Boundary-layer turbulent processes and mesoscale variability represented by Numerical Weather Prediction models during the BLLAST campaign

Fleur Couvreux[1], Eric Bazile[1], Guylaine Canut[1], Yann Seity[1] , Marie Lothon[2], Fabienne Lohou[2], Françoise Guichard[1], Erik Nilsson[3]

[1]CNRM (Météo-France and CNRS), Toulouse, 31057, France
[2]Laboratoire d'Aérologie, University of Toulouse, CNRS, Toulouse, France
[3]Uppsala University, Uppsala, Sweden

Correspondence to: Fleur Couvreux (fleur.couvreux@meteo.fr)

Abstract. This study evaluates the ability of three operational models, with resolution varying from 2.5 km to 16 km, to predict the boundary-layer turbulent processes and mesoscale variability observed during the Boundary Layer Late-Afternoon and Sunset Turbulence (BLLAST) field campaign. We analyse the representation of the vertical profiles of temperature and humidity and the time evolution of near surface atmospheric variables and the radiative and turbulent fluxes over a total of 12 Intensive Observing Periods (IOPs) each lasting 24h. Special attention is paid to the evolution of the turbulent kinetic energy (*tke*), which was sampled by a combination of independent instruments. For the first time, this variable, a central one in the turbulence scheme used in AROME and ARPEGE, is evaluated with observations.

In general, the 24-h forecasts succeed in reproducing the variability from one day to another in terms of cloud cover, temperature and boundary-layer depth. However, they exhibit some systematic biases, in particular a cold bias within the daytime boundary layer for all models. An overestimation of the sensible heat flux is noted for two points in ARPEGE and is found to be partly related to an inaccurate simplification of surface characteristics. AROME shows a moist bias within the daytime boundary layer, which is consistent with overestimated latent heat fluxes. ECMWF presents a dry bias at 2 m above the surface and also overestimates the sensible heat flux. The high-resolution model AROME resolves the vertical structures better, in particular the strong daytime inversion and the thin evening stable boundary layer. This model is also able to capture some specific observed features, such as the orographically-driven subsidence and a well-defined maximum that arises during the evening of the water vapor mixing ratio in the upper part of the residual layer due to fine scale advection. The model reproduces the order of magnitude of spatial variability observed at mesoscale (a few tens of kilometers). AROME provides a good simulation of the diurnal variability of the turbulent kinetic energy while ARPEGE shows the right order of magnitude.

1 Introduction

Limited area numerical weather prediction models are used routinely for operational weather forecasting across the world. Their increasing resolution is making it important to evaluate their capability to reproduce the low-troposphere vertical profiles of temperature and moisture and their surface turbulent and radiative fluxes as they being increasingly used for numerous applications such as predictions of black ice on roads or agro-meteorology. Here we present the performance, which has remained largely unexplored so far, of these models in representing near-surface variables and boundary-layer turbulent kinetic energy (*tke*).

The evaluation and improvement of models is often a motivation for deploying instruments in field campaigns. However, field campaign observations are less often extensively used to evaluate the representation of surface and boundary-layer processes by operational models. Atlaskin and Vihma (2012) used observations from a field campaign to evaluate NWP models. They focused on the representation of very stable conditions at very low temperatures (<-10°C) in northern Europe

and showed a systematic positive bias for the 2-m temperature, due to an underestimation of the stratification during the coldest nights characterized by very stable conditions. Many studies have used field campaign data to evaluate the behaviour of various non-operational limited-area models. Steeneveld et al. (2008) used data from three particular days of the CASES-99 field campaign to evaluate the impact of the boundary-layer scheme and the radiative scheme on the performance of three different limited-area models. LeMone et al. (2013) used CASES-97 observations to evaluate the boundary-layer schemes and their diagnostics based on mesoscale model simulations. In parallel, models have been evaluated over permanent observing sites such as the ground-based remote sensing observations from the Swiss plateau (Collaud Coen et al., 2014), the Atmospheric Radiation Measurement (ARM, Morcrette 2002 or Guichard et al., 2003) or the Cloudnet sites (Illingworth et al., 2007). In particular, the Cloudnet project has allowed a systematic evaluation of clouds in different operational forecast models. For instance, Bouniol et al. (2010) showed that models tended to overestimate cloud occurrence at all levels.

The Boundary Layer Late Afternoon and Sunset Turbulence (BLLAST) field campaign was conducted from 14 June to 8 July 2011 at Lannemezan in southern France, in an area of complex and heterogeneous terrain. A wide range of instrument platforms including full-sized aircraft, remotely piloted aircraft systems (RPAS), remote sensing instruments, radiosoundings, tethered balloons, surface flux stations, and various meteorological towers were deployed over different types of surface (Lothon et al., 2014). During this campaign, twelve fair-weather days were extensively documented by Intensive Observing Periods (IOPs). These days corresponded mainly to high-pressure fair-weather situations. In this study, we take advantage of the large dataset provided by this campaign to evaluate the vertical structure of the boundary layer and its diurnal evolution as represented in NWP models. Here, we also focus on the mesoscale variability that can occur in the area and how this impacts the observations locally as well as how this is reproduced by the model. Acevedo and Fitzjarrald (2001) used observations complemented by a Large-Eddy Simulation (LES) to show that the spatial variability peaked in the evening transition and that land use and orography played a crucial role in setting temperature anomaly patterns. This highlights the important role of fine resolution in defining the right orography in the model. They also found that, around sunset, horizontal advection played a secondary role compared to vertical divergence.

Several recent studies have also assessed the behaviour of single-column models (a single column of the atmosphere that integrates the same suite of parameterizations as a full 3D simulations) when representing the entire diurnal cycle by comparison to LES. Single-column runs are often used as a simplified configuration of a full 3D simulation in order to highlight some deficiencies in the physics parametrization of the model and to test new developments. By comparing the 1D model to the LES for a case based on observations at Cardington, UK (Beare et al., 2006), which covered the transition from early afternoon to the next morning, Edwards et al. (2006) showed that the 1D model had difficulties in correctly representing turbulence diffusivity during the afternoon transition; this impacted the mean profiles. More recently, Svensson et al. (2011) compared LES and single column models on the entire diurnal cycle of a CASES-99 case and showed a faster decrease of the temperature in the afternoon compared to LES. However, this type of evaluation has not been carried out for operational NWP models and has not used observations of turbulence in the entire boundary layer. For example, observations of *tke* profiles, are quite rare, as they are made only during field campaigns. Therefore the boundary-layer parametrization based on a prognostic equation of the turbulent kinetic energy, which has been shown to perform better than a first-order scheme (Holt and Raman, 1988), has only been evaluated via comparisons with LES results (Cuxart et al., 2006, for instance). Here, we carefully analyse the turbulent kinetic energy, which is a key parameter of the turbulent scheme (Cuxart et al., 2000) used in the two French models evaluated.

Our objectives are i/ to evaluate the skills of operational NWP models in predicting the whole diurnal cycle of the boundary-layer temperature and moisture and in particular the afternoon transition, ii/ to assess the representation of the turbulent kinetic energy by models in which the boundary-layer parametrization is based on a prognostic evolution of the turbulent kinetic energy, iii/ to evaluate the variation of surface thermodynamic parameters for different covers. The

observations and the models evaluated are described in Section 2 together with the methodology used to carry out the comparison. Results are presented in Section 3, focusing on the general representation of the entire diurnal cycle: we provide separate analyses of the reproduction of the energy balance at the surface, the surface meteorological variables and the boundary-layer characteristics, and we end the analysis with a specific focus on the behaviour of the models during the afternoon transition. Discussion and conclusion end the paper.

2 Methodology

2.1 Observations

The observations used in this study were acquired during the BLLAST field campaign and have been described in details by Lothon et al. (2014). Here, they are briefly summarized. They consist of measurements made by remote sensing (Doppler lidar, aerosol lidar, UHF wind profiler) and in-situ (automatic meteorological stations, soundings, remotely piloted aircraft systems, manned aircraft) instruments. They were not used in the assimilation system and could therefore be used for evaluation purposes without ambiguity. Table 1 summarizes all the types of data and measurements used in this study, giving details on the resolution of the raw data, the estimated parameters and their sampling. In the following, we use the observations from the 12 IOPs of the field campaign (Lothon et al., 2014).

In total, 7 different sites were instrumented with eddy covariance systems and radiometers, documenting various types of covers (wheat, grass, forest, moor (an area of open wasteland with grass and heath), corn and more heterogeneous sites). Forest and grassland were the two main land types of the area while moor and urban surface types were intermediate and corn, wheat and bare soil were minority covers (Hartogensis, 2015). A common procedure to retrieve surface heat fluxes from the raw data acquired at 10Hz was applied to all surface stations measuring turbulence and provided surface turbulent and radiative fluxes at 30 minutes' resolution (De Coster and Pietersen, 2012). These observations were used to evaluate the radiative and turbulent fluxes and also the meteorological parameters simulated by the models close to the surface. Their locations are indicated in Fig. 1b by small yellow dots. For these sites, the wind was measured at different altitudes above the ground and was interpolated to 10 m for comparison with the models using a logarithmic profile and the measure of the wind stress close to the surface.

To describe the vertical profile of the boundary layer, we used the data from i/ radiosondes (MODEM, M10 probes) launched four times per day (0000, 0600, 1200 and 1800 UTC - note here that UTC time was the same as solar time as the sites were very close to the Greenwich meridian) from the north-easternmost site (“main site” in the following, indicated by large orange dots in Fig. 1b), ii/ radiosondes (Vaisala RS92 probes) in the lower troposphere (up to 3 to 4 km, Legain et al., 2013) launched hourly from the southern most launching site (4 km from the main site) and iii/ the vertical profiles obtained from the remotely piloted aircraft system (RPAS) SUMO (Reuder et al., 2012) that flew around the main site and provided 4 to 10 soundings of the lower troposphere during the afternoons of the IOPs. These measurements provided vertical profiles of temperature, water vapour content and horizontal wind. Boundary-layer depths were derived from these profiles as detailed in section 2.3. Boundary-layer depths derived from UHF and aerosol lidar data were also used.

The combination of various measurements that provided estimates of the turbulent kinetic energy was a unique aspect of this field campaign. The Doppler lidar (Windcube, manufactured by Leosphere, Gibert et al., 2012), measurements from ground towers, aircraft and the turbulence probe mounted on the tethered balloon (Canut et al., 2016) all contributed estimates of the variance of horizontal and/or vertical wind at high sampling rates (every 4 s for the lidar and 0.1s for the turbulence probe) and thus estimates of the turbulent kinetic energy.

2.2 Numerical weather prediction models

In this study we evaluate the behaviour of three Numerical Weather Prediction (NWP) models:

- two NWP models from Météo-France: (i) a global model, ARPEGE (Courtier and Geleyn, 1988) with a stretched horizontal grid of about 10 km x 10 km over France and a 4Dvar assimilation system and (ii) a limited-area non-hydrostatic model, AROME (Seity et al., 2011), with a grid of 2.5 km x 2.5 km and a 3Dvar data assimilation system;
- the operational ECMWF IFS model with a horizontal grid size of around 16 km x 16 km (Simmons et al., 1989).

Table 2 presents the main characteristics (horizontal resolution, number of vertical levels, boundary-layer scheme, initialization time and forecast period, initialization of the land-surface properties) for the three models.

For this field campaign, the AROME model was run in near-real time over a smaller domain (about a quarter of France) using lateral boundary conditions and initial conditions from the operational AROME, which uses ARPEGE for the lateral boundary conditions. This provided specific outputs for the 16 grid points surrounding the main site (Fig 1b).

All models employed a terrain following hybrid sigma-pressure vertical coordinate. However, the vertical grid differed from one model to another (Table 2): ARPEGE had 70 vertical levels with about 11 levels within the first km (first level at 16 m), AROME had 60 vertical levels with about 15 levels within the first km (first level at 10 m), and ECMWF has 91 vertical levels with about 11 levels within the first km (first level at 10 m). The time step varied from 1 min for the AROME model to about 10 min for ARPEGE and ECMWF. The models also differed by their parametrizations. For the boundary-layer turbulence, AROME uses an Eddy-diffusivity Mass flux concept with the local turbulence (small eddies) represented by a turbulent kinetic energy (*tke)* prognostic scheme (Cuxart et al., 2000) with a non-local length-scale (Bougeault and Lacarrere, 1989) and the boundary-layer thermals and shallow convection represented by a mass-flux scheme (Pergaud et al., 2009). ARPEGE uses the same *tke* prognostic scheme (Cuxart et al., 2000) and uses a mass-flux scheme only when shallow convection is active (Bechtold et al., 2001). ECMWF uses an Eddy-diffusivity Mass flux based on two updraughts (Köhler et al., 2011) and a non-local K profile for the boundary layer while shallow convection is handled by a separate bulk mass-flux scheme (Tiedtke 1989). The surface scheme is ISBA in ARPEGE (Noilhan and Planton, 1989; Giard and Bazile, 2000), AROME uses the surface platform SURFEX (Martin et al., 2014) and ECMWF uses the HTESSEL model (Balsamo et al., 2009). All models have the same longwave radiation scheme, the RRTM parametrization (Mlawer et al., 1997) but differ for the shortwave component: ECMWF uses the SRTM parametrisation while AROME and ARPEGE has the Morcrette at al. (2001) code. The radiation scheme is called every hour for ARPEGE and every 15 min for AROME. Concerning the cloud scheme, ARPEGE uses a distribution of relative humidity based on Smith (1990), AROME a distribution of the saturation deficit based on Bougeault (1982) and ECMWF uses a prognostic scheme (Forbes et al., 2011). In ARPEGE, there are 12 different vegetation covers and one grid point can have only one given vegetation cover while in AROME each grid is associated with a certain fraction of various vegetation types (crops, land, town, mixtures of crops and woodland, Landes forest or broad-leaf forest).

2.3 Comparison methodology

This section gives a detailed description of how the comparison was conducted, focusing on the temporal and spatial resolution of the different variables obtained from models and observations.

Due to the coarse grid spacing of each model, real surface heterogeneities, topography and local circulation are not expected to be reproduced by the models. The real orography and the one present in each model are shown in Figure 2, from which it can be seen that high-resolution (2.5 km) is needed to resolve the north-south valleys of the Pyrenees. Large variability of surface fluxes exists among the sites (Fig 1) at scales smaller than 2.5 x 2.5 km², which corresponds to the size of a grid box in AROME (see for example in Fig 7 of Lothon et al (2014) the differences between the moor and the corn sites, or the grass and the wheat sites, which are a few hundred metres apart). This is mainly due to surface cover as noted by Lothon et al. (2014). However, the variability among observations and the differences between model outputs and observations provide clues as to the main drawbacks of the models. The simulated grid points (and associated columns) surrounding the locations of the measurement sites were extracted and are shown in Figure 1: 3 neighbouring grid points are

extracted for ARPEGE, 16 neighbouring grid points for AROME (a box of 10 km x 10 km including all sites) and 9 neighbouring grid points for ECMWF. Table 3 presents the main physiographic characteristics (altitude, albedo, vegetation fraction and roughness length) of these points.

For ECMWF we evaluated both the analysis available every 6 hours and the operational forecast with 3-hourly outputs for the surface characteristics from the run launched at 0000 UTC while for the two other models we show the forecast launched at 0000 UTC with hourly outputs. The forecast length analysed here was chosen to be 24h. The atmospheric variables corresponded to instantaneous fields sampled every hour for AROME and ARPEGE and every 6 hours for ECMWF. The diagnostics T2m (temperature at 2 m), rh2m (relative humidity at 2 m) and ws10m (horizontal wind speed at 10 m) were obtained using a vertical interpolation following Geleyn (1988) based on the Monin-Obukhov theory between the surface and the first model level for ARPEGE and IFS or calculated using a prognostic surface boundary-layer scheme for AROME (Masson and Seity, 2009).

In the model, the boundary-layer depth is the first level where the *tke* is below 0.01 $m^2\ s^{-2}$. In the observations, various diagnostics allowed the boundary-layer depth to be derived:

i/ the height of maximum air refractive index structure coefficient (Jacoby-Koaly et al., 2002) is obtained from UHF data; it usually provides an estimate of the inversion height based on the vertical gradient of the relative humidity

ii/ the first level below the height diagnosed through i) where the *tke* dissipation rate becomes greater than a threshold (10^{-3} m^2/s^{-3}) is also derived from the UHF data; this criterion gives an estimate of the top of the turbulent layer,

iii/ the height of the largest gradient of aerosol backscatter from the aerosol lidar data (Boyouk et al., 2010); this is another way to estimate the inversion height and

iv/ the best (determined manually) of four criteria applied to the various vertical profiles from soundings and RPAS (Remotely Piloted Airplane Systems) (Lothon et al., 2014), using the height where the virtual potential temperature exceeds the averaged value over the lower levels plus 0.2, or the height of maximum relative humidity, or the height of maximum first derivative of the potential temperature or the height of minimum first derivative of the specific humidity. Often, the criterion based on the virtual potential temperature is chosen. A comparison of different boundary-layer depths derived from various instruments is presented in Bennett et al. (2010).

The decrease of the boundary-layer depth in the afternoon transition is a delicate process and in practice, its estimation is sensitive to the criteria used to derive the boundary-layer depth as already shown by Angevine and Grimsdell (2002) and Bennett et al. (2010). Details of this will be given in Sect. 3.5. The diagnostic used in the model was compared to the criteria iv) applied to the model profiles. These two diagnostics were consistent but in ARPEGE, the model diagnostic tended to overestimate the value derived from the profiles by about 200 m while, in AROME, there was a very good agreement except for 14 June after 1500 UTC and 15 June after 1400 UTC due to the presence of clouds (discussed later). In the following, we will use the model diagnostic discarding these hours of disagreement as it depicts the turbulent layer, in particular during the afternoon transition.

When comparing observations and modelling, we considered the fact that the horizontal and temporal average in observations should be as consistent as possible with the time step and resolution of simulations. In the latter, the surface turbulent and radiative fluxes at a given hour h correspond to the average value between hour h-1 and hour h. In the observations, values were processed every 30min and then averaged to provide the 1-hr average for the comparison. Furthermore, it should be kept in mind that the area (footprint of a few hundred metres) of the surface sampled in the measured surface turbulent fluxes was small relative to the grid size of the three NWPs.

In the observations, the *tke* was estimated for 20 min time windows for the 60-m tower, the Doppler lidar and the tethered balloon; 10 min windows for the 10-m tower (sensitivity to a computation with 20 min windows did not change the results); and for horizontal legs of 25-30 km for the aircraft measurements (corresponding to 5-8 min cf Table 1 and Canut et al., 2016 for more details). This is a compromise between having the same time window as the other measurements and

minimizing the influence of the mesoscale heterogeneities. Note that a 5 km high-pass filter was applied only to the aircraft raw data before the calculation of the *tke* to filter out the mesoscale variability. This is the current treatment used for flux computation, but it induces an underestimation of the *tke* of about 20%. We also tested the *tke* estimates obtained with a 2.5 km high-pass filter but it was affected by a large time-variability, indicating that the samples were not large enough. The estimation of the *tke* with the Doppler lidar (Gilbert et al., 2012) assumed that the turbulence was isotropic and derived the value from the measured vertical velocity variances. To evaluate this hypothesis, we computed the ratio $A = 1.5\frac{\overline{w'^2}}{tke}$, a coefficient from the tower measurements (both from the 60 m tower and the 10 m tower) and from the tethered balloon. A=1 if the turbulence is isotropic, when A>1, the contribution of the vertical velocity variance is dominant (A=3 if the horizontal velocity variances are zero), and when A<1, the contribution of horizontal variance is dominant (A=0 if the vertical velocity variance is zero). Both the tower measurements and the tethered balloon (the tethered balloon never reached heights above 500m) measurements indicated that above 0.1 to 0.2 zi (zi being the boundary-layer height) and in the middle of the boundary layer, this coefficient was between 1 and 2 suggesting that the variance of the vertical velocity was often the main contributor to the *tke* at that height and the *tke* could be estimated from the $\overline{w'^2}$ as $tke = 1.5\,\overline{w'^2}$. Aircraft measurements indicate that closer to the top of the boundary layer this coefficient decreased again taking values between 0.75 and 1. Below 0.1 zi, the variance of horizontal wind was significant and the coefficient A was mostly below 0.6 (see Canut et al., 2016 for more details). Therefore, in the following, we only use Doppler lidar estimates from altitudes above 100 m. More complex computations taking the day-to-day and vertical variation of the anisotropy factor derived from the tethered balloon or aircraft into account could be performed in a future study. Note also that, as we derive the *tke* as 1.5 $\overline{w'^2}$, the observed *tke* tends to be overestimated most of the time but may be underestimated on days with more wind, conditions in which horizontal wind fluctuations are expected to be larger.

In the models, a horizontal resolution of 2.5 km in AROME and 10 km in ARPEGE is equivalent to 9 and 30 min respectively if a wind speed of around 3-5ms^{-1} is considered in the boundary layer. This is consistent with the 20 min used to derive the *tke* from surface point observations. We checked that none of the models directly resolved boundary-layer eddies - even the model with the finest resolution (due to its effective resolution of ~9 Δx, see Ricard et al., 2013). The contribution of the mass-flux scheme in AROME was taken into account by adding the mass-flux contribution, estimated as $0.5 * a_{up} * w_{up}^2$, where a_{up} is the coverage fraction of the thermals and w_{up} the thermal vertical velocity, to the subgrid *tke*. This contribution is small close to the surface and reaches about 20% of the total in the middle of the boundary layer.

Eventually, in order to characterize the afternoon transition (AT), the time at which the buoyancy flux became negative was determined in both observations and models. This was done by finding the 0 cross-over from the interpolation of hourly flux outputs.

Below, we evaluate the representation of the diurnal cycle of the boundary-layer characteristics and surface energy budgets over all 12 IOPs. As shown in Lothon et al. (2014), these days correspond to mainly high-pressure fair-weather conditions with no cloud cover, or, for 14, 15, 24, and 30 June, a small amount of clouds. Most of the days experienced a typical mountain breeze circulation with nocturnal southerly down-slope wind and north-westerly to north-easterly up-slope wind during the days. The 25, 26 and 27 June did not register such circulation (cf Lothon et al., 2014, Fig 6) and were characterized by easterly winds. These three days also showed higher temperature and stronger wind; this was due to the presence of a low pressure system in the Gulf of Lion (for more details see Nilsson et al., 2016a). In the following, these three days will be referred to as hot days.

3 Results

In this section, we compare surface fluxes, meteorological variables, boundary-layer structure and turbulent kinetic energy for the 12 IOPs.

3.1 Radiative and surface fluxes

Figure 3 presents series of 24h sequences of the observed and simulated surface downwelling solar radiation, sensible heat fluxes and latent heat fluxes for the 12 different IOPs (from 14 June to 5 July 2011). The mean value and the maximum range (computed at each time step as the difference between the maximum and the minimum over all the points of either of the models or the observations), averaged for daytime and night-time respectively as a measure of the horizontal variability, are plotted. The cloudy days are clearly depicted by an increase in the horizontal variability of the observed surface downwelling solar radiation (Fig 3a) consistently with Lothon et al. (2014). ARPEGE and AROME mostly distinguish between the clear days (noted 'o') and the cloudy days (noted by triangles) indicated by an increased horizontal variability. For at least two observed clear days (20 June, 27 June), ECMWF depicts a decrease of downwelling solar radiation from 1030 to 1330 UTC; this suggests the presence of clouds in the model. There are some clouds from 1500 UTC to 1900 UTC on 26 June, while ECMWF predicts variability in the downwelling solar radiation from 1030 to 1330 UTC. There are high clouds in ARPEGE throughout the day of 27 June, while observations only registered thin cirrus after 1700 UTC (not shown). Stratocumulus is present in the morning of 30 June, clearing up through the afternoon. Cloud cover remains quite variable in the afternoon, whereas ARPEGE and ECMWF predict a cloud-free atmosphere. The spatial variability is slightly overestimated for 14, 15, 30 June in AROME and underestimated for 24 June but is otherwise in good agreement with observations. In summary, all models capture the spatial and temporal variability in downwelling solar radiation in general with, however, better behaviour for AROME in terms of cloud occurrence and spatial variability.

There is more discrepancy in the simulation of sensible heat fluxes with biases reaching more than 100 Wm^{-2} (Fig 3b). For instance, ECMWF overestimates the surface sensible heat fluxes. The variability from one IOP to another (Fig 3b) is correctly reproduced by all three models with, for instance, a decrease of the maximum sensible heat flux during the hot days. They also all predict more negative sensible heat flux during the nights of the hot period (from 25 to 27 June) even though ECMWF and ARPEGE underestimate this negative sensible heat flux while AROME overestimates its value in the first night (25 to 26 June). Concerning the spatial variability, the large value obtained from the surface sites is noteworthy. The observed range is computed either for all the stations (full black line) or by removing the forest stations (dash-dotted black line). The forest stations induce larger observed ranges especially during the first part of the period. The spatial variability among the various ECMWF grid points is much smaller; this is partly explained by a coarser horizontal grid-size while the value for ARPEGE and AROME is of the same order of magnitude as the observations but slightly underestimated at the end of the period. As shown in Fig 4a, ARPEGE predicts very large sensible heat fluxes for two of the three points (ARP1 and ARP3 mainly differ from ARP2 in terms of altitude and roughness length as shown in Table2). They are of the same order of magnitude as observations recorded at forest sites (dashed and dash-dotted black lines) and are characterized by forest cover, which has a lower albedo (0.12 against 0.2). They are also at higher altitude. However, these simulated sensible heat fluxes are too large to be representative of a 10-km-wide grid box over the area, which, according to Figure 1, cannot be characterized by a uniform forest cover; indeed, there is a large variability of surface covers at scales below 10 km. The third point (northernmost, ARP2) is in better agreement with the non-forest sites (indicated by the black error bars).

There is also discrepancy in the simulation of latent heat fluxes. AROME systematically overestimates the observed values by up to 100 Wm^{-2} (Fig 3c) and this may be related to the soil moisture content being too large (however, no observations were available at various sites to evaluate this variable). The two high-vegetation points of ARPEGE (Fig 4b) do not show evidence of greater evaporation as could have been expected from the larger net radiation (due to the lower albedo). ECMWF correctly reproduces the range of observations. The variability among the various IOPs is also correctly reproduced, with higher latent heat fluxes during the hot days (Fig 3c). The spatial variability is of the same order of magnitude as observed in AROME, slightly underestimated in ARPEGE and strongly underestimated in ECMWF.

Interestingly, when the latent heat fluxes are plotted against the sensible heat fluxes at 1200 UTC, the models reproduce the -1 slope related to an almost constant available energy (cf Supplementary Fig 1), in agreement with LeMone et al. (2003). This is more valid for the clear days (cyan or blue symbols) than the cloudy days (green and purple symbols), in agreement with Lohou et al. (2014). Most of the observations also record a negative relationship (though with a less steep slope) except the observations at 60m on the tower (grey squares) and observations at 30 m over the forest (dots).

To sum-up, we note an overestimation of the sensible heat flux by ARPEGE for the two points covered with forest and, to a lesser extent, by ECMWF and an overestimation of the latent heat flux by AROME (strong bias). All models reproduce the day-to-day variability with in particular the characteristics of the hot period. The observed spatial variability is underestimated for ECMWF probably because of the larger horizontal grid-size and more expanded area for the 9 extracted grid points.

3.2 Meteorological variables

Figure 5 presents the same figures as Figure 3 for the observed and simulated 2-m temperature, 2-m water vapour mixing ratio and the 10-m wind speed. First, all models reproduce the variability of the 2-m temperature through the period with, in particular, a warming from 24/06 to 27/06. In AROME and ARPEGE, the maximum of daytime temperature occurs earlier (by about one hour) than in the observations (note that this could not be analysed in ECMWF with 3-hourly outputs). The main discrepancies occur during the night where the models tend to have a cold bias consistently with common deficiencies of NWP models (Svensson et al., 2011). The spatial variability in night time temperature among sites is smaller for the hot period; this is probably due to higher wind speed during this time (as shown in LeMone et al., 2003 and Acevedo and Fitzjarrald, 2001). The models do not reproduce this behaviour: during the hot period, the models predict both an increasing variability of both night-time sensible heat fluxes and 2 m temperature. The underestimation of the spatial variability by AROME and ARPEGE during most days is not due to a misrepresentation of the wind, which was relatively weak over the whole period and more or less in agreement with observations. ECMWF overestimates the spatial variability. This is partly explained by the westerly grid points being warmer (not shown). Also the diurnal cycle of the spatial variability in ECMWF is inverted compared to observations with higher daily variability than nightly variability. This needs further investigation.

Concerning the 2-m water vapour mixing ratio, the models reproduce the progressive moistening before a precipitating event (the days with precipitation were not IOPs and thus correspond to an interruption of time in Figure 4, indicated by the double vertical dotted lines). Often, observations show morning and evening maxima (e.g. 19 June, 27 June, 30 June, 1 July, 2 July) associated with latent heat flux within a shallow boundary layer and this is reproduced by the models. The models also reproduce the increase in spatial variability during the hot period. There is no clear diurnal cycle in observations and models except in ECMWF which presents a drying at midday leading to a dry bias during daytime especially in the second part of the period. It can be seen that the overestimation of the latent heat fluxes by AROME has no clear consequences in the reproduction of the 2-m water vapour mixing ratio. Concerning the 10-m wind speed ARPEGE and AROME reproduce higher wind speed (greater than 2-3 ms^{-1}) during the hot period with also a larger spatial variability. ECMWF does not reproduce this shift.

In summary, the surface meteorological variables were well simulated in AROME and ARPEGE but were slightly less accurate in ECMWF especially for wind speed and water vapour mixing ratio. In the following sections, we focus only on the French models for which we have hourly outputs.

3.3 Vertical structure

Figure 6 presents scatterplots of the simulated versus observed values of the potential temperature and water vapour mixing ratio averaged over the first 500 m deep layer. First, there is good agreement among all types of observations for potential temperature. Then, the MODEM soundings are drier than the others by about 1 g kg^{-1} consistently with the findings of Agusti-Panareda et al. (2009). AROME and ARPEGE display a cold bias of about 1.5 K. In ARPEGE, the temperature bias is dependent on the average temperature with less bias for temperatures higher than 305 K. ARPEGE does not present a warm bias despite its overestimation of the sensible heat flux for two of the grid points. AROME presents a moist bias, which is consistent with the latent heat flux being too high, while ARPEGE exhibits a dry bias. The AROME moist and cold biases are not clear in the time evolution of 2-m variables, indicating distinct reproduction of the surface layer and the boundary layer.

Figure 7 illustrates the time evolution of the vertical profiles of potential temperature and water vapour mixing ratio (sampled every two hours for clarity) from 12 to 20 UTC for two clear IOPs on 27 June 2011 (one of the hot days) and 1 July 2011. AROME captures the strong inversion in potential temperature that occurs at the top of the boundary layer (at 1400 UTC on 27 June or 1 July) better and this is true for most of the IOPs. This may be due to the finer vertical grid. In both models, there is more spatial variability during the hot period than otherwise and this remains true throughout the day, and is consistent with the results at the surface (higher variability in terms of surface heat fluxes and 2-m meteorological variables) as shown previously. In particular in AROME, on 27 June, the variability among the 16 columns is larger than the variability among the 3 ARPEGE columns even though the area covered by the 16 AROME points is equivalent to the size of one grid of ARPEGE. For 1 July, note the maximum in water vapour mixing ratio in the upper part of the boundary layer simulated by AROME; this maximum is also observed in the radiosoundings. Analysis of the moisture budget indicated that this maximum was mainly related to fine scale advection not resolved at 10 km (not shown).

To further assess the representation of the vertical structure of the boundary layer, we compare the boundary-layer depth estimated by the model with that estimated from observations. The boundary-layer depth is a useful diagnostic to evaluate the representation of boundary-layer evolution in models as it results from the interplay of surface flux, turbulence and subsidence (LeMone et al., 2013). Figure 8 presents the time evolution of the different boundary-layer depth estimates for all the IOPs. The overestimation of the boundary-layer depth by AROME and ARPEGE (more pronounced in ARPEGE) on 14 and 15 June 2011 is explained by the modelled boundary-layer depth criterion based on significant *tke,* which marks the top of the shallow cumulus layer. Both AROME and ARPEGE are able to reproduce days with higher boundary layers compared to days with shallower boundary layers, with, for instance, a shallower boundary layer during the hot days and, the highest on 30 June, 1 July and 2 July (if we discard the 14 and 15 June). The model forecasts are initialized every day so part of the variability among the IOPs is forced through the initial state, but the existence of variability of the boundary-layer depth among the IOPs shows that the physics of the models responds correctly to these differences in weather. Lothon et al. (2014) identified three types of growth of the boundary layer occurring in the morning of the day: typical growth on 20, 24, 25, 30 June and 2 July, slow growth on 26 June, 27 June and 5 July and rapid growth on 14, 19 June and 1 July. The causes of the different types of morning boundary-layer growth are related to the initial profiles, the intensity of the sensible heat fluxes and the intensity of the subsidence as explained in Lothon et al. (2014). This distinction is reproduced by the models. Evaluating the decrease of the boundary layer in the afternoon is more complex. The aerosol diagnosis based on the lidar measurement always shows the top of the inversion layer in the afternoon while the profile diagnosis and the reflectivity gradient from the UHF indicate either the top of the stable layer or the top of the residual layer depending on the case. The model diagnosis depicts the top of the turbulent layer; this is also the case when the boundary-layer depth is diagnosed from the dissipation rate measured by the UHF. The difference between those diagnoses in the afternoon indicates the existence of a pre-residual layer between the top of the turbulent layer and the top of the inversion layer as detailed in Nilsson et al. (2016b). Concerning the decrease of the turbulent layer, ARPEGE predicts a later decrease than AROME most of the time. AROME is in better agreement with the boundary-layer depth diagnosed from the dissipation rate even though AROME

tends to give slightly higher values; this could be explained by the fact that the turbulence variable used to diagnose the boundary-layer depth is different: *tke* instead of dissipation. Also worth noting is the large spatial variability among the model grid points in particular on 26, 27 June and 2, 5 July. However, the highest boundary layer is not systematically over the same grid point, so this can not be explained by particular surface characteristics.

3.4 Turbulent kinetic energy

A unique feature of this campaign was the existence of various simultaneous measurements of the turbulent kinetic energy at various heights in the atmosphere. We used these measurements to evaluate the reproduction of the *tke* by the subgrid turbulence scheme in AROME and ARPEGE. We remind here that despite its fine resolution of 2.5 km, no resolved eddies were simulated in AROME and that we included the mass-flux contribution to the total *tke*.

Figure 9 presents the time evolution of the *tke* for all the IOPs close to the surface and higher in the boundary layer. In the upper panel, the *tke* observed close to the surface, at ~ 8m, is compared to the *tke* modelled at the first level (at 11 m in AROME and 17.5 m in ARPEGE). Often, observations show significant *tke* in the morning, which is not simulated except for a few days (25, 26 and 27 June for AROME and 24 June for ARPEGE), characterized by a greater wind speed and therefore stronger shear production (Fig 5c). There is also significant *tke* in the evening with a minimum around sunset that is also not simulated except for a few days (20, 25, 26 June and 5 July for AROME and 5 July for ARPEGE). This minimum of *tke* is associated with a minimum of wind speed and is present for most days with weak wind. Note that the maximum measured on the evening of the 27 June was associated with convective storms and is reproduced by the models. Those morning and evening *tke* values are related to slope-wind and also potentially to the effect of the nocturnal low-level jet in the early morning. ARPEGE tends to present a Gaussian diurnal cycle of the *tke* for most days (except 3 days: 24 June, 27 June and 05 July, where maximum *tke* exists in the morning or the evening) but with a maximum value consistent with observations. AROME systematically underestimates the maximum value but records a variable diurnal cycle from one day to another. This underestimation is in apparent contradiction with a larger sensible heat flux, at least near the end of the period. The higher value in ARPEGE can be explained by a higher model level (17.5 m versus 11 m, as less turbulence is expected close to the ground) and a larger grid size (9 km versus 2.5 km). Higher in the atmosphere, the modelled and observed *tke* are in better agreement. Note that the various types of observations agree in terms of intensity. The temporal variability at these levels is well reproduced by the models with smaller values during the hot period in agreement with lower buoyancy flux, which is the main source of *tke* during the day (see also Nilsson et al., 2016a). At 60m and higher up, AROME systematically has less *tke* than ARPEGE, as expected from a smaller grid size.

Figure 10 illustrates the time evolution of vertical profiles of the turbulent kinetic energy modelled and observed for 1 July (this was the only day where we had enough observations to retrieve a time-varying vertical profile of the *tke*). AROME has larger *tke* than ARPEGE around mid-day and it decreases the turbulence more rapidly. The shape of the vertical profiles is consistent between each model and the observations. The lidar observations (triangles, note that this is a *tke* estimate deduced from the turbulent variance of the vertical velocity) indicate a more or less stationary value in the middle of the boundary layer from 1400 to 1600 UTC; this is not simulated by the models. However, it should not be forgotten that the lidar only measures the vertical velocity variances by assuming A=1 (same contribution from vertical and horizontal velocity variances). But a comparison of the square (tethered balloon) and the triangle (Doppler lidar) symbols of the same colour and at the same altitude gives an idea of the error on this estimation: A is underestimated during daytime with values more around 1.3-1.8 (smaller contribution from vertical wind variances) while A is overestimated in late afternoon (1700 and 1800 UTC) with A around 0.4-0.8 (stronger contribution from horizontal wind variances). This deserves further investigation with more measurements of the vertical profiles. Also, comparison of the shear contribution with the buoyancy contribution in the creation of *tke* and the *tke* budget in general could be further analysed in observations and models.

3.5 Afternoon transition

In this section, we focus on the afternoon transition period. During this period, the turbulence regime changes from the fully convective regime of turbulence, close to homogeneous and isotropic, towards more heterogeneous and intermittent turbulence. Most of the terms in the TKE equation -buoyancy production, shear production, dissipation and vertical transport- are small (Nilsson et al., 2016b).

Concerning the evolution of the boundary layer in the afternoon, the IOPs can be separated into the two categories proposed by Grimsdell and Angevine (2002) as defined by the behaviour of the UHF reflectivity with 24/06, 30/06, 1/07 and 2/07 pertaining to the inversion layer separation cases (ILS, so-called by Grimsdell and Angevine, 2002, where the height of the reflectivity gradient stays more or less at the same height as the maximum registered during the day) and 25/06, 26/06, 27/06 pertaining to the descent cases (where the height of the reflectivity gradient decreases with time in the evening). As in Grimsdell and Angevine (2002), the ILS cases are colder and drier days characterized by strong inversion of potential temperature at the top of the boundary layer and associated with strong shear as shown in Nilsson et al. (2016a). These cases have also a strong inversion reproduced by the models (not shown except for 1 July). The descent cases are warmer and moister days corresponding to the hot period. However, the height of the strongest gradient in the UHF reflectivity is more representative of the top of the inversion layer and does not really determine the top of the turbulent layer, which is better indicated by the height derived from the dissipation rate (in pink in Fig 8). This height is more comparable to the boundary-layer depth diagnosed in the models, which makes sense as *tke* and dissipation rate are closely related. AROME always predicts an earlier decrease of turbulence than ARPEGE and agrees better with the evolution of the height derived from the dissipation rate. The layer between the pink and the red symbols was named the pre-residual layer by Nilsson et al., (2016b). It is characterized by very low turbulence and results from the adjustment of turbulence to the decreasing surface fluxes (Darbieu et al., 2015).

Figure 11 presents the variations of the time when the virtual temperature flux (which is a combination of the surface sensible heat flux and the latent heat flux) becomes negative, t_Hv0, through the IOPs and the various points. This time varies strongly from one surface to the other in the observations as already shown by Lothon et al. (2014, their Fig 8 and black symbols in Fig 9), suggesting that the vegetation partly drives the delay of the transition from one site to the other. The range of t_Hv0 among the three points of ARPEGE (blue symbols) is less than one hour except during the hot period (26 and 27 June) and 1 July. The range of t_Hv0 is much larger in AROME (green symbols) with a range varying from 2 hours to 6 hours with, however, no systematic behaviour for a given point (indicated by a given symbol). AROME systematically has an earlier t_Hv0 than ARPEGE, consistently with an earlier decrease of turbulence. Also this occurs earlier during the hot period than on the other days and this is reproduced by the models. In observations and models, the spatial variability is the strongest during the hot period.

In summary, the models do a relatively good job during the afternoon. This could be related to the quasi-stationary behaviour discussed in Darbieu et al. (2015) and Nilsson et al. (2016a), where no changes in turbulence structure or characteristics are evident after normalization by the decreasing surface sensible heat fluxes. The difficulties increase in the very late afternoon. We have also noted more difficulties when the models attempt to reproduce the varying characteristics of close-to-surface variables at night. This highlights the models' difficulties in reproducing stable conditions.

4. Conclusions

The BLLAST field campaign gathered a large dataset, in particular high-frequency observations of the vertical structure of the boundary layer and observations of the turbulent kinetic energy; this enabled us to extensively evaluate three numerical weather prediction models. In summary, all models reproduced the temporal variability observed among the different IOPs in terms of variations of the cloud amount (clear versus partly cloudy conditions), maximum height of the boundary layer, and variations of temperature. This is also a necessary first step if we want to use such models further to derive the large-scale fields, *e.g.* large-scale advection, that are needed for smaller scale modelling studies. For instance,

during the hot period, models and observations produced lower sensible heat fluxes, higher temperature, stronger winds, and weaker *tke* than during the other days. The different types of growth of the boundary layer encountered during the field campaign and detailed in Lothon et al. (2014) were correctly distinguished by AROME and ARPEGE. However, systematic biases appeared over the 12 IOPs: too-large latent heat fluxes in AROME, a too-large diurnal amplitude of relative humidity at 2 m and a dry bias during the day for ECMWF (especially at the end of the period). For two ARPEGE points, the surface fluxes were similar to measurements over forest; but the satellite data do not indicate a homogeneous forest patch over 10 x 10 km² in this 10 x 10 km² area. AROME reproduced the vertical structures better and also the variability in boundary-layer depth among the different IOPs in terms of daily maximum value or growth in the morning. The spatial variability reproduced by AROME was similar to the one derived from the various in-situ surface sites.

For the first time, turbulent kinetic energy, the prognostic variable of the turbulence scheme in AROME and ARPEGE, has been evaluated. Both models reproduced the right order of magnitude. AROME reproduced the variation from one day to another of its diurnal cycle better while ARPEGE always predicted a similar bell shaped evolution. However, AROME underestimated the value while ARPEGE was in better agreement with the observed intensity. Note that we took the contribution of the mass-flux scheme to the *tke* into account here. This may be due to differences not only in grid-size but also in physical parametrization. In a future study, we could gain some insight by evaluating the different simulated terms of the near-surface *tke* budget that have also been derived from observations by Nilsson et al. (2016a).

In summary, this study is a first attempt to analyse the improvements provided by high-resolution numerical weather prediction. AROME seemed to depict the mesoscale spatial and temporal variability better. However, future studies are needed to determine the exact role of the increase in resolution versus the change in physical parametrization.

Acknowledgements

The authors thank F. Said for providing the tower measurements, J. Reuder for providing the SUMO measurements, F. Gibert for providing the lidar measurements, P. Augustin for the boundary-layer diagnostic derived from the lidar, E. Pardyjak, D. Alexander and C. Darbieu for the forest flux measurements, D. Legain for the contribution of the CNRM to the field campaign, P. Durand for the Piper Aztec turbulence data process, B Piguet for the tethered-balloon turbulence data process, LEOSPHERE for providing the Doppler lidar to the field campaign. The BLLAST field experiment was made possible thanks to the contribution of several institutions and supports: INSU-CNRS (Institut National des Sciences de l'Univers, Centre National de la Recherche Scientifique, LEFE-IMAGO program), Météo-France, Observatoire Midi-Pyrénées (University of Toulouse), EUFAR (European Facility for Airborne Research), BLLATE-1&2, COST ES0802 (European Cooperation in the field of Scientific and Technical). The field experiment would not have occurred without the contribution of all participating European and American research groups, which all have contributed in a significant amount. The Piper Aztec research airplane is operated by SAFIRE, which a unit supported by INSU-CNRS, Météo-France and the French Spatial Agency (CNES). BLLAST field experiment was hosted by the instrumented site of Centre de Recherches Atmosphériques, Lannemezan, France (Observatoire Midi-Pyrénées, Laboratoire d'Aérologie). This research has also been carried out in the framework of the DEPHY2 project supported by INSU-CNRS through the LEFE-IMAGO program and the Ministery for Environment, Energy and Sea**.**

References

Acevedo O. C. and Fitzjarrald, D. R.: The Early evening surface-layer transition: temporal and spatial variability. J Atmos Sci, 58, 2650-2667, 2001

Atlaskin E., Vihma T.: Evaluation of NWP results for wintertime nocturnal boundary-layer temperatures over Europe and Finland. Q J R Meteorol Soc, 138, 1440-1451, 2012

Balsamo, G., P. Viterbo, A. Beljaars, B. van den Hurk, M. Hirsch, A. Betts, and K. Scipal, 2009: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. J. Hydrometeor., 10, 623–643

Bennett L J, Weckwerth T, Blyth A M, Geerts B, Miao Q, Richardson Y, 2010: Observations of the evolution of the nocturnal and convective boundary layers and the structure of open-celled convection on 14 June 2002. Mon Wea Rev, 138, 2589-2607

Bougeault P.: Cloud ensemble relations based on the gamma probability distribution for the higher-order models of the planetary boundary layer. *J Atmos Sci* 39:2691–2700, 1982

Bouniol D, Protat A, Delanoé J, Pelon J, Piriou JM, Bouyssel F, Tompkins A, Wilson D R, Morille Y, Haeffelin M, O'Connor E J, Hogan R, Illingworth AJ, Donovan D P, Baltink HK : Using continuous ground-based radar and lidar measurements for evaluating the representation of clouds in four operational models. J Appl Meteorol Clim, 49 : 1971-1991, 2010

Boyouk N, Leon JF, Delbarre, H, Podvin T, Deroo C: Impact of the mixing boundary layer on the relationship between PM2.5 and aerosol optical thickness, Atmos Env, 44, 271-277, 2010

Canut, G., Couvreux F, Lothon M, Legain D, Piguet B, Lambert A, Maurel W, Moulin E: Turbulent fluxes and variances measured with a sonic anemometer mounted on a tethered-balloon, in revision for Atmospheric Measurement Techniques, 2016

Collaud Coen M, Praz C, Haefele A, Ruffieux D, Kaufmann P, Calpini B: Determination and climatology of the planetary boundary layer height above the Swiss plateau by in situ and remote sensing measurements as well as by the COSMO-2 model; Atmos Chem Phys, 14, 13205-13221, 2014

Courtier, P. and Geleyn, J.-F.: A global numerical weather prediction model with variable resolution – Application to the shallow-water equations. Q. J. R. Meteorolog. Soc.114, 1321-1346, 1988

Cuxart J, Bougeault P, Redelsperger, JL.: A turbulence scheme allowing for mesoscale and large-eddy simulations. *Q. J. R. Meteorol. Soc.* 126: 1-30, 2000

Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek M, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld GJ, Svensson G, Taylor P, Weng W, Wunsch S, Xu KM: Single Column model intercomparison for a stably stratified atmospheric boundary layer. Boun Lay Meteorol, 118, 273-303, 2006

De Coster O, Pietersen, H: BLLAST- uniform processeing of Eddy-Covariance data, http://bllast.sedoo.fr/documents/reports/H-Pietersen_O-de-Coster_BLLAST-surf_flx-uniform-processing.pdf, 2012

Forbes R, Tompkins A M, Untch A.:A new prognostic bulk microphysics scheme for the IFS. ECMWF Tech Memorandum No 649, 28pp, 2011

Geleyn J.F.: Interpolation of wind, temperature and humidity values from model levels to the height of measurement. Tellus, 40A, 347-351, 1988

Giard D, Bazile E, 2000: Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. Mon Wea Rev, 128, 997-1015

Gibert F, Arnault N, Cuesta J, Plougonven R, Flamant P: Internal gravity waves convectively forced in the atmospheric residual layer during the morning transition. *Q. J. R. Meteorol. Soc.* 137: 1610-1624, 2011

Gibert F, Dumas A, Thobois L, Bezombes Y, Koch G, Dabas A, Lothon M: Afternoon transition turbulence decay revisited by Doppler Lidar, Symposium on boundary layer and turbulence, Boston, USA, 2012

Grimsdell A, W Angevine: Observations of the afternoon transition of the convective boundary layer. *J Appl Meteorol*, 41: 3-11, 2002

Guichard, F., D. B. Parsons, J. Dudhia, and J. Bresch: Evaluating mesoscale model predictions of clouds and radiations with SGP ARM data over a seasonal timescale. Mon. Wea. Rev., 131, 926–944, 2003

Hartogensis O. K: BLLAST Flux maps, http://bllast.sedoo.fr/workshops/february2015/ presentations/Hartogensis-Oscar_area-averaged-flux.pdf., 2015

Holt, T. Raman S: A review and comparative evaluation of multilevel boundary layer parameterizations for first-order and turbulent kinetic energy closure schemes. Rev Geophys, 26, 761-780, 1988

Illingworth AJ, Hogan RJ, O'Connor EJ, Bouniol D, Brooks ME, Delanoé J, Donovan DP, Eastment JD, Gaussiat N, Goddard JWF, Haeffelin M, Klein Batink H, Krasnov O A, Pelon J, Piriou JM, Protat A, Russchenberg HWJ, Seifert A, Tompkins AM, Van Zadelhoff G-J, Vinit F, Willen U, Wilson DR, Wrench CL: Cloudnet, Continuous evaluation of cloud profiles in seven operational models using ground-based observations. Bull Am Meteorol Soc, 883:898, 2007

Koehler M, Ahlgrimm M, Beljaars A: Unified treatment of dry convective and stratocumulus topped boundary layer in the ECMWF model. QJ R Meteorol Soc 137:43-57, 2010

Legain D., Bousquet, O., Douffet, T., Tzanos, D., Moulin, E., Barrie, J. and Renard, J.-B.: High frequency boundary layer profiling with reusable radiosondes. Atmos. Meas. Tech. Discuss., 6, 3339-3365, 2013.

LeMone M, A, Grossman R, L, Chen F, Ikeda, K, Yates D: Choosing the averaging interval for comparison of observed and modeled fluxes along aircraft transects over a heterogeneous surface, J Hydrometeorol,4, 179:195, 2003

LeMone M, A, Tewari M, Chen F, Dudhia J: Objectively determined fair-weather CBL depths in the ARW-WRF model and their comparison to CASES-97 Observations. Mon Weather Rev, 141, 30-54, 2013

Lohou F, Patton E,G: Surface energy balance and buoyancy response to shallow cumulus shading. J Atmos Sci, 71, 665-682, 2014

Lothon M,et al: The BLLAST field experiment: Boundary-Layer Late Afternoon and Sunset Turbulence, ACP, 2014

Masson, V. and Y., Seity: Including atmospheric layers in vegetation and urban offline surface schemes, Journal of Applied Meteorology and Climatology, 48, 7, 1377-1397, 2009

Morcrette, J.-J.: Radiation and cloud radiative properties in the ECMWF operational forecast model. J. Geophys. Res., 96, 9121–9132, 1991

Nilsson E, F Lohou, M Lothon, E Pardyjak, L Mahrt, C Darbieu: Turbulence kinetic energy budget during the Afternoon transition, Part A: Observed surface tke budget and boundary layer description for 10 Intensive Observation Period Days, in revision for Atmos Chem Phys, 2015a

Nilsson E, M Lothon, F Lohou, E Pardyjak, O Hartogensis, C Darbieu: Turbulence kinetic energy budget during the Afternoon transition, Part B: A simple TKE model, in revision for Atmos Chem Phys , 2015b

Pergaud J, Masson V, Malardel S, Couvreux F.: A parameterization of Dry thermals and shallow cumuli for mesoscale numerical weather prediction. *Boundary-Layer Meteorology*. **132**, 83-106. DOI 10.1007/s10546-009-9388-0, 2009

Reuder, J., Jonassen, M., and Olafsson, H.: The Small Unmanned Meteorological Observer SUMO: Recent Developments and Applications of a Micro-UAS for Atmospheric Boundary Layer Research, Acta Geophys., 60, 1454–1473, 2012

Ricard D, Lac C, Riette S, Legrand R, Mary A: Kinetic energy spectra characteristics of two convection-permitting limited-area models AROME and Meso-NH: Q J Royal Meteorol Soc, 139, 1327-1341, 2013

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France Convective-Scale Operational Model, Mon. Weather Rev., 139, 976–991, 2011

Simmons, A J, Burridge, D M, Jarraud, M, Girard, C, Wergen W: The ECMWF Medium Range Prediction models development of the numerical formulations and the impact of increased resolution. Meteorol Atmos Physics, 40, 28-60, 1989

Smith RNB: A scheme for predicting layer clouds and their water content in a general circulation model. *Q J R Meteorol Soc*. 116 : 435–460, 1990

Steeneveld G J, Mauritsen T, De Bruijn E I F, Vila-Guerau de Arellano J, Svensson G and Holstlag A A M: Evaluation of limited-area models for the representation of the diurnal cycle and contrasting nights in CASES-99. J Applied Meteorology and Climatology, 47, 869-887, 2008

Svensson G, Holtslag AAM, Kumar V, Mauritsen T, Steeneveld GJ, Angevine WM, Bazile E, Beljaars A, de Bruijn EIF, Cheng A, Conangla L, Cuxart J, Ek M, Falk MJ, Freedman F, Kitagawa H, Larson VE, Lock A, Mailhot J, Masson V, Park S, Pleim J, Söderberg S, Weng W, Zampieri M: Evaluation of the diurnal cycle in the atmospheric boundary layer over land as represented by a variety of singlecolumn models: the second GABLS experiment. Boundary-Layer Meteorol,140, 177–206, 2011

Tiedtke M.: A comprehensive mass flux scheme for cumulus parametrization in large-scale models. Mon Weather Rev. 117: 1779–1800,1989.

Tables:

Table 1. List of the instruments and their spatial and temporal resolutions

Instrument	Used measured parameters	Derived diagnostics	Time resolution/range	Spatial resolution/range	Location
Standard radiosoundings (MODEM, M10 probes)	q, q_v, wind speed	h_{BL}	0000, 0600, 1200, 1800 UTC	~10-15 m/0-20k m	Main site
Low-troposphere radiosoundings (VAISALA RS92 probes)	q, q_v, wind speed	h_{BL}	Hourly from 1200 to 2200 UTC in IOP	~10-15 m/0-2 km	
Turbulence station (eddy-covariance system)	T2m, q2m, ws10m, sensible & latent heat flux, u'^2, v'^2, w'^2		30 min from 20 Hz (except the forest site that has 10 Hz) sampling rates		7 stations over wheat, grass, forest, moor, corn
Radiative flux station (radiometers)	incoming & outgoing shortwave and longwave radiation		1 Hz sampling rates		Moor, Corn, Forest, main tower sites
UHF	refractive index structure coefficient, Turbulent energy dissipation rate	h_{BL}	5 min consensus (2 cycles over 5 beams)	~75 m /175 m-4000 m	
Doppler lidar	Vertical velocity	tke	4s time resolution; turbulence moments calculated on *20* min	50 m	
Aerosol lidar	Aerosol backscatter	h_{BL}	4s time resolution but diagnostic derived every 15 min	15 m	Main site
French Piper Aztec aircraft	3-D wind	tke	25 Hz high rate measurements moments calculated on 5-7 min samples	~3m spatial resolution of the high rate measurements; aircraft velocity of 70 m/s; turbulence moments calculated over 30-40 km legs stabilized in attitude & altitude	
Remote piloted aircraft system SUMO	q, q_v, wind speed		2Hz for thermo and 100 Hz for wind		Main site
Tethered Balloon with a turbulence probe	u'^2, v'^2, w'^2	tke	20 min from 10 Hz sampling rates		Main site

Table 2. Description of the three models

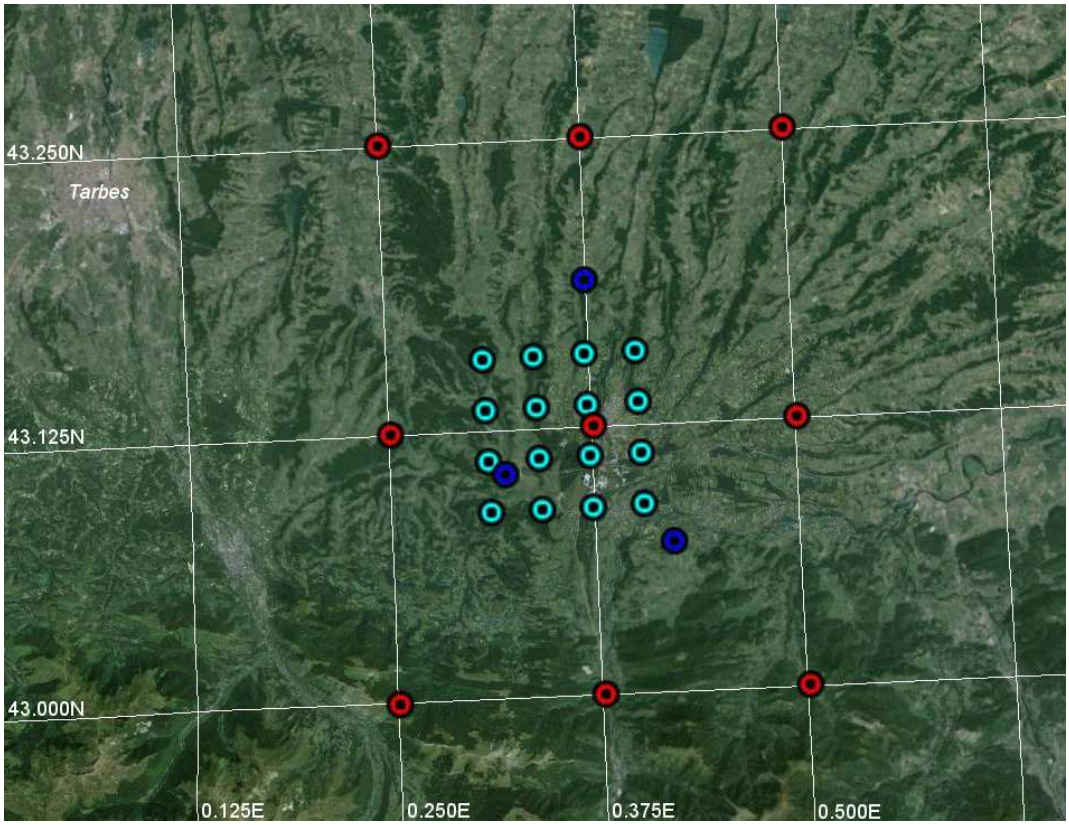
Model	Horizontal resolution	Number of vertical levels (in the 1st km)/ 1st level altitude	time step (mn)	Surface scheme	PBL scheme	Initialization time/ model run length (hours)	Initialisation of land-surface properties
AROME	2.5 km	60 (15) / 10m	1	SURFEX	TKE prognostic scheme + Mass flux scheme for dry and cloudy thermals	00TU; 30	From a surface reanalysis

ARPEGE	10 km	70 (11) / 16m	10	ISBA	TKE prognostic scheme + mass-flux scheme when cumulus are present	00 TU; 36	From a surface reanalysis
ECMWF	16 km	91 (11)/ 10m	10	HTESSEL	Non-local K profile; mass-flux scheme	00-06-12-18 TU; 06	From a surface reanalysis

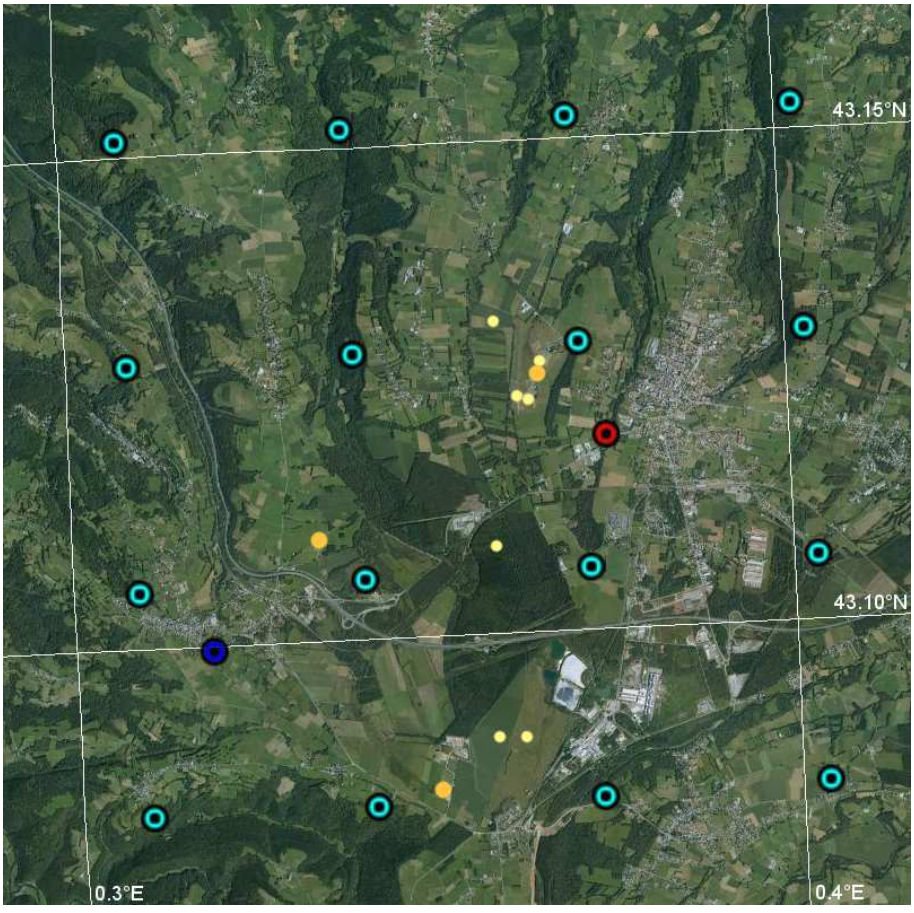
Table 3. Surface characteristics of the various points extracted from the models: the surface characteristics,i.e. albedo, vegetation fraction (the complementary being bare soil), LAI and roughness length correspond to the total value for the grid point. In ARPEGE and ECMWF the roughness length takes into account the subgrid orography.

Points	Altitude (m)	Albedo	Vegetation fraction	LAI	Roughness length (m)	Dominant vegetation type
ARO-1	535	0.18	0.95	3.4	0.78	Broad leaved forest (62%), land (38%)
ARO-2	611	0.19	0.93	3.5	0.53	Cultures (38%): Broad leaved forest (37%), land (25%)
ARO-3	595	0.19	0.92	3.2	0.26	Land(38%), Cultures (25%), Town (25%), Broad leaved forest (12%)
ARO-4	558	0.20	0.92	3.4	0.16	Cultures(67%), land (33%)
ARO-5	552	0.20	0.92	3.5	0.24	Cultures(67%), land (25%), Broad leaved forest (8%)
ARO-6	605	0.19	0.93	3.4	0.38	Cultures(42%), land (33%), landes-forest (8%)
ARO-7	609	0.16	0.85	3.3	0.45	Land (42%), Landes forest (33%) Town (25%)
ARO-8	593	0.17	0.94	3.2	0.39	Land (56%) Landes forest (33%) Cultures (11%)
ARO-9	532	0.19	0.93	3.5	0.49	Cultures (42%) Land (25%) Broad Leaved Forest (33%)
ARO-10	567	0.19	0.91	3.7	0.37	Cultures (83%) Broad leaved forest (17%)
ARO-11	579	0.20	0.91	3.3	0.18	Cultures (60%), Town (20%) Land (20%)
ARO-12	575	0.19	0.91	3.5	0.47	Cultures (35%) Mixtures (27%) Broad leaved forest (18%) Town (10%)
ARO-13	505	0.18	0.93	3.8	0.83	Broad leaved forest (58%) Cultures (42%)
ARO-14	521	0.18	0.92	3.7	0.64	Cultures (58%) Broad leaved forest (42%)
ARO-15	529	0.19	0.88	3.2	0.23	Cultures (78%) Mixtures (22%)
ARO-16	527	0.19	0.90	3.5	0.38	Cultures (75%) Broad leaved forest (17%) Landes forest (8%)
ARP-1	701	0.12	0.86	3.7	1.8	Forest
ARP-2	477	0.2	0.84	3.2	0.17	Cultures
ARP-3	778	0.12	0.85	3.6	1.93	Forest
ECMWF-1	1068	0.15	Not available	Not available	6.2	Not available
ECMWF-2	894	0.15	Not available	Not available	5.1	Not available
ECMWF-3	772	0.15	Not available	Not available	4.8	Not available
ECMWF-4	510	0.15	Not available	Not available	0.65	Not available
ECMWF-5	491	0.15	Not available	Not available	0.62	Not available
ECMWF-6	463	0.15	Not available	Not available	0.88	Not available
ECMWF-7	282	0.15	Not available	Not available	0.65	Not available
ECMWF-8	314	0.15	Not available	Not available	0.62	Not available
ECMWF-9	325	0.15	Not available	Not available	0.62	Not available

Figures



(a) Big Domain



(b) Small Domain

1

Figure 1: (Upper panel) Map of the different points extracted from the models (red for ECMWF, blue for ARPEGE and cyan for AROME).(Lower panel) zoom of (upper panel) with surface sites shown by small yellow dots and radiosoundings launching site in large orange dots. Note that the western most site was the site for launching the few GRAW soundings that were not used in this study (Google Earth Source).

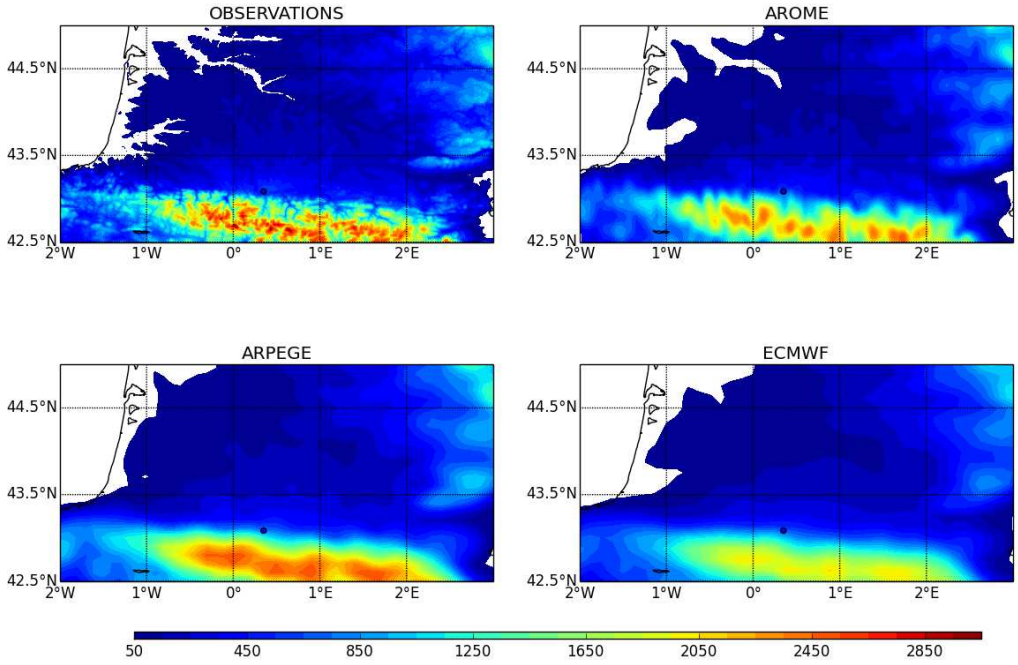


Figure 2: Orography as (upper left) observed (dataset ETOP01, 1 arc-minute global relief of Earth's surface: doi:10.7289/V5C8276M) or modeled in AROME (upper right), ARPEGE (lower left) and ECMWF (lower right), isocontours every 100m are drawn.

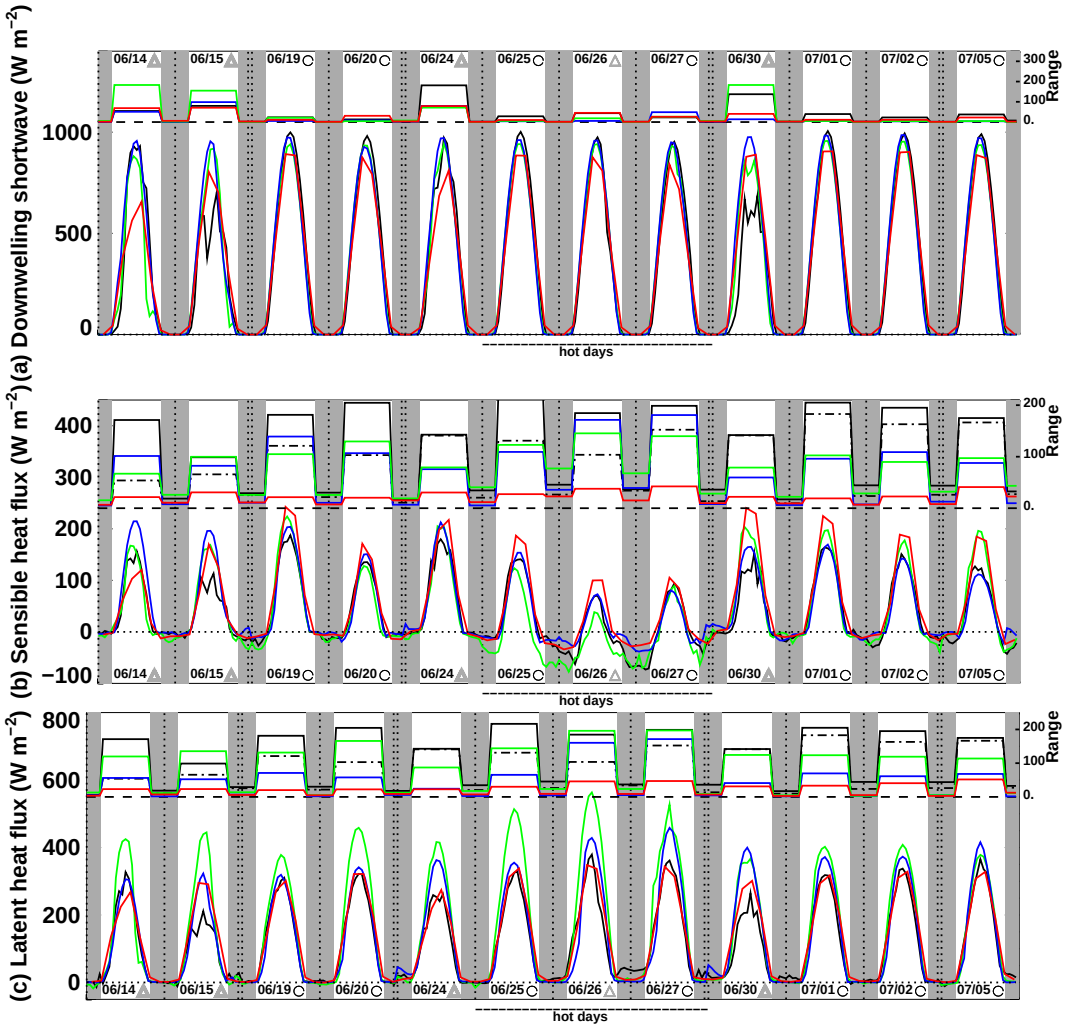


Figure 3: Time series of 24h sequences for the 12 IOPs of (a) surface downwelling solar flux, (b) sensible heat flux and (c) latent heat flux, measured over surface sites in black, simulated by ARPEGE in blue, by AROME in green and ECMWF in red with the mean value (left axis) and the maximum horizontal range (right axis), computed as the difference between the maximum value and the minimum value for all sites or all grid points of a given model but averaged respectively over day and night; for observations both the range computed with all sites (full line) or by removing the forest stations (dash-dotted lines). The vertical grey shading marks the nighttime. Two consecutive vertical dashed lines indicate interruption in the days. Note that for ARPEGE, due to the different behaviour of ARP1 and ARP3, only ARP2 is plotted as the mean while the spatial variability is computed with the three points.

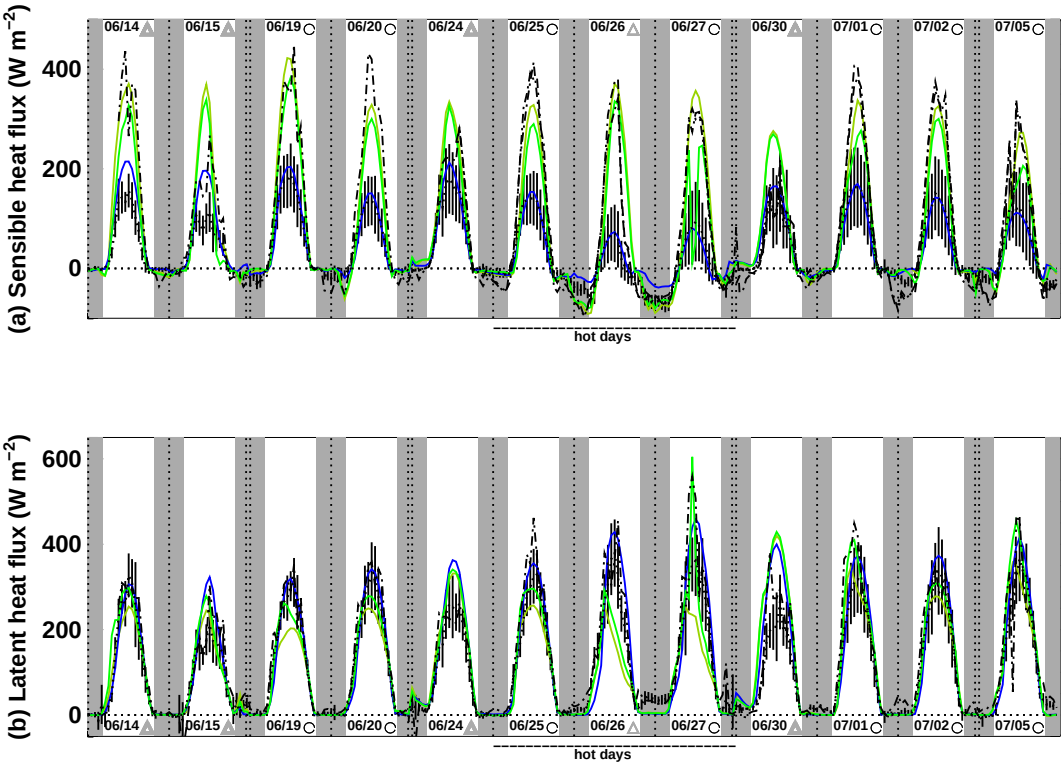


Figure 4: Time series of 24h sequences for the 12 IOPs of (a) sensible heat flux and (b) latent heat flux. Measurements over several surfaces are indicated in black curve for the mean with horizontal standard deviations indicated by error bars; the dashed and dot-dashed black lines correspond to the observations over the forest sites that are not included neither in the mean nor in the horizontal standard deviations. Values simulated by ARPEGE are indicated in dark blue for point 2, light-green for point 1 and green for point 3.

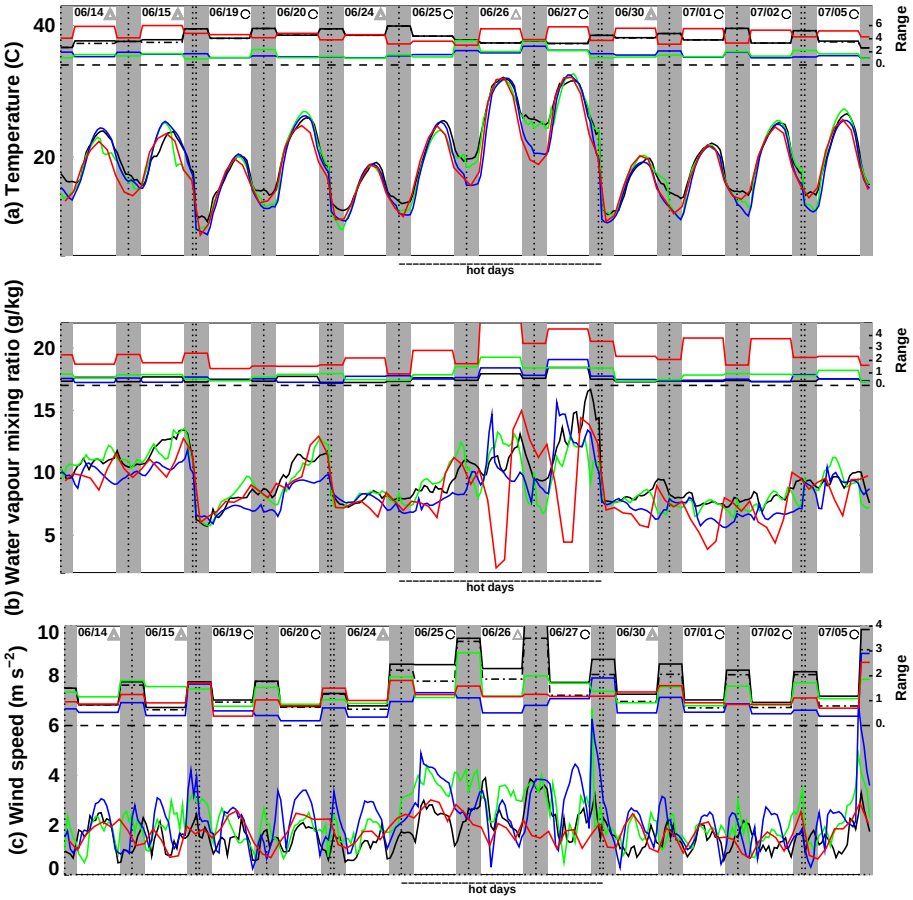


Figure 5: Time series of 24h sequences for the 12 IOPs of (a) 2m temperature, (b) 2m water vapour mixing ratio and (c) 10m-wind speed, measured over several surfaces in black, simulated by ARPEGE in blue, by AROME in green and ECMWF in red with the mean value (left axis) and the maximum horizontal range (right axis, computed as the difference between the maximum value and the minimum value for all sites or all grid points of a given model but averaged respectively over day and night). The vertical grey shading marks the nighttime. Two consecutive vertical dashed lines indicate interruption in the days. Note that for ARPEGE, due to the different behaviour of ARP1 and ARP3, only ARP2 is plotted as the mean while the spatial variability is computed with the three points.

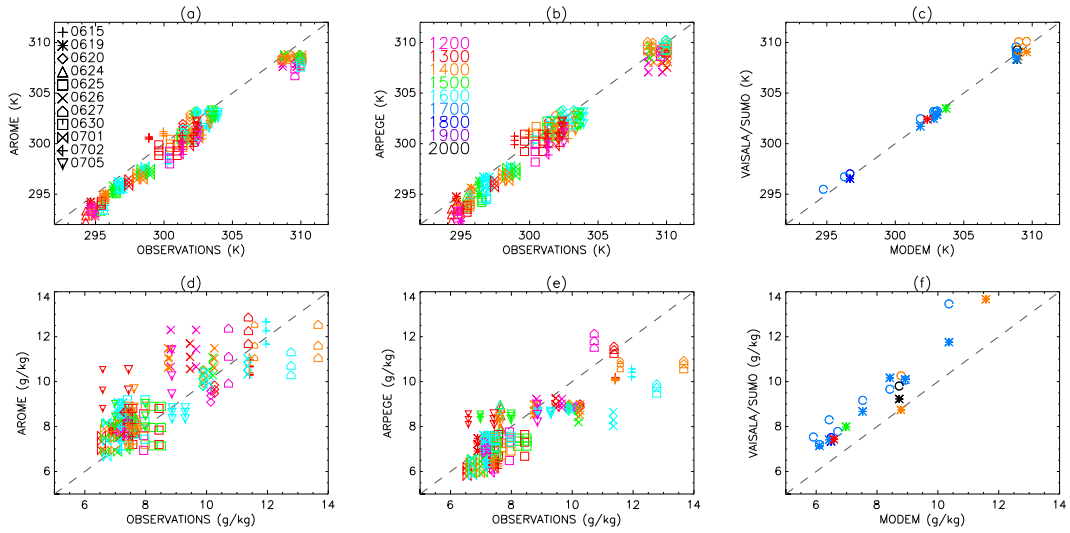
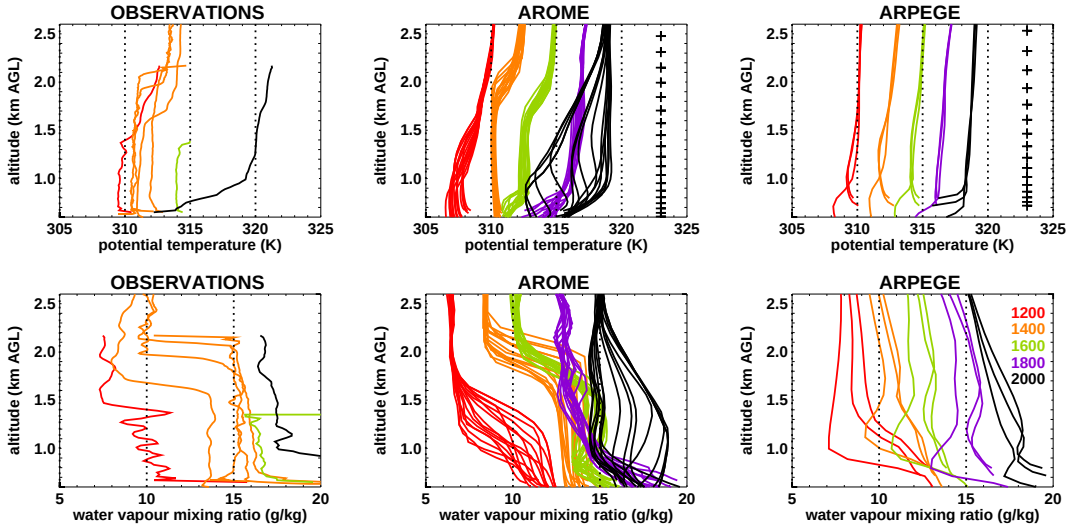
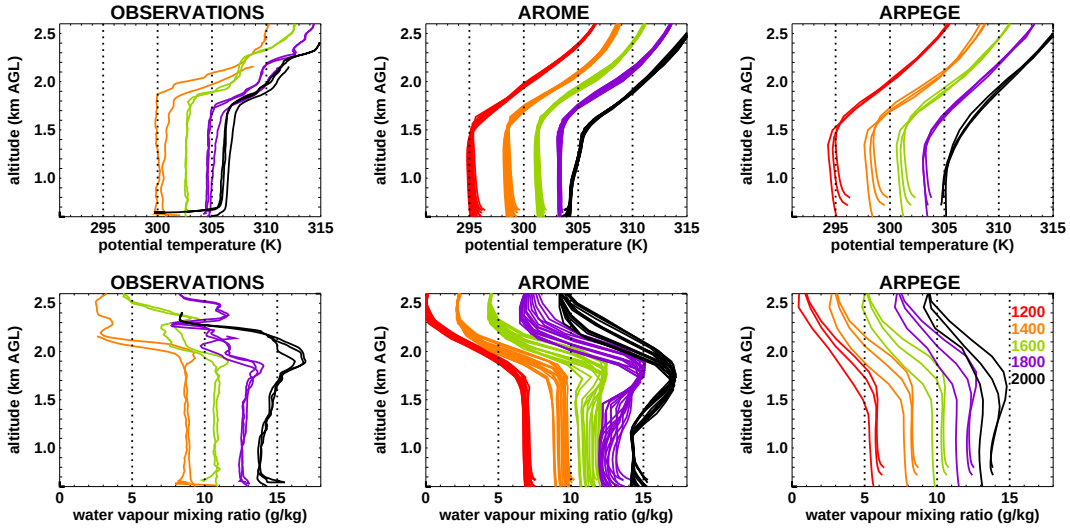


Figure 6: Scatterplot for (a,b,c) the potential temperature and (d, e, f) the water vapour mixing ratio averaged over the first 500 m deep layer: (a and d) AROME values versus the observed values, (b and e) ARPEGE values versus the observation values and (c and f) values obtained from the Vaisala and the SUMO profiles versus the values obtained from the MODEM profiles. Symbols vary from one day to the other and color from one time to the other (see legend).



(a) 27 June 2011



(b) 01 July 2011

Figure 7: Vertical profiles of potential temperature and water vapour mixing ratio for observations (left panels), AROME (middle panels) and ARPEGE (right panels) for two days the 27 June 2011 (upper panels) and the 01 July 2011 (lower panels). For visibility purposes, the vertical profiles are offset by adding 2K or 2 g/kg every two hours from 1400 to 2000.

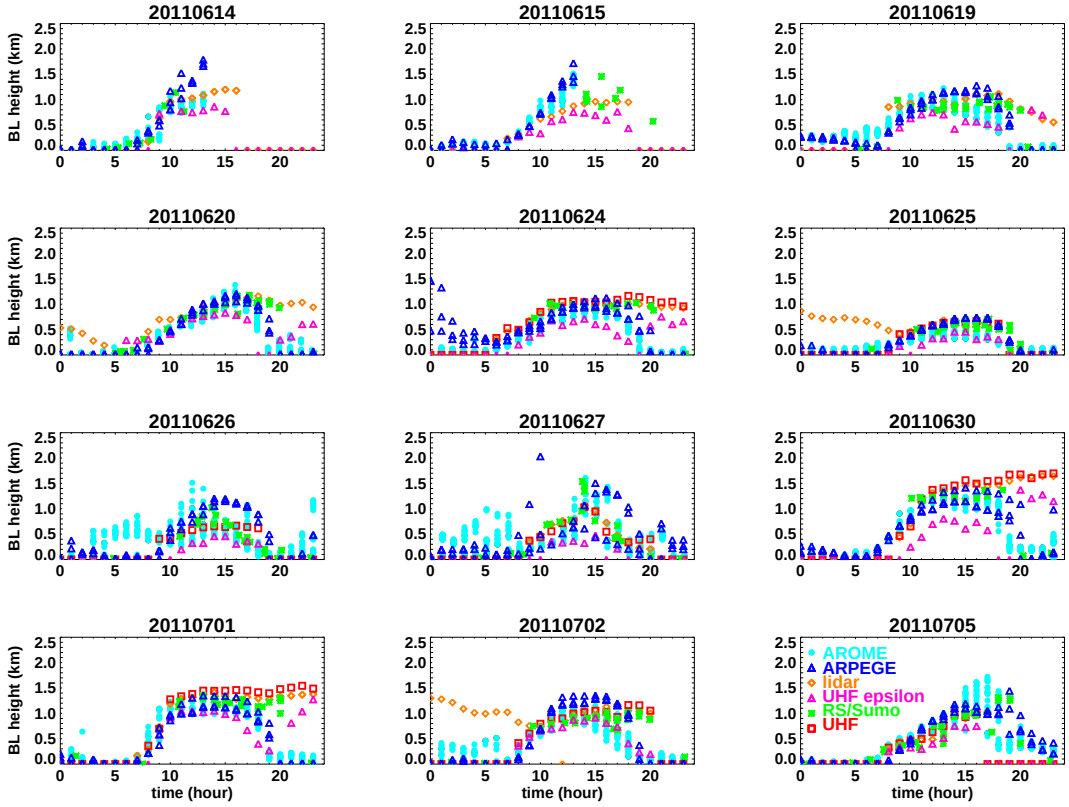


Figure 8: Time series of boundary-layer height for each IOP observed by aerosol lidar(orange diamonds), UHF (from reflectivity in red squares and from the dissipation in pink triangles), radiosoundings or SUMO profiles (green stars) or simulated by ARPEGE (blue triangles) or AROME (cyan full circles). As indicated in the text, no value is drawn from ARPEGE and AROME after 1400 UTC on 14 and 15 June as the existence of clouds induce that the boundary-layer height diagnostic depicts in fact the top of the shallow clouds.

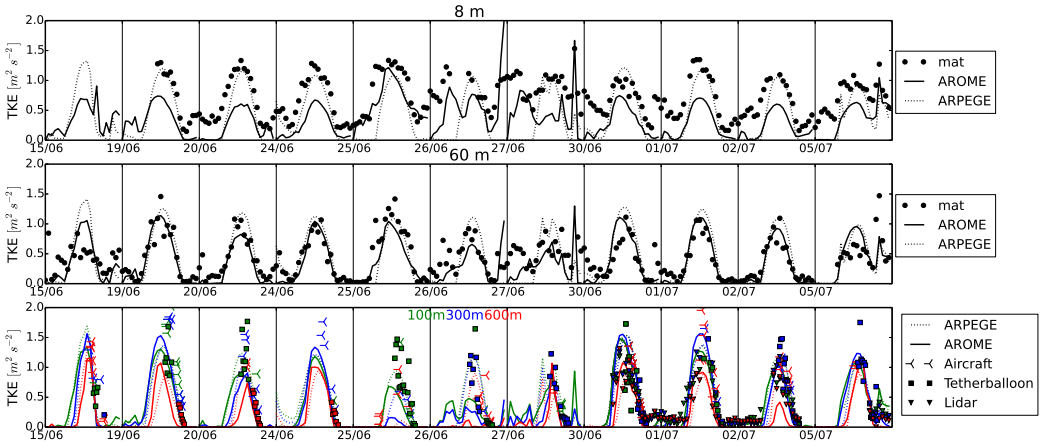


Figure 9: Time series of turbulent kinetic energy observed (in symbols) or simulated by AROME (full line) and ARPEGE (dotted line) at (a) 8m above ground level for observations, 11m for AROME and 17.5m for ARPEGE, (b) 60m above ground level and (c) 100m, 300m and 600m above ground level for the different IOPs from 15 June to 5 July.

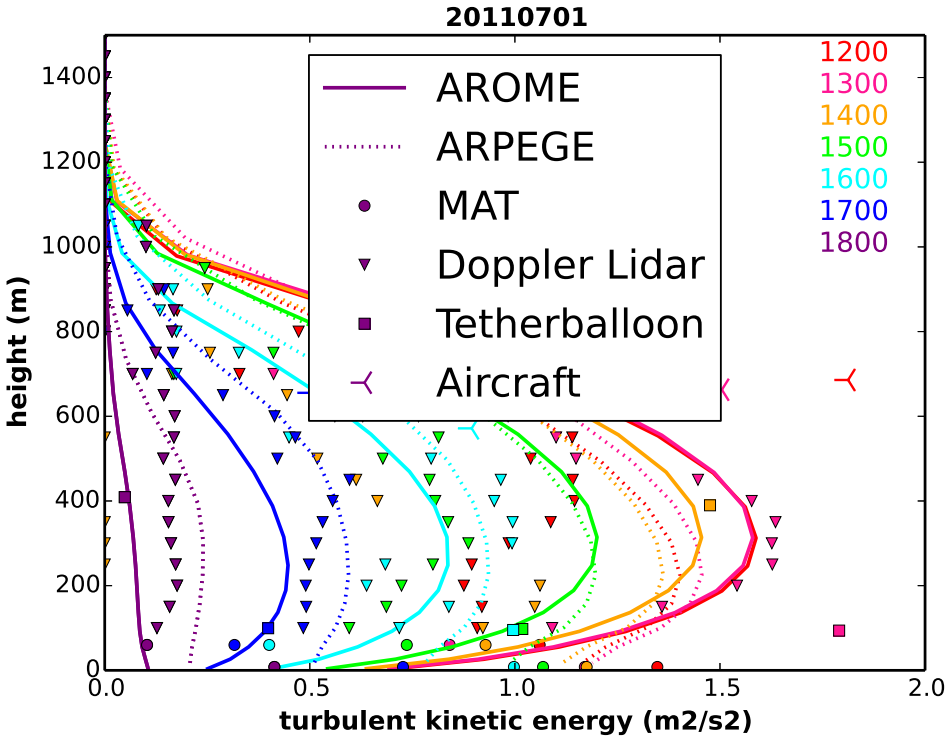


Figure 10: Vertical profiles of the turbulent kinetic energy modelled by AROME (full lines) and ARPEGE (dotted lines) from 1200 to 1800 UTC (see legend), when available, observations are overplotted.

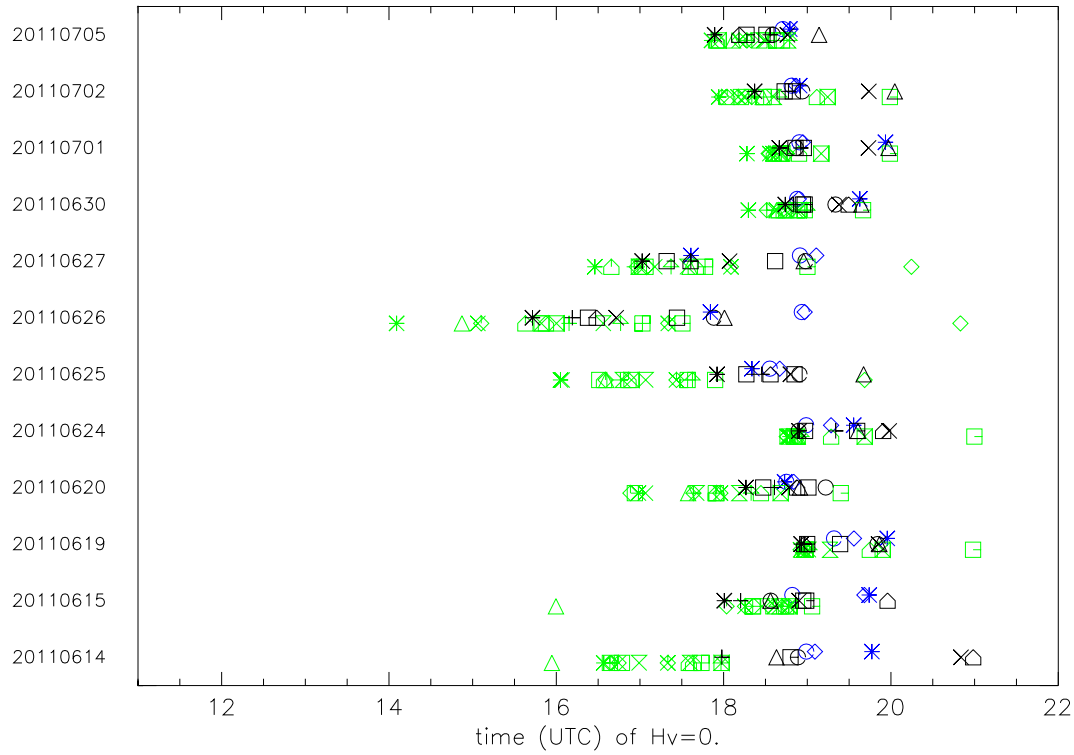
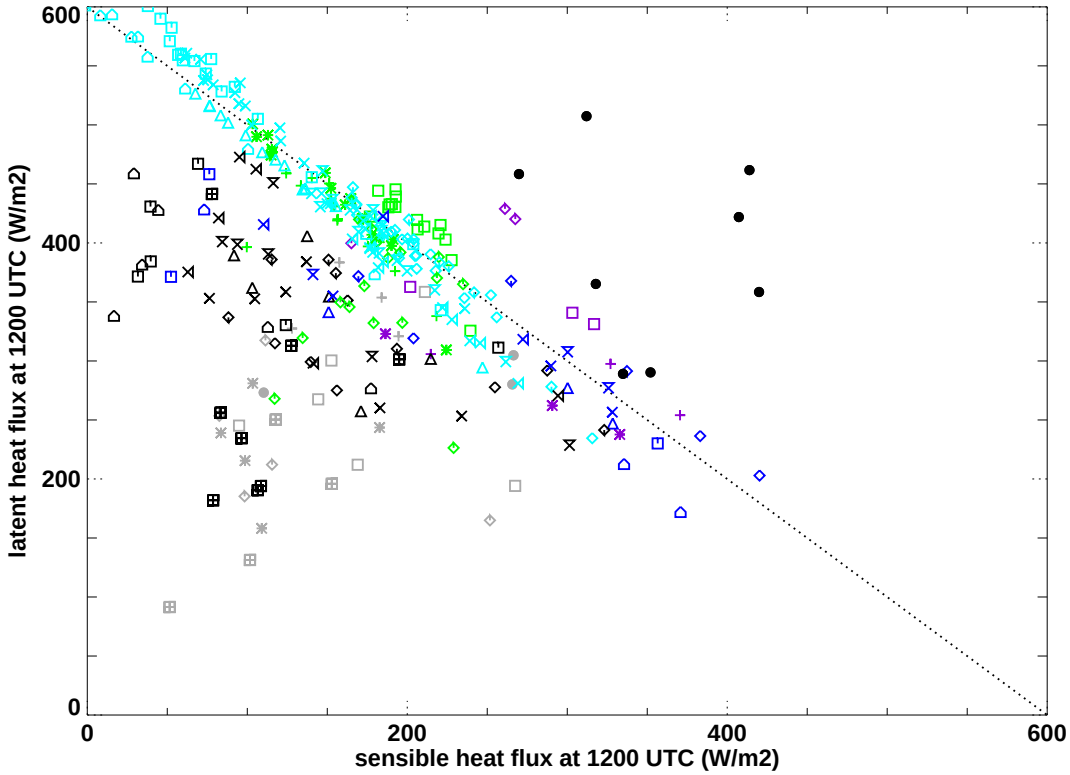


Figure 11: Time (on the x-axis) when the virtual temperature flux becomes negative for surface stations observations (black symbols), the ARPEGE grid-points (blue symbols) or the AROME grid-points (cyan symbols) for each IOP plotted on the y-axis.



Supplementary Figure 1: Latent heat flux versus Sensible heat flux at 1200 UTC in observations (in black for clear days and grey for cloudy days; the dots correspond to the observations over the forest, while the crossed squares correspond to the observations at 60m in the 60m-tower) and models (AROME in cyan for clear days and green for cloudy days and ARPEGE in blue for clear days and purple for cloudy days). One symbol is plotted for each IOP.