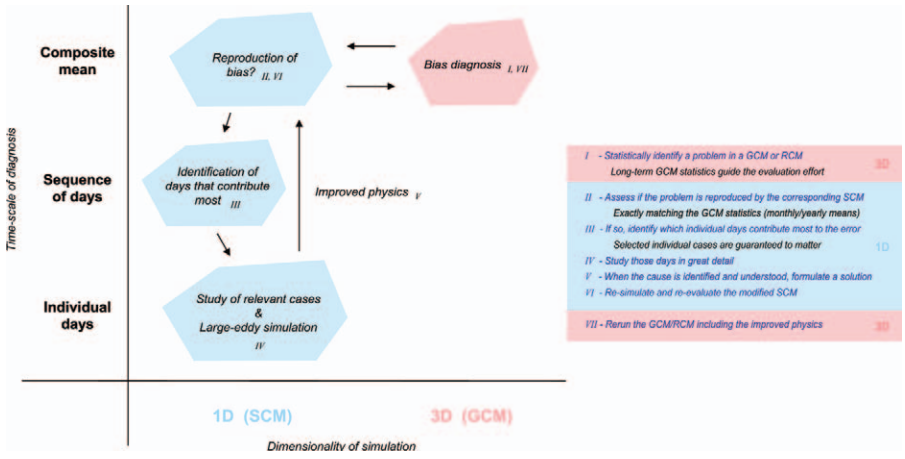

Fig. 4. Schematic illustration of the evaluation strategy followed in KPT. The pink box indicates a GCM activity; the blue boxes represent SCM activities. The subscript Roman numerals indicate the steps as listed in the panel on the right. Further interpretation is provided in the text.

with the SCM. If the same bias is reproduced, then it is possible that the fast physics are the cause, and it makes sense to continue. The subsequent step is then to look more closely at the individual days in the composite, and to identify the day or days that contribute most to the bias in the long-term composite. This step ensures that the cases selected for further study are actually representative of the problem in the GCM; this way, we also preserve the benefits (i.e., model transparency) of single-case studies. Studying these relevant cases in more detail, paying close attention to what exactly happens at the process level and simultaneously evaluating multiple relevant model parameters against measurements and LES, should give better insight into the exact cause of the bias and give inspiration for a solution. If an improvement has been formulated, the improved SCM can be rerun to regenerate the long-term composite. This should reveal if the long-term bias has reduced, and if the modification is generally applicable. If so, the final step is to run the GCM with the improved physics, to establish if the 1D results carry over to the 3D world.

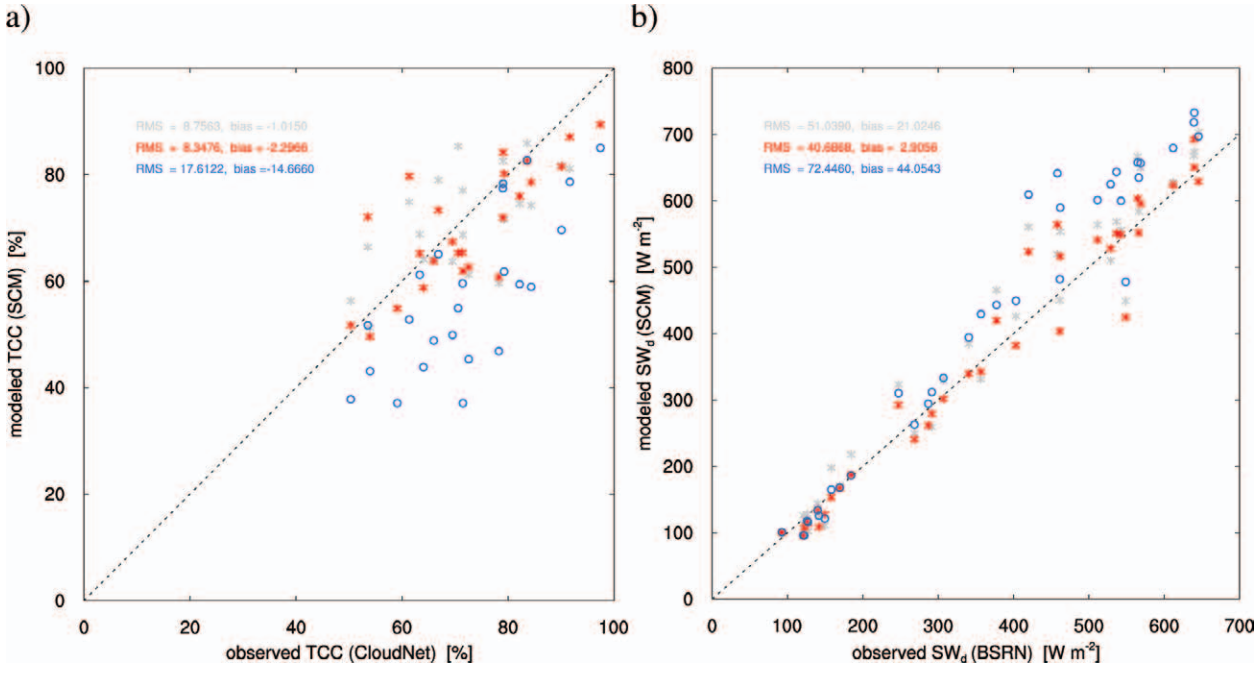

Fig. 5. Scatterplots of monthly-mean Cabauw observations (abscissa) against equivalent model results (ordinate) at 1200 UTC for the period 2007–09. (a) Total cloud cover (TCC), including the CloudNet column Ca product. (b) Downward shortwave radiation at the surface (SW_d), including measurements by the Baseline Surface Radiation Network (BSRN) station. Gray represents the GCM, red represents its SCM, and blue represents its SCM with a different boundary layer scheme. The annotations indicate the root-mean-square error (rmse) and the bias of each model relative to the diagonal.

ILLUSTRATION. We now briefly illustrate the KPT strategy by means of three examples, each demonstrating a different stage in the sequence of steps as outlined in Fig. 4.

Long-term SCM statistics. The first example, Fig. 5, demonstrates the stage of SCM evaluation on long time scales against multiple independent datasets (corresponding to step II in Fig. 4). It concerns the evaluation of the multiyear cloud-radiative model climate against two independent measurements at the Cabauw site. The point of this example is to illustrate how SCM evaluation can be linked to and guided by GCM statistics, and how the availability of multiple independent observational datasets can play a crucial role in this process. Figure 5a shows the TCC, while Fig. 5b shows the SW_d for the 3-yr period 2007–09. Each data point is a combination of a monthly-mean model result (ordinate) and its observed equivalent (abscissa). The model value represents a mean over about 30 daily simulations, the exact number depending on the length of the month. Accordingly, all points together can be interpreted as representing approximately 1,000 individual case studies. The observed values are the total cloud cover of CloudNet and the SW_d as measured by the BSRN, respectively. Three different models are evaluated: a GCM, its own SCM, and its SCM including an experimental version of a new boundary layer scheme.

The results illustrate some important aspects of the test bed approach. First, in this example the SCM more or less reproduces the cloud-radiative climate of its native GCM, which implies that the SCM is representative of GCM behavior and can be used for further study at the process level. Second, the cloud-radiative climate of the SCM with different boundary layer physics differs significantly. Apparently, in this setup, the subgrid physics are free enough to create their own unique state, which is essential for establishing which code does best. It also shows that boundary layer physics can have a large fingerprint on cloud-radiative climate, as was also found by Bony and Dufresne (2005) using GCM data. The biases of the modified SCM against the two independent measures of cloud presence have opposite signs, which is consistent with the known physical impact of the one on the other (i.e., more cloud cover reduces downward shortwave radiation). Such consistency over multiple independent signals can increase confidence in the quality of the evaluation and thus in any conclusion it suggests (in this case, which model has the best cloud-radiative climate).

An attractive way of quantifying the model performance for multiple parameters is the Taylor diagram (Taylor 2001). The idea of these diagrams is to assess how closely a simulated pattern matches the observed pattern, with a pattern being a spatial and/or temporal field. The similarity between two patterns is quantified in terms of their correlation, their variance, and their centered root-mean-square difference. By normalizing the variances with the observed (reference) value, the results for multiple parameters can be plotted in one single figure. Figure 6 is an example for the Cabauw site, in which the models already discussed in Fig. 5 are confronted with eight independent measurements of variables reflecting the heat budget of the coupled boundary layer–soil system; the TCC, the surface downward radiation in the SW_d and longwave LW_d, the soil temperature at 0 cm (T_{soil}), the surface sensible (SHF) and latent (LHF) heat fluxes, and the temperature at 2 m (T_{2m}) and 200 m (T_{200m}). Shown are the monthly-mean values at noontime for the period 2007–10. Although a lot of information can be read from a Taylor diagram, we now focus on the distance to the "REF" point, which represents the situation in which the modeled pattern perfectly matches the observed pattern in terms of correlation and variance. The distance to REF, as indicated by the gray circles, then corresponds to the centered root-mean-square difference between the simulation and the measurement. The red model always has a smaller centered RMS difference, implying that its simulated pattern agrees better with the measurements for all variables.

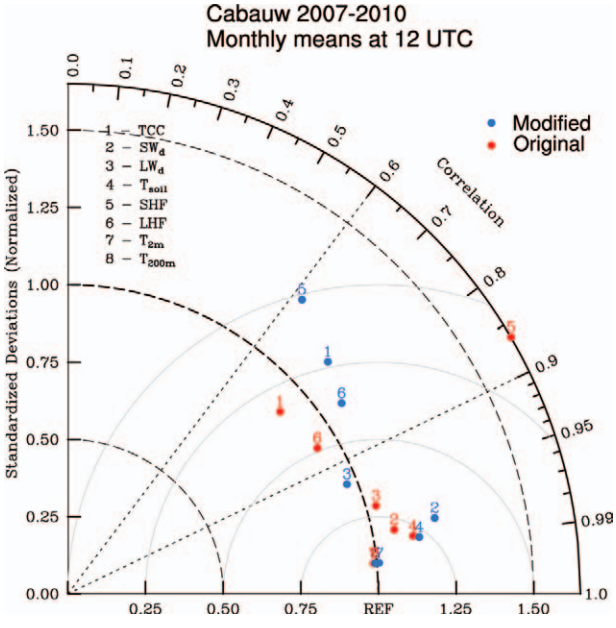

Fig. 6. A Taylor diagram quantifying the model performance at Cabauw for the period 2007–10 for eight parameters. The legend and interpretation are explained in the text.

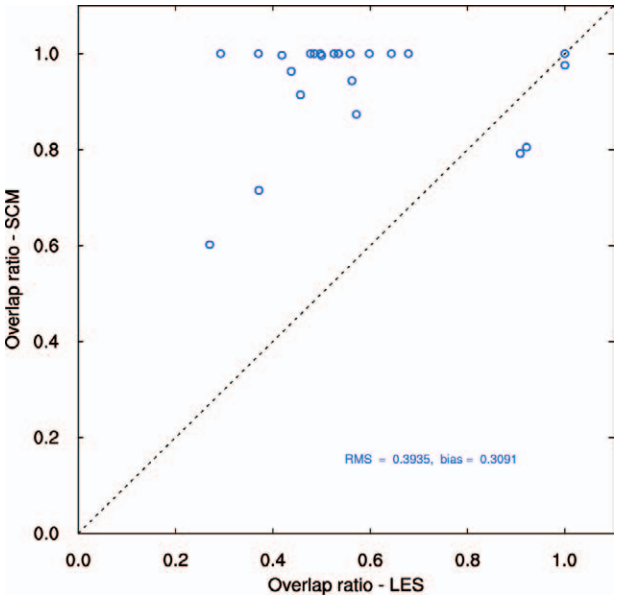

Fig. 7. An evaluation of the cloud overlap in an SCM against LES results for Jun 2008. Plotted is the boundary layer cloud overlap ratio, defined as the ratio of the maximum cloud fraction to the projected cloud cover within the boundary layer. Each point represents the ratio at 1200 UTC on a single day.

Long-term LES statistics. The role of LES in the evaluation of parameterizations is to provide information that is yet hard to measure using present-day instrumentation. Good examples are the three-dimensional structure of a convective cloud field, and the higher moments of statistical distributions describing the turbulent convective variability. A downside of LES can be its significant computational load, which until recently has limited the period of simulation to a few days at most. A key goal of the KPT is to apply LES on a continuous basis and simulate multiyear periods, enabled for the first time by the use of GPUs. Figure 7 is a demonstration of the opportunities brought by long-term LES, showing an evaluation of the vertical cloud overlap in the boundary layer at Cabauw as represented in an SCM against LES results. This SCM is the model already shown in blue in Figs. 5 and 6. The fine horizontal and vertical discretizations applied in the LES mean that it can resolve cumuliform cloud overlap, providing a relevant dataset for the evaluation of parameterizations. In this example the LES model is simulated for the whole month of June 2008. Cloud overlap is here expressed by the ratio of the maximum cloud fraction to the total cloud cover, both diagnosed over the boundary layer. An overlap ratio of 1 implies maximum vertical overlap; a ratio smaller than 1 points to more random (i.e., inefficient) overlap. The figure illustrates that the overlap function in the SCM fails to reproduce the inefficient overlap as diagnosed in LES, which is the probable cause of the underestimation of the monthly-mean total cloud cover as seen in Fig. 5, as well as the related worse performance for the other variables as quantified in Fig. 6. In a related study, inspired by this KPT result,

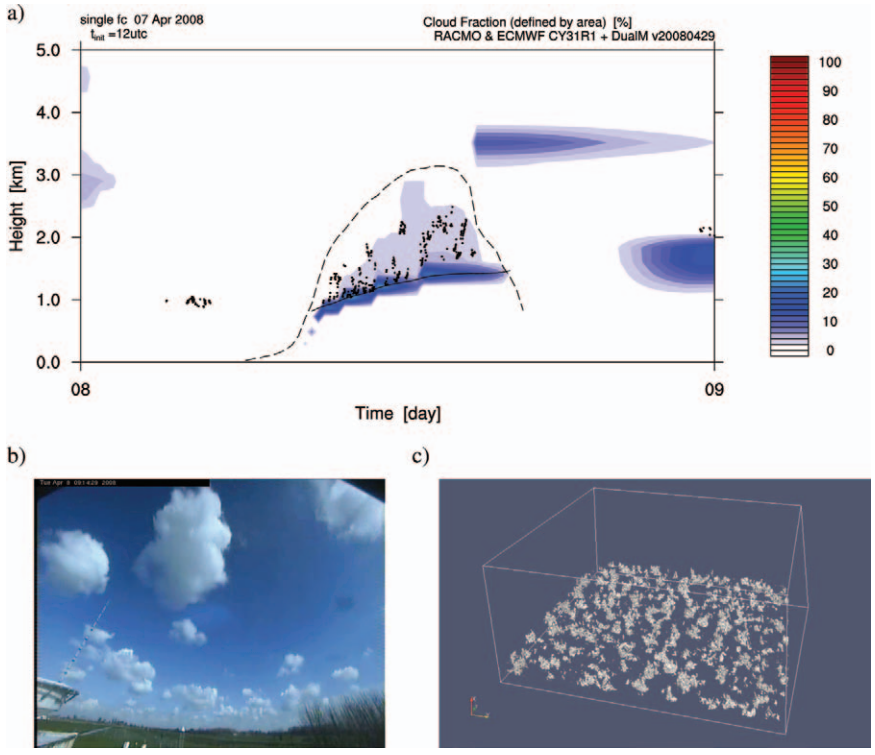

Fig. 8. An example of a single-case process-level study with KPT, showing model output and measurements on 8 Apr 2008 at Cabauw. (a) Time–height contour plot of an SCM's cloud fraction (shaded) overplotted by the lowest cloud-base height as observed by the CT75k ceilometer (black dots). The lifting condensation level (solid line) and the termination height (dashed line) of the strongest model updraft are also shown, for reference. (b) A photo taken by the north-looking Cabauw webcam on 8 Apr 2008. The tower can be seen on the left, with the cloud radar in the foreground. (c) A snapshot of the 3D cloud field as produced by the LES model for this day.

the topic of cumuliform cloud overlap is explored further (Neggers et al. 2011).

Process-level study. The third example demonstrates the stage of SCM evaluation on short time scales that are at or close to the model integration time step (corresponding to step IV in Fig. 4). This stage corresponds to the classical method of single-case process-level study using SCM and LES that has long been practiced by, for example, GCSS working groups, but is now supplemented by a multitude of high-frequency observations.

Figure 8a evaluates the cloud structure and time development at Cabauw on 8 April 2008, featuring a diurnal cycle of shallow cumulus convection, as modeled by the "blue" SCM code as evaluated in the previous figures. Cloud location is evaluated by overplotting the model cloud fraction with high-frequency observations of the lowest cloud-base height by the CT75k ceilometer. What captures the eye is that the time development of the height of the cumulus cloud base is reproduced reasonably well by the boundary layer scheme in the model. Also, the passage of individual cumuli can clearly be distinguished in the high-frequency ceilometer observations. Figure 8b is a snapshot by the Cabauw web camera of the actual cloud field on this day, while Fig. 8c is a snapshot of a virtual cloud field as produced by DALES.

The evaluation of the vertical thermodynamic structure of the boundary layer has always been a key part of model intercomparison studies at the process level, because i) it is mainly established by and therefore reflective of the subgrid transport model in a GCM, and ii) it is strongly linked to the eventual representation of clouds. Figure 9 is an example of such an evaluation, showing the vertical thermodynamic and cloudy structure of the shallow cumulus-topped boundary layer as simulated and observed on 16 June 2008 at Cabauw. The evaluation of a vertical structure requires atmospheric profiling; at Cabauw, both in situ datasets (radiosondes and tower sensors) and remote sensing datasets (profilers, radars, and lidars) are available (note that the radiosonde used in this example is launched at a location about 30 km away from the Cabauw site, which probably explains the offset in the mixed layer humidity). What the figure emphasizes is that relatively small deviations in the vertical thermodynamic structure can be associated with large deviations in cloud state. In this case, the blue model is more successful than the red in reproducing

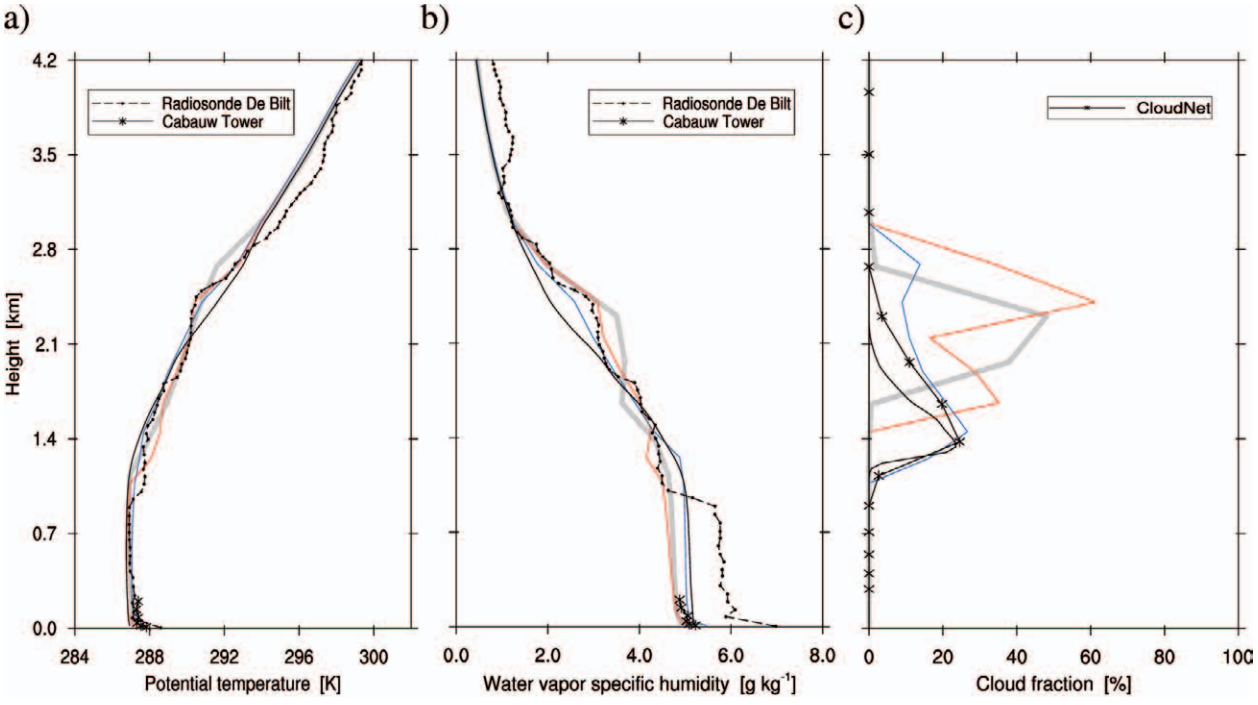

Fig. 9. The vertical structure of thermodynamic and cloudy state of the shallow cumulus-capped boundary layer as observed and simulated at 1200 UTC 16 Jun 2008 at Cabauw. (a) Potential temperature, (b) water vapor specific humidity, and (c) cloud fraction (area averaged). The solid-colored lines refer to the model simulations as shown in Fig. 5, the solid black line represents LES, while the marked black lines represent observational data streams as annotated in the legend.

the observed decreasing cloud fraction with height, a phenomenon that is considered typical of fair-weather cumulus cloud layers. The next step would be to improve the statistical significance of the evaluation result by averaging over many more days with a similar cloud regime. What the figure also illustrates is the benefit of having all types of data streams available in one interface for on-demand plotting and mutual intercomparison; for example, tower measurements can be compared to radiosonde profiles, CloudNet profiles of cloud fraction can be compared to LES and SCM results, etc.

FURTHER DISCUSSION. It is important to consider the role of spatial variability around a meteorological site when comparing models to observations. A first problem can concern representativeness. While a grid box in a numerical model represents a mean over a certain area, a point measurement at a certain location is only a local sample. How can one achieve an honest comparison of model results to such measurements? One way is to make use of area-covering measurements, such as networks of instruments and remote sensing satellite data; another is to focus the evaluation on long time averages, by which time averages become equal to spatial averages (the ergodic principle).

A second problem with spatial variability is its potential impact on local weather. Although Cabauw is a flat land site, the surrounding surface is by no means homogeneous. This has been illustrated by a number of studies [see, e.g., Fig. 1 of Verkaik and Holtslag (2007)]. For example, in the case of spatial heterogeneity in the roughness length of the surface, the behavior of the low-level wind can become dependent on the prevailing wind direction. Another example is the stable nocturnal boundary layer, in which the imposed forcing can reflect local features but also dominate the energy and heat budgets near the surface (Baas et al. 2008). To address this issue the model hierarchy (described in the section "Model hierarchy") is applied in a flexible way, depending on the problem of interest. To this purpose we provide both i) the prescribed forcings and ii) area averages of spatially covering measurements at a range of different scales. This allows the simulation and evaluation of parameterized physics in host models at a range of different horizontal resolutions.

One could include many detailed submodels in the SCM and LES setup (e.g., concerning the representation of the local soil and terrain) for the purpose of making the simulation better reflect local conditions at Cabauw. Although interesting in itself, it is not the intention of the KPT to create the perfect simulation. Instead, the goal is to evaluate with the SCM the subgrid physics exactly as they are in their host model, including all their shortcomings. Otherwise, the SCM might no longer be representative of its host model, which would complicate the attribution of biases as diagnosed in a host model to its subgrid physics.

It should finally be mentioned that the application of continuous nudging in SCM simulations as described above has strong analogies with the so-called initial tendency approach as sometimes applied in three-dimensional forecast models to study the behavior of parameterizations (Rodwell and Palmer 2007). What both approaches share is that they are designed to visualize the fingerprint of fast parameterized physics. An important difference is that in continuously nudged SCM simulations, this fingerprint remains visible throughout the simulation, while in the initial tendency approach it is only visible during the first few time steps. Another argument for applying continuous nudging is that it can reduce the impact of errors in the prescribed forcing, for example, when achieved through assimilation of a locally observed atmospheric state.

SUMMARY AND OUTLOOK. The KPT is designed to be a platform where models and observations come together and can easily be accessed, visualized, and intercompared at a range of different time scales. The primary purpose is to improve the statistical significance and representativeness of process-level evaluation of fast atmospheric physics, with an emphasis on the planetary boundary layer. We propose a new strategy that consists of applying continuous long-term SCM and LES simulation, in combination with comprehensive evaluation against observations at multiple time scales. The examples included in this paper illustrate that it then becomes possible to reproduce typical long-term mean behavior of fast physics in larger-scale models, while still preserving the benefits (e.g., model transparency) of single-case studies. It is argued that this strategy facilitates the tracing and understanding of errors in parameterization schemes, which should eventually lead to a reduction of related uncertainties in numerical predictions of weather and climate.

The extensive use of both model and observational datasets situates the KPT project directly at the interface between two classical communities in the atmospheric sciences, namely, the modeling community and the observational community. The expertise in both communities is essential for making the comprehensive evaluation of models against

observational datasets successful. We therefore hope that this article can convey the opportunities created by an evaluation infrastructure such as KPT for both communities, and that it may encourage future collaborations. Thus, by increasing the efficiency of process-level evaluation studies, we hope to shorten the considerable turnover time that currently still exists between atmospheric observation on the one hand and improvement in numerical weather and climate prediction on the other.

Presently KPT is operational on a permanent basis as a KNMI internal project. However, work is in progress to make KPT accessible to external participants by means of a dedicated server. Detailed information to this purpose is provided on the KPT website (www.knmi.nl/~neggers/KPT). Another ongoing effort is to extend the KPT database to include forcings and measurements at other meteorological sites, such as the CloudNet sites (e.g., Chilbolton, Palaiseau, and Lindenberg) and the ARM sites. The available range of observational data streams is also continuously being extended, with high priority being given to products that have better spatial coverage than point measurements, in order to make time averages equal to spatial averages also on short averaging time scales. Examples are local networks of surface instruments and satellite remote sensing products. A related effort is the application of instrument simulators for both ground-based and satelliteborn instruments.

The KPT project takes part in the ongoing European Union Cloud Intercomparison, Process Study and Evaluation project (EUCLIPSE; www.euclipse.eu/), as well as the Fast-Physics System Testbed and Research project (FASTER; www.bnl.gov/esm/) as funded by Earth System Modeling (ESM) program of the U.S. Department of Energy. Finally, in a forthcoming companion paper the results of a first evaluation study using KPT will be presented, featuring the models used for illustration in this paper.

ACKNOWLEDGMENTS. The KPT project would not have been possible without the continuous and diligent effort by the Cabauw team at KNMI to keep the site operational. Of particular value has been the work by Henk Klein Baltink, Fred Bosveld, and Wouter Knap at KNMI in organizing and maintaining the archive of observational data streams measured at Cabauw. We are much obliged to Erik van Meijgaard at KNMI and Martin Köhler at ECMWF/DWD for generating and providing the advective forcings from RACMO and ERA-Interim, respectively. We greatly appreciate the valuable feedback from various members of the GCSS BLCWG and EUCLIPSE communities during the development stage of the KPT. Wim de Rooij and Cisco de Bruijn at KNMI are thanked for their continuing contributions to the KPT project, and Wayne Angevine at NOAA is acknowledged for providing the WRF model. Our thanks go out to Nils Wedi at ECMWF, whose request for local gliding forecasts inspired the automated daily SCM project that later acted as a prototype for the development of the KPT. We would like to thank two anonymous reviewers for their constructive and knowledgeable comments on this manuscript. The research presented in this paper has received funding from the European Union, Seventh Framework Programme (FP7/2007–2013) under Grant Agreement 244067.

REFERENCES

Ackerman, T. P., and G. M. Stokes, 2003: The Atmospheric Radiation Measurement program. *Phys. Today,* **56,** 38–44.

Baas, P., F. C. Bosveld, G. J. Steeneveld, and A. A. M. Holtslag, 2008: Towards a third intercomparison case for GABLS using Cabauw data. *Extended Abstracts, 18th Symp. on Boundary Layers and Turbulence,* Stockholm, Sweden, Amer. Meteor. Soc., 8A.4. [Available online at http://ams.confex.com/ams/18BLT/techprogram/paper_139844.htm.]

——, ——, G. Lenderink, E. van Meijgaard, and A. A. M. Holtslag, 2010: How to design single-column model experiments for comparison with observed nocturnal low-level jets. *Quart. J. Roy. Meteor. Soc.,* **136,** 671–684, doi:10.1002/qj.592.

Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.,* **112,** 693–709, doi:10.1002/qj.49711247308.

——, J. H. Ball, A. C. M. Beljaars, M. J. Miller, and P. A. Viterbo, 1996: The land surface–atmosphere interaction: A review based on observational and global modeling perspectives. *J. Geophys. Res.,* **101,** 7209–7225.

Bony, S., and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of cloud feedback uncertainties in climate models. *Geophys. Res. Lett.,* **32,** L20806, doi:10.1029/2005GL023 851.

Browning, K. A., and Coauthors, 1993: The GEWEX Cloud System Study (GCSS). *Bull. Amer. Meteor. Soc.,* **74,** 387–399.

Deardorff, J. W., 1972: Numerical investigation of neutral and unstable planetary boundary layers. *J. Atmos. Sci.,* **29,** 91–115.

Driedonks, A. G. M., H. van Dop, and W. Kohsiek, 1978: Meteorological observations on the 213 m mast at

Cabauw, in the Netherlands. Preprints, *Fourth Symp. on Meteorological Observations and Instrumentation,* Denver, CO, Amer. Meteor. Soc., 41–46.

Heus, T., and Coauthors, 2010: Formulation of and numerical studies with the Dutch Atmospheric Large-Eddy Simulation (DALES). *Geosci. Model Dev. Discuss.,* **3,** 99–180.

Illingworth, A. J., and Coauthors, 2007: Cloudnet: Continuous evaluation of cloud profiles in seven operational models using ground-based observations. *Bull. Amer. Meteor. Soc.,* **88,** 883–898.

Jakob, C., 2003: An improved strategy for the evaluation of cloud parameterizations in GCMs. *Bull. Amer. Meteor. Soc.,* **84,** 1387–1401.

——, 2010: Accelerating progress in global atmospheric model development through improved parameterizations: Challenges, opportunities, and strategies. *Bull. Amer. Meteor. Soc.,* **91,** 869–875.

Neggers, R. A. J., T. Heus, and A. P. Siebesma, 2011: Overlap statistics of cumuliform boundary-layer cloud fields in large-eddy simulations. *J. Geophys. Res.,* **116,** D21202, doi:10.1029/2011JD015650.

Randall, D. A., K. Xu, R. J. C. Somerville, and S. Iacobellis, 1996: Single-column models and cloud ensemble models as links between observations and climate models. *J. Climate,* **9,** 1683–1697.

Rodwell, M. J., and T. N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. Roy. Meteor. Soc.,* **133,** 129–146, doi:10.1002/qj.23.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM-5. Part I: Model description. Max-Planck-Institut für Meteorologie Rep. 349, 137 pp. [Available online at www.mpimet.mpg.de/fileadmin/publikationen/Reports/max_scirep_349.pdf.]

Russchenberg, H., and Coauthors, 2000: Ground-based atmospheric remote sensing in the Netherlands: European Outlook. *IEICE Trans. Commun.,* **E88-B,** 2252–2258, doi:10.1093/ietcom/e88-b.6.2252.

Schalkwijk, J., E. J. Griffith, F. H. Post, and H. J. J. Jonker, 2012: High performance simulations of turbulent clouds on a desktop PC: Exploiting the GPU. *Bull. Amer. Meteor. Soc.,* **93,** 307–314.

Simmons, A. J., D. M. Burridge, M. Jarraud, C. Girard, and W. Wergen, 1989: The ECMWF medium-range prediction models: Development of the numerical formulations and the impact of increased resolution. *Meteor. Atmos. Phys.,* **40,** 28–60.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech Note, NCAR/TN-468+STR, 88 pp. [Available online at www.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf.]

Sommeria, G., 1976: Three-dimensional simulation of turbulent processes in an undisturbed trade wind boundary layer. *J. Atmos. Sci.,* **33,** 216–241.

Stokes, G. M., and S. E. Schwartz, 1994: The Atmospheric Radiation Measurement (ARM) program: Programmatic background and design of the cloud and radiation test bed. *Bull. Amer. Meteor. Soc.,* **75,** 1201–1222.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106,** 7183–7192.

Tiedtke, M., 1977: Numerical tests of parameterization schemes for an actual case of transformation of Arctic air. ECMWF Internal Rep. 10, 21 pp. [Available online at www.ecmwf.int/publications/library/ecpublications/_pdf/internal_reports/IR10.pdf.]

Van Ulden, A. P., and J. Wieringa, 1996: Atmospheric boundary-layer research at Cabauw. *Bound.-Layer Meteor.,* **78,** 39–69, doi:10.1007/BF00122486.

Verkaik, J. W., and A. A. M. Holtslag, 2007: Wind profiles, momentum fluxes and roughness lengths at Cabauw revisited. *Bound.-Layer Meteor.,* **122,** 701–719, doi:10.1007/s10546-006-9121-1.