Answer to the reviewer 1 about the manuscript entitled : 'Boundary-layer turbulent processes and mesoscale variability represented by Numerical Weather Prediction models during the BLLAST campaign' by F. Couvreux et al.:

First, we wish to thank the reviewer for her careful and very detailed review. Below is our response (in blue) to the comments on a point-by-point basis. Reference to how we plan to modify the text is indicated in italic.

General comments :

I already made several in my overall quick review.

The answer to this quick review is attached at the end of this document.

However, I think it is very important to focus more on the impacts of the different grids, in terms of what model TKE is most comparable to observations, in terms of shadowing (and its impact on surface fluxes, especially in the evening and early morning), and in terms of resolved boundary layer structures, which can account for an important part of the TKE (w of order of 1 m/s in Ching et al 2014 MWR and LeMone et al. 2013 MWR).

We have look at the AROME forecasts and verified that no spurious convectively induced secondary circulations were present in those forecasts (horizontal maps of the temperature at different vertical levels are available on the BLLAST website: http://boc.sedoo.fr/nwp/lammodel/arome). Indeed, the effective resolution of the AROME model is around 9 Δx as shown in Ricard et al (2013). Note that simulations with Meso-NH, a research model, that has a smaller effective resolution, more on the order of 3-5 Δx do shown spurious circulations at 2km resolution. For ARPEGE and ECMWF, with horizontal resolution greater than 10 km, the boundary-layer structures are entirely parameterized. For any of those 3 models, the resolved vertical velocity is very small. So here, the resolved boundary layer structures are not an issue. However, we modify the text and now reference the above papers, to stress that in other situations resolved spurious boundary layer structures can be an issue.

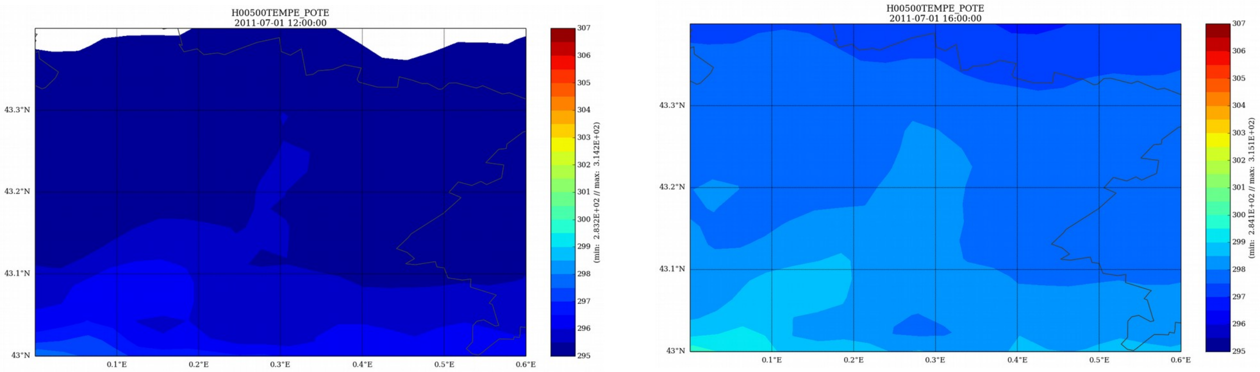


Figure 0: horizontal map of potential temperature at 1200 (left figure) and 1600 (right figure) for the 1st July 2011

Also, the impact of the different model terrain, particularly on heterogeneity at night. Acevedo and Fitzjarrald and LeMone et al. (2003, JAS) both show terrain plays a role in nighttime horizontal heterogeneity.

As you suggested, we have now included a figure showing the terrain represented in the different models as well as the real terrain. We also have included more discussion relative to the role of terrain on night heterogeneity (see response to detailed comments below).

I spend a lot of time writing what model variables might be directly comparable to the TKE measured in the atmosphere. This would be unambiguous if all PBL transport were proportional to the local gradient (i.e., don't need mass flux in the PBL schemes) and there are no resolved PBL eddies (possible with large horizontal grid spacing). It starts to get ambiguous when you have the resolved eddies (I'd just add their TKE to the subgrid TKE), and when you have mass flux in your EDMF schemes. What I don't know is whether the "MF" in the mass flux scheme is by TKE is completely separate from that in the “TKE” part of the scheme. In my comments, I assumed that it was, i.e., that the model TKE was the sum of the subgrid TKE + MF TKE + resolved-eddy TKE.

As said previously, there is no resolved vertical eddies with a 2.5km resolution in AROME. The mass-flux scheme is a more important issue that we partly discarded. The budget analysis of this contribution indicates that the mass-flux scheme provides a small contribution close to the surface, less than 10% of the total tke but a stronger one in the middle of the convective boundary layer where it reaches 20-25%. We therefore revised the comparison by including the mass-flux scheme contribution to the total tke and modified the text accordingly. The figures below present the time evolution of the total turbulent kinetic energy (subgrid turbulence scheme + mass-flux scheme + resolved eddies) at two different altitudes for two different days. We can clearly see that the resolved eddies contributions is null for the 16 different points (dash-dotted lines). The mass-flux scheme contribution is smaller than the subgrid turbulence scheme and accounts for around 10% of the total turbulent kinetic energy at 60m and around 20% in the middle of the boundary layer (illustrated here at 250m).

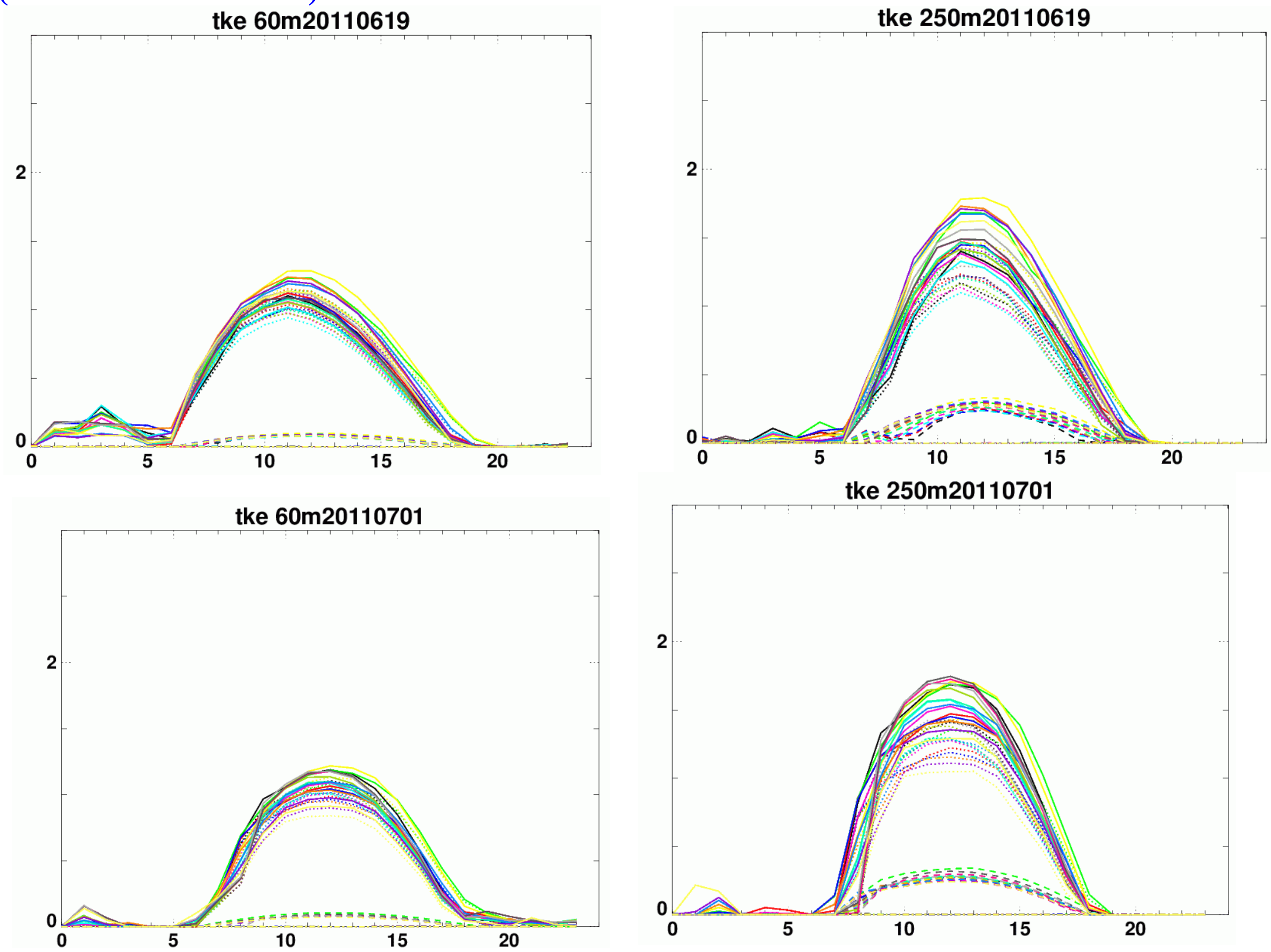


Figure 1: time evolution of the turbulent kinetic energy (total in full line, subgrid turbulent scheme contribution (dotted line), mass-flux scheme contribution (dashed line) and resolved eddies (dash-dotted line) for the 19 June 2011 on the upper panels and the 01 July 2011 on the lower panels and at 60m on the left panels and at 250m on the right panels

Specific Comments:

P1 L26. Should be 24-h forecasts

Done

P2 L1-2 (=P1 L22): Not sure what you mean here. Do you mean that there were more forests in the model or in reality? You could clarify by being more specific, for example, “related to identifying mixed forest and meadows as “forest” in two grid scales in the model.” (If there is too much forest in the model).

This is now better explained in the text. Here, we meant that in ARPEGE there is an overestimation of the sensible heat fluxes as if the grid boxes were entirely covered by forest while the analysis of the Land-Use map indicates that only part of the 10-km grid box (less than 25%) is covered by forest. We removed the words 'over-predominance of forest' from the abstract.

P2 L10-11. How about "The model reproduced the range of variables to within an order of magnitude." (This is more compact; you don't need to write that it was analyzed).
We have made the proposed change.

P2 L29. Don't need "the" before "Europe"
Done

P2 L29. It is interesting that this model has a warm bias in cold and stable conditions. Don't most models show cold biases under such circumstances?
In the study of Atlaskin and Vihma (2012), they tested four models, among which AROME and ECMWF. They showed a positive bias for the 2-m temperature under very low temperature (T<-10°C; a negative bias is observed for less cold temperature at night). We modified the text to make it clearer: '*They focused on the representation of very stable conditions at very low temperature (<-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.*'

P3, L3-4. LeMone et al. evaluated PBL schemes and their diagnositics.
We changed this sentence to '*LeMone et al (2003) used CASES-97 observations to evaluate boundary-layer schemes and their diagnostics based on mesoscale model simulations'.*

L25, work of Acevedo and Fitzjarrald. LeMone et al. (2003, JAS) showed from CASES-97 data and evaluation of results of earlier field programs that the timing of maximum horizontal variability depends on the scale of the terrain. This is because of how long it takes for drainage currents to flow from high points to low points. This timing has to be also affected by "frost hollows" that are sheltered from the wind. This makes LES of limited value unless it has a very large domain with very fine mesh. Just a comment; doesn't need a response. And it also implies an important role of model resolution.
We added a sentence in the text highlighting the importance of model resolution : '*This highlights the important role of fine resolution in order to get the right orography in the model.* '. There is also now a new figure that includes both the orography and the modelled terrain. See response to major comment n°2.

P4, L11-14. I don't understand this sentence, especially the use of the word "punctual," which means "on time", as in "She was punctual" -- she arrived just when we expected her to. Maybe you should just write that observations of TKE profiles, being made only during field campaigns, are quite rare.
Thanks for your suggestion we included the sentence following your proposition : «*For example, observations of tke profiles, being made only during field campaigns are quite rare, therefore the boundary layer parametrization based on a prognostic equation of the turbulent kinetic energy, which has been shown to perform better than first-order scheme (Holt and Raman, 1988), has only been evaluated via comparison with LES results (Cuxart et al, 2006 for instance)."*

P5, L14, L16. Could you describe "moor" in more detail? Is that a specific kind of vegetation or mix of vegetation?
Indeed, this is a specific kind of vegetation. Moor is an area of open wasteland, often overgrown with grass and heath. We have included this information after the first use of this term.

Figure 1. This figure is extremely hard to read. Need bigger range of color or lighter colors. And maybe larger size.

We have enlarged this figure and hope that now it is clearer.

P 6, L7. Suggest “unique aspect” rather than “specificity.”
Done

Section 2.2. Suggest details regarding numerical models in a brief table. This helps the reader refer back to model physics (especially the PBL schemes), grid spacing, run length, beginning of runs, etc.
This was already added in the second version of the submission material after your first quick review.

Table 2. Why not include vegetation type, rather than a lot of the detail here, since readers will know, for example, that “forest” has a larger roughness length and LAI than “grassland,” and “forest” has a lower albedo than “grassland.” And then you could include a column describing what you consider to be the land cover. Or, if not, at least you could refer back to the table when noting result mismatches due to mischaracterization of vegetation. Also, it would be instructive to including a four-frame figure showing the terrain contours for the three models and what it really looks like. A fussy comment: should be “grid points” not “grid-points.”
We have included the dominant vegetation type in a supplementary column of Figure 2 in order to help the reader's interpretation. However, we decided to keep the surface characteristics (albedo, roughness length, vegetation fraction,..) of the different points as this corresponds to the values that are used in the computation of the energy budget. For AROME and ARPEGE, they have been calculated taking into account the subgrid variability of the land use as explained in Giard and Bazile (2001). A four-frame figure showing the terrain contours for the three models and in the real world has now been included in the manuscript (new figure 2).
Throughout the text, 'grid-points' was changed into 'grid points'.

Section 2.2. Also, did you run the ECMWF model or download output? As to vertical grid points, you could put them in your profile figure to give the reader an idea of where they are.
We did not run the ECMWF model, it was run operationally by the European Center. We retrieved the model outputs from the ECMWF archive and analyzed them. There is no figure showing vertical profiles of ECMWF runs but the information concerning the vertical resolution is already included in Table 2.

Section 2.3
P 8, L8-10. Are you referring to Lothon et al? If so, refer to it.
We have modified the sentence to be more explicit : « *A large variability of surface fluxes exists among the sites (Fig 1) at scales smaller than 2.5x2.5 km², which corresponds to the size of a grid box in AROME (see for example the differences between the moor and the corn sites, or the grass and the wheat sites) that are mainly due to surface cover; this was also shown in Lothon et al (2014)*”

P8 L11. “Clues as to” rather than “inferences on”?
Done

Need to give conversion from UTC to local time, which is what drives PBL development. At this location (Lannemezan, France; lon=0.38°E) the longitude is very close to the 0° Greenwich meridian. So, UTC time is very close (~2min) to solar time. However, in France the local time is postponed so that 1400 LT is equal to 1200 UTC time and 1200 solar time. So we have kept the UTC time in the paper but we also indicated that this is very close to solar time. We have included the following text:
'… note here that UTC time is the same as solar time as very close to the Greenwich meridian'

P8, L26. Replace “an” with “a vertical interpolation”
Done

P8 L30. Which model? All three? Also – why don’t you try using some of your observational criteria on the model profiles? In some sense, you are often comparing apples and oranges rather than apples to apples, since different criteria can give different PBL heights. (See LeMone et al. 2013 – we very rapidly abandoned the diagnosed values because they were often inconsistent with the model theta profiles).
Here we only analysed ARPEGE and AROME models as the ECMWF finer available time sampling (3 hours) was too coarse to investigate the temporal evolution. It is not always straightforward to use the same boundary-layer diagnostics for observations and models. Indeed, in observations we use different types of diagnostics derived either from the UHF (two different diagnostics), from an aerosol lidar (one diagnostic) or from thermodynamical profiles (four diagnostics). As you suggested, we applied to the models the diagnostics based on thermodynamical profiles and we now state in the text the results of the comparison of those diagnostics to the model diagnostic (based on *tke*). However, during the afternoon transition, the diagnostics based on thermodynamical vertical profiles sometimes depicts the top of the residual layer rather than the top of the still convectively active shallower layer. In the figures below (illustrated for four IOP days), we compare the diagnostic computed online based on the vertical profiles of the turbulent kinetic energy in black/grey for AROME/ARPEGE with the diagnostic based on the virtual potential temperature in green/blue for AROME/ARPEGE.

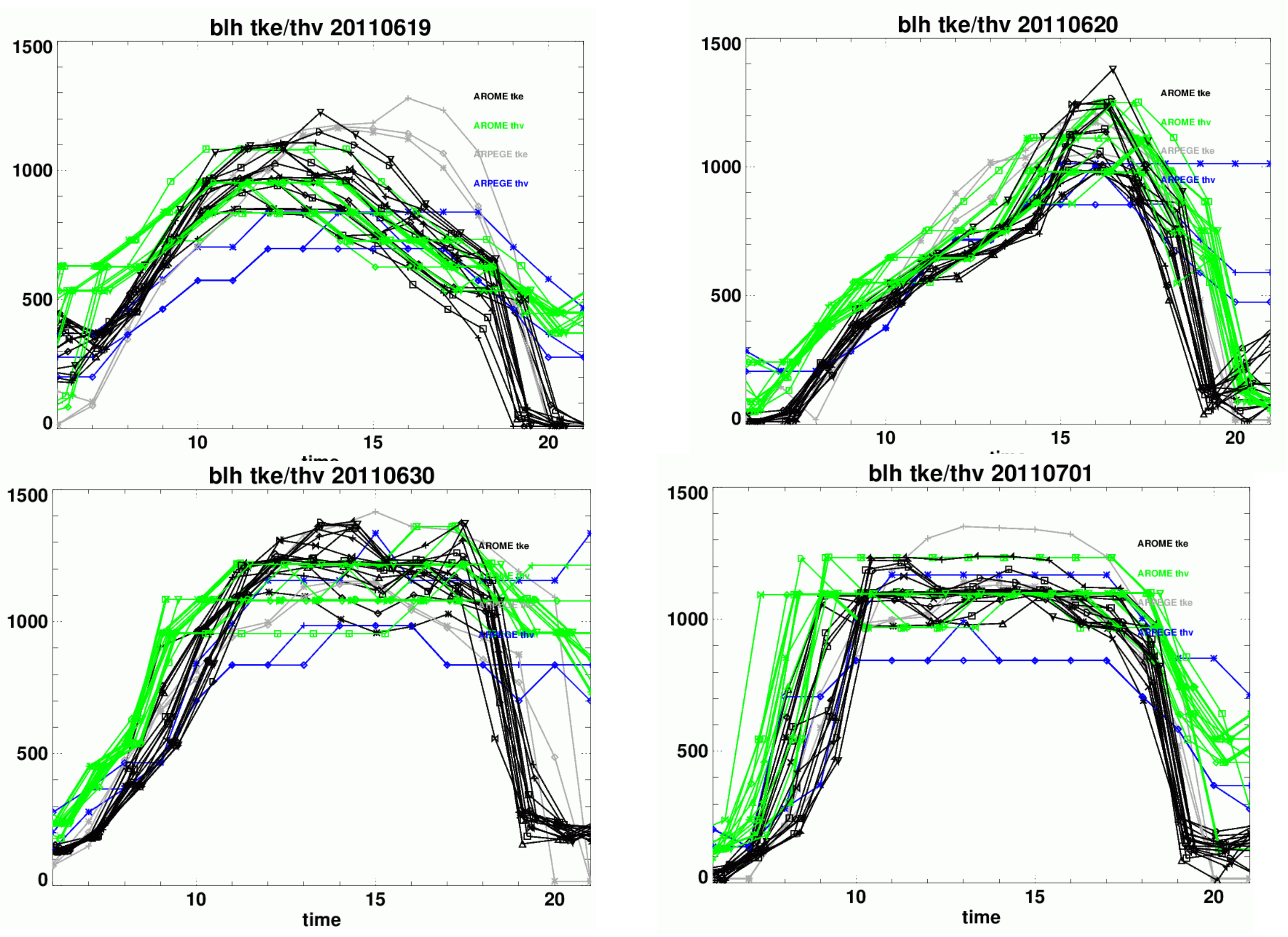


Figure 2: time evolution of boundary-layer height diagnosed by the model (based on tke) for AROME (black) and ARPEGE (grey) or diagnosed from the vertical profile of the virtual potential temperature for AROME (green) and ARPEGE (blue) for the 16 points.

There is consistency between both diagnostics for most of the models with however some discrepancy for some times (in particular during the afternoon transition). We therefore decided to keep the model diagnostics (discarding however the time where it is not relevant due to the presence of shallow clouds, this diagnostic depicts the top of the shallow clouds : two hours for the 15 June) as well as time where strong shear induces a decoupling between the boundary-layer and the tke profiles (morning of the 27 June) as illustrated in LeMone et al (2013) for the shear case. Eventually, also not that with observations we derive different diagnostics with the idea to analyse what each diagnostic depicts in particular during the transition.

P9, L4. Can delete “previous”
Done
As noted in earlier general comments, a look at the paper by Lindsey Bennett et al. (MWR, 2010) might be helpful.
After your first quick review, we included a reference to this paper in this section. This paper is quoted twice in page 6 :'A *comparison of different boundary-layer depths derived from various instruments has been presented in Bennett et al (2010).*' and '*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).'*
P9, L23-5. Why not apply the different criteria to the model profiles to see how they relate within the model? (I.e., different criteria give different PBL depths).
See response to comment (P8 L30) above.

P10, L3-13. Evaluation of TKE in the PBL is hard; and comparing it to TKE in the model is even more challenging. Averages are probably too short; and aircraft high-pass filtering eliminates important scales. See Grossman et al. (1992) and Kelly et al. (1992), both in J. Geophys. Research for flux profiles. In the CBL, you should expect to see large eddies of scale of the order of 1.5-3 times the depth of the CBL; a 5-km cutoff will diminish these eddies significantly. In fact, use of such a short averaging time (and cutoff) is not consistent with the 30-min averages for surface fluxes, which are designed to capture all the fluxes.
The 5-km cutoff is what is usually used in the program computing fluxes from the high-frequency aircraft data. We analysed the sensitivity of turbulent fluxes to the choice of this cutoff length for BLLAST and other field campaigns (AMMA & HYMEX) and found that increasing this cutoff length did not strongly modified the fluxes estimations. However, as expected the computation of the variance is decreased by the use of the 5-km cutoff as illustrated in the figure below, and this effect is stronger for the variance of the horizontal wind compared to the variance of the vertical wind.

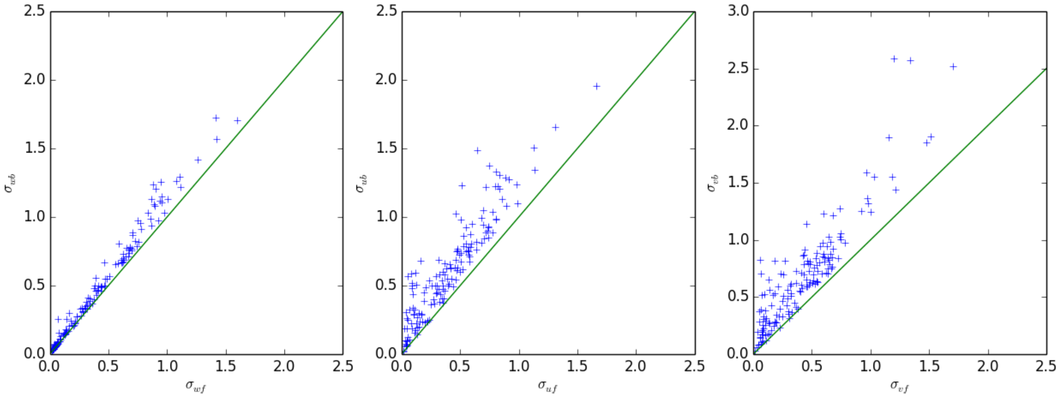


Figure 3: Comparison of the variance computed from filtered signal (x-axis) or raw data (y-axis) for (left figure) vertical velocity variance, (middle figure) zonal wind variance and (right figure) meridional wind variance.
For the turbulent kinetic energy, the 5-km cutoff induced a reduction of 20-22% as shown in the figure below:

This is now commented in the text as :
'...; this is the current treatment used for flux computation, it however induces an underestimation of the tke of about 20%'
The wind during BLLAST is relatively weak, typically from 1 to 3 m/s so a 30 min average correspond to 30 min ~ 2-5 km and is therefore consistent with a 5km cutoff length. Eventually, during BLLAST, the boundary-layer height was usually around 1 km so the scale of the large eddies should be broadly resolved with such measurements. The segments used to compute the turbulent kinetic energy used for comparison to the models are on average 31km-long and last for 7.5 min (450 s).

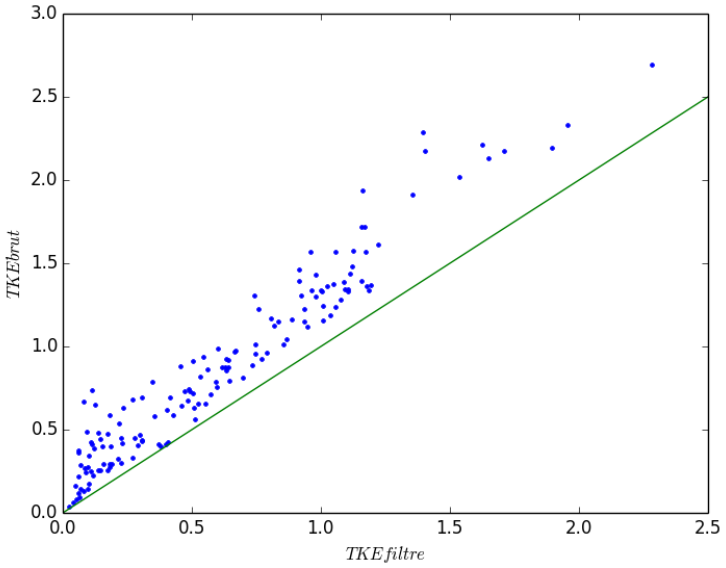


Figure 4 :Comparison of the turbulent kinetic energy computed from filtered signal (x-axis) or raw data (y-axis).

Of course this is for capturing the total TKE. For 16-km horizontal grid spacing, this would represent the TKE from the PBL scheme, plus the TKE associated with the parameterized mass flux. For 2.5-km grid spacing, this would be the TKE from the PBL scheme plus the TKE associated with mass flux, plus the TKE associated with partially-resolved eddies.
It's not surprising that you get larger measured TKE than model TKE simply because you don't include the "MF" part (which is only w). Perhaps one meaningful comparison could be made at mid-PBL when the horizontal TKE is smallest and vertical TKE the largest.
Or you could derive a rough representation of mass flux in the PBL scheme by developing an empirical relationship between mass flux and TKE from the aircraft data (for larger scales). Not sure this would work – the relationship between w in TKE and w in mass flux in EDMF schemes is not clear to me.
At 2.5km, we checked that there were no resolved eddied and we have added the contribution of the mass-flux scheme which is negligible close to the surface but more significant in the middle of the boundary layer (see also response to general comment).
A really tough but useful test (but more doable than TKE from the model point of view) would be to compare the moisture flux from the model and from the observations, since the model uses total flux divergences (at least for the two coarse-grid models; you would need to add flux by the resolved eddies in the 2.5-km grid spacing model). Heat flux also – but that is so tightly constrained that it doesn't give you as much information. (We tried this in Tastula et al. QJRMS 2015 or 2016).
It is reassuring that your TKE is typically larger than the model TKE – that is what one would expect, given the above discussion. I'd expect the discrepancy to be even larger if you used averages that included the larger scales.
Unfortunately the moisture flux was not an output of the model so this is a really tougher test that we decided not to carry out.

P11, L 19-20. Ambiguous. I thought these four days had no clouds or a few clouds (and by implication, the other days had more clouds). Suggest rewording, perhaps like this.
Those days correspond to mainly high-pressure fair-weather conditions with no cloud cover, or, for 14, 15, 24, and 30 June, a small amount.
Done, we included your suggestion.

In Figure 2 caption, don't use "range" since range means maximum minus minimum. Suggest "black curve with horizontal standard deviations indicated by error bars" instead of "black curves … shaded in grey" since you already have nighttime shaded in gray and this avoids the use of the word "range"). I put in "horizontal" since I think that is what you mean. Also, you should replace "variability" with "range," which is the correct label – and you have room for a bigger font, which is important. Print is very marginal in size for readability. Finally, you need to explain the dashed lines in the figure (they are explained in Figure 3). Also, if you label your points in Table 1 with land use, it would help interpretation here as well as in the text.
Eventually, following reviewer 2, we decided to simplify this figure and only showed the mean curves. However, to be able to illustrate the over-estimation of two grid points of ARPEGE we have added a figure (now Figure 3) showing the comparison of sensible and latent heat fluxes between all the observation sites and the different points of ARPEGE. In this caption, we have used your proposition ('*black curve for the mean with horizontal standard deviations indicated by error bars*'). 'Range' was already replaced 'variability' with a bigger font following your early review.

P12 L1. Similarly, you don't need the "gray shading that indicates the envelope containing the different surface sites," since (a) it's described in the caption, (b) the "gray shading" is confusing, since the error bars look black on the graph and the gray shows night, and (c) an "envelope" typically describes the range (maximum – minimum).
Now Figure 2 only shows the mean and the range. Figure 3 presents the horizontal variability but vertical error bars are plotted and the gray shading has been removed.

P12 L3. What does "for a given type" mean? Don't you mean for a given day?
We have modified the sentence to "*this is computed at each time step by the difference between the maximum and the minimum over all the points of either one model or the observations*"

P12 L3. This is correct use of "range." So why not use that instead of "variability" on the right side of figure 2. This word is also shorter, so you can make the letters bigger. (I can't read it easily unless I enlarge the electronic version)
Following your first quick review, we already changed 'variability' into 'range' in the last submitted version. This has been changed for the 3 figures.

P12 L4. Suggest (no C in Figure) after "cloudy days" since you do not explain what the C means (I thought it meant "cloudy"!). Maybe – if possible, you could include a circle with cloud fraction instead of the C. That way it would be less ambiguous (since both "cloudy" and "clear" start with C.
We agree that the 'C' was ambiguous. It has been changed by an empty circle for cloud-free days and a grey triangle for cloudy days. We did not have quantitative observations of cloud fraction so we could not include this information.

P12 L7. Either "clear" or "cloud-free" but not "clear-free" Figure 3.
We have changed clear-free to clear in the text.
You should repeat the labels on the plot that you put in Fig. 2; also replace "variability" with a "range" in a larger font. Also label the "hot" days, since you discuss them.
Following your first quick review, we already labeled the 'hot days' in Figures 2, 3 and 4**.**

P12, L23-4 “which has similar range as observations above the forest”. I am not sure what you mean by this. When you say range, are you referring to range in time, since there is only one curve? If you are referring to the difference between two forest sites in model and observations, should point out that they are not shown in the graph. Again, a vegetation type label would be useful.
In fact, we wanted to state that the values predicted by ARPEGE for the high-vegetation grid points are of the same order of magnitude as the observations above the forest. However, these simulated sensible heat fluxes are too large to be representative of a 10km wide grid box over an area which is characterized by much more surface heterogeneity and is far from being entirely covered by forest (cf Fig 1).We changed the text to :
'...However, these simulated sensible heat fluxes are too large to be representative of a 10km wide grid box over an area which cannot be characterized, according to Figure 1, by a uniform forest cover; indeed, there is a large variability of surface covers at scales below 10km '*… 'For two ARPEGE points the surface fluxes are similar to measurements over forest, but the satellite data does not indicate a homogeneous forest patch over 10x10km² in this 10x10km² area. '*

P12. L29-30. I THINK you are saying that the model assumes more trees in the grid box than there actually is. That is not captured by “much more surface heterogeneities at this size.” Also, reference to Fig. 1 doesn’t help since you really can’t see much (it might if you improve the figure). If you put surface type in the table, this would help. And perhaps label the points in the figure that you discuss in the text. (I.e., you don’t have to label all of them).
This was not clear and we modified the text as '*However, these simulated sensible heat fluxes are too large to be representative of a 10km wide grid box over an area which cannot be characterized, according to Figure 1, by a uniform forest cover; indeed, there is a large variability of surface covers at scales below 10km.*' The points which are referred to in the text are now labelled in Figure 1. Figure 1 has also been enlarged.

P12 L31. The only gray shading I see is the nighttime.
According to your previous comment, we change the gray envelope into error bars so now there is indeed only gray shading for nighttime.

P13 L3. Again, please label the hot days somehow on Figure 3.
Following your first quick review, we already labeled the 'hot days' in Figures 2, 3 and 4**.**

P13 L17. Have you looked into the “coupling constant”? I.e., the coefficient in the bulk formula used to calculate flux? We have found it sometimes to be off in the model when compared to the observed value. This could account for both latent and sensible heat flux being too high, since the solar radiation doesn’t look that far off.
Ideally to more fully explore the surface energy budget, we should look at the G component as well, but it was not available in the models. In fact, in the different models the coefficient used in the bulk formula used to calculate the flux is not constant but is computed iteratively and is a function of stability so it is tough to look at this 'coupling constant' and we did not do it.

P13 L18. What is “high vegetation”?
ARPEGE uses a criterion to separate 'high vegetation' from 'low vegetation' in term of stomatical resistance and roughness length. However, this information is not really necessary here. We have removed the term 'high vegetation' in the text.

P13 L24. This is a new thought, so should start a new paragraph.
Thanks for the comment. We started a new paragraph.

I noted in my earlier set of comments the citation to LeMone et al., which you appear to have in the references but not obviously in the text. As noted previously, a negative slope in the plot means a

constant available energy, not a constant Bowen ratio. For a constant Bowen ratio:
Bowen ratio = B = SH/LH. If it were constant, LH = SH/B, which would mean that the slope would be positive, not negative.
After your quick first review, this reference has been included in the text (cf P9 l30: 'Interestingly, when plotting the latent heat fluxes as a function of 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).')

P14, L17-19. We found that wind reduced horizontal variability during the night for CASES-97 in LeMone et al. (2003). I would guess Acevedo and Fitzjarrald did as well for their data; because the BL remains coupled to the ground. In strong winds, we found theta almost constant at night. Curious that the model didn't – but then you wouldn't get as much terrain-induced variability with the coarser-grid models. (Again, would be nice to see what the terrain looks like with the coarse-grid models).
Concerning the horizontal variability at night we have added these two references: '*The spatial variability in night time temperature among sites is smaller for the hot period; this is probably be due to larger wind speed during this period (as shown in LeMone et al 2003 and Acevedo and Fitzjarrald 2001).*'
We have also included a figure showing the terrain in models and observations (cf answer to the second general comment and new figure 2).

Figure 4 caption: should note what the double vertical lines are. Did the rain occur at the same time every day, as the figure implies? Regarding diurnal cycles for mixing ratio (bottom, P 14). It does look as though you get the morning and evening maxima at least at some sites (associated with large latent heat flux into a shallow BL). This is a good marker for the creation of the shallow PBL in the evening locally. If the terrain is complex, perhaps this happens at different times at different sites. This feature is strongest for weak winds and strong LH.
The double vertical dotted lines indicate interruptions in the days as only the IOP days are plotted and not all the days from 14 June to 2nd of July. This is now added in the caption of new figures 2 and 4. There is no explicit mention of the time of the rain, the rain often occurs at night but not always at the same time. This is also mentioned in the text as :' *the days with precipitation were not IOPs and corresponds therefore to an interruption of time in Figure 4, indicated by the double vertical dotted lines*'.
We have also included a comment regarding the morning and evening maxima: '*Often, observations indicate a morning and evening maxima (e.g. 19 June, 30 June, 01 July, 02 July) associated with large latent heat flux into a shallow boundary layer; this is correctly simulated by the models.*' However, the relationship with the intensity of winds and surface latent heat fluxes is not so obvious.

End, section 3.2 – yes, mixing ratio is the most difficult!
Figure 5. Regarding warm and cold biases in the lowest 500 m for the models. Have you factored in differences in PBL depth? For example, if the PBL depth were underestimated by the models, the mixing ratio would be greater. (Of course, horizontal advection – and initial conditions – could also have an effect).
For AROME and ARPEGE, there is no obvious biases in terms of the PBL depth. Concerning the ECMWF dry bias, we have checked that it is not related to a too high PBL depth either.

Figure 6. Suggest taking advantage of this figure to show where the lowest grid points are. One could do this by putting points on one profile for each of the models, or you could mark grid points in a three columns within one of the frames – (top right figure would be excellent for this).
As you proposed, the vertical grid of the models is indicated by crosses in this figure.

For 27 June, I am intrigued by the large horizontal variability even though the skies are clear. Do you have resolved PBL eddies?

For sure, there is no resolved PBL eddies in the ARPEGE simulation due to the coarse 10km-resolution. We have also checked in the AROME simulation and AROME does not either present resolved PBL eddies (see also response to first general comment). In addition, ARPEGE, as AROME, shows a large horizontal variability for this day. This large variability seems to be related to the synoptic conditions as the 25 June (during the hot period) large wind may have prevented the establishment of the mountain-plain circulation or at least delays it.

These can affect surface fluxes, and especially humidity and wind (also temperature, depending on PBL scheme). On strong-wind days you could be getting model rolls as well as observed rolls, which are associated with strong horizontal changes (see e.g., Weckwerth et al. MWR, 1996). Also Ching et al. (2014, MWR) and references therein.

Thanks for this reference, but we have checked and there is no convectively induced secondary circulations during this day.

P16L11. “Mesoscale circulation” very vague. It would be good to give a scale and perhaps a likely cause. Do you mean terrain-induced circulations? Or something larger in scale?

We have added a scale. In fact, carrying a simulation with the same physics as ARPEGE but at 2.5 km horizontal resolution also reproduces the maximum in the upper part of the boundary layer as shown in the figure below. This indicates that this feature is related to fine scale advection not resolved with a 10-km grid : '*Analysis of the moisture budget indicates that this maximum is mainly related to fine scale advection not resolved at 10 km (not shown).*'

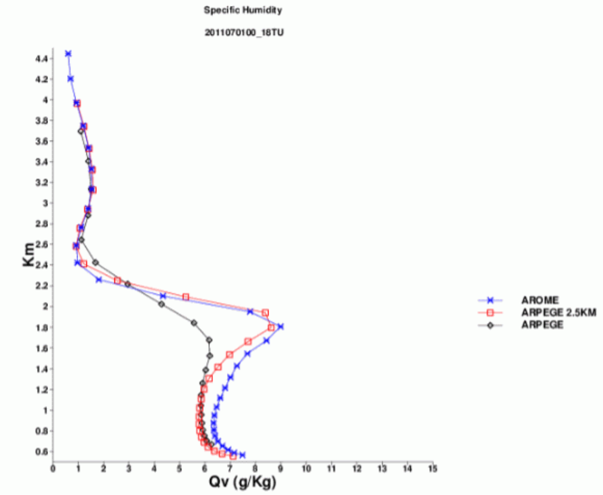


Figure 5: vertical profiles of specific humidity the 1st of July simulated by AROME (2.5km resolution, in blue), ARPEGE (10 km resolution in black) and ARPEGE physics using the 2.5km dynamics of AROME (in red).

In Fig. 7, label the hot days.

The hot days are now labeled in this figure and also the previous figures.

It would be more meaningful to compare similarly-diagnosed PBL depths from observations and model. And to compare differently diagnosed PBLs internal to the model and internal to the observations.

First, concerning the evaluation of the boundary layer we have added a reference to the paper of LeMone et al (2013): '*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).*' Concerning the PBL depth diagnostics as explained in the answer to P8 L30 we decided to keep the tke based diagnostic due to the better behaviour during the afternoon transition.

P16, bottom. Have you an idea what causes different types of morning PBL growth? Subsidence? Strength of inversion? And as you note, different criteria can give you different PBL depths. One you might mention is RH max, which will give you the top of the PBL in the absence of cumulus clouds, but will give you cloud base in the presence of cumulus clouds. (Curiously, we have found that PBL turbulence statistics scale well with cloud base, but it can be argued that the true PBL depth is somewhere in the cumulus cloud layer.)

The rapid growth in the morning is due to presence of a residual layer that remained close to neutral as for instance for the 1 July (cf Lothon et al (2014); Blay-Carreras et al (2014)). The days with limited growth have smaller sensible heat fluxes and experience subsidence of warm air that made it very difficult for the CBL to grow. *We have now stated that 'The causes for different types of morning PBL growth is related to initial profiles, intensity of the sensible heat fluxes and of the subsidence as explained in Lothon et al (2014).'*

P17, L1. Isn't the top of the stable layer and the top of the inversion layer the same thing?

We change 'top of the inversion layer' to 'top of the residual layer'. The idea, here, is that the measurement during the afternoon can either detect the top of the stable boundary layer that is forming or the top of the residual layer corresponding to the trace of the convective boundary layer of that day.

P17, L7. Better prediction for AROME makes sense to me if you include shading, which could be a factor in decreasing surface buoyancy flux, especially near sunset. Do you?

I don't exactly understand what you meant by shading. Concerning shading by the orography, Orographic shading is now included in the operational AROME version, but it was not in the 2011 version used during BLLAST. But as the Pyrénées are mainly oriented East-West we do not expect a strong impact in the late afternoon as the sun set westward which will induce very small shading. Concerning shading by the vegetation, this is not directly included in the code, but accounted for via the albedo. Concerning the shading by the clouds, this is included in the radiation code through the cloud fraction for each grid, similarly as in ARPEGE.

P17 L16. "unique feature" rather than "specificity."

Done

Figure 8. Bigger font on right side. I can barely read the labels in my printed version – I'm working off my computer screen.

This figure has been redrawn and the labels have been enlarged. Sorry for that.

P17L30-31. Measured ON the evening … and IS reproduced.

Done

P 18. How do model and observed TKE compare if you count the TKE associated with resolved PBL eddies in AROME? And including some representation of the mass flux associated with the PBL scheme?

P18 L7, L14-15 While the difference in height might be a factor (at the lowest level), you would expect more parameterized TKE in ARPEGE compared to AROME because the PBL scheme in ARPEGE has to do almost all the transport (because the resolved eddies grow much more slowly for 10-km grid than for 2.5-km grid), see Ching et al. 2014, also LeMone et al. 2013). If it's an EDMF scheme, it should account for all the TKE. (Again one has to include somehow the mass flux in the TKE estimate).

P18, L19-20. Resolved PBL eddies grow during the day until saturation is reached. It could be that horizontally-averaged model TKE starts to go down as the resolved eddies grow. Though I am obviously skeptical that you will even achieve exact agreement, it is encouraging that the trends are similar.

There are no resolved PBL eddies in AROME. This might be due to the fact that the effective resolution is ~9 Dx (Ricard et al, 2013) which is 23.5 km (as also said to the answer of the general comment). The mass flux contribution is significant in the middle of the boundary layer and is now accounted for in the analysis.

P19L5. “physical processes … are small “.. You mean terms in the TKE equation are small? If you don’t want to write out the equation, you could write something like “Most of the terms in the TKE equation, -- buoyancy production, shear production, dissipation – are small.” ? Or are all the terms small?
You are right. We modified the text with your suggestion: '*Most of the terms in the TKE equation-buoyancy production, shear production, dissipation, vertical transport- are small (Nilsson et al, 2016).*'

P19L15. “where the height of the reflectivity gradient decreases with time …”?
We have changed the sentence to ' *where the height of the reflectivity gradient decreases with time in the evening*'

P19L25-6. This makes sense, since dissipation and TKE are closely related.
We have added this comment in the text: *'… , which makes sense as tke and dissipation rate are closely related.'*

P20 paragraph 1. The earlier time at which sensible heat flux goes negative at the surface is consistent with large latent heat fluxes. This makes sense both from the point of view that more of the total energy is going into LH. But it also means that the buoyancy flux remains positive after the sensible heat flux goes negative. I would guess that the time when the buoyancy flux goes negative is also earlier for AROME, and this would be more directly related to turbulence generation than sensible heat flux. It would be good to see what a plot similar to Figure 10 for buoyancy flux looks like.
After your first quick review, we have modified the Figure 10 to show the time when the buoyancy flux becomes negative. Indeed, for most of the cases this delays the time of about 5-15 min; however the meaning of the figure is unchanged. Figure 10 of the revised manuscript is computed with the buoyancy flux as the buoyancy (and not the sensible heat flux) is the term that controls the intensity of turbulence (also see response to point 4 of your previous quick review showing both figures).

P20 paragraph 2.Because of the large latent heat fluxes, it might be useful to normalize thing in terms of buoyancy flux rather than sensible heat flux.
I am sorry but I did not understand what should be normalized by buoyancy flux. As explained above the time of the end of the transition, namely the time at which heat flux goes negative is now based on buoyancy flux instead of sensible heat flux.

P20L32-P21L1, suggest .. Models and observations produce lower sensible heat fluxes, higher temperatures, stronger winds, and weaker TKE than (what? For the other days?).
Following your advice we have changed the text to '*For instance, during the hot period, models and observations produce lower sensible heat fluxes, higher temperature, stronger winds, and weaker tke than during the other days.*'

P21 L22-3. From P7, L12-14, I thought that ARPEGE had the same PBL scheme as AROME. This sentence implies there is no “MF” in the ARPEGE PBL scheme. This could be clarified by listing the PBL physics schemes in a table and describing them more carefully. If there is no “MF” in the ARPEGE scheme, then the TKE should be pretty comparable to the total (no high-pass filtering) TKE. (Though I would expect some discrepancy since pure TKE schemes don’t really represent what is going on in the CBL). Please clarify. As noted earlier, the parameterized TKE in an EDMF scheme should be smaller than measured, particularly for fine-mesh model runs (because of the contribution of resolved eddies to the TKE).

This was not clear enough in the manuscript. ARPEGE and AROME do have the same *tke* scheme (Cuxart et al, 2000). However, in AROME, there is also a mass-flux scheme, based on the eddy-diffusivity mass-flux concept (Pergaud et al, 2009). In ARPEGE, there is only a mass-flux scheme active when shallow cumulus are present so this scheme is not active for clear days. There is already a table (table 2) that describes the different parameterizations; it was added after your first short review. As said previously there is no contribution of resolved eddies to the *tke* in AROME.

P21, end of 2nd paragraph. Estimation of some terms in the TKE budget might be simpler than the estimation of the TKE, at least in terms of direct comparison of model with observations, for reasons discussed earlier.
We agree with you. This is what we propose as a next step for this study. This is complicate to handle as all the runs have to be redone as the different TKE budget terms were not saved.

Supplementary Figure 2. Is this discussed?
This figure was discussed in the appendix but we decided to suppress the appendix.

P10, bottom. You refer to an Appendix here (which isn't part of the paper). Perhaps you should just refer to supplementary figure 2? Was there an appendix?
It seems that you may have had access to an earlier version of the text, possibly the one initially submitted before your earlier review. But after your quick review, we provided a new document. The appendix was added during this first step of revision but eventually we decided to suppress it.

References:

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

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

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

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

Responses to Earlier Review:

General comments

The paper is interesting and I think will be in a publishable form with some
modifications. I include here only some major thoughts (in no particular order).

1. The SH-vs-LH graphs having a slope close to -1 doesn't indicate a constant Bowen ratio- quite the opposite, it shows horizontal variation in Bowen ratio. Rather, it shows a constant available energy (i.e. LH+SH=constant). This is discussed for CASES-97 in LeMone et al 2003 (J Hydromet, choosing the averaging interval) and discussed as a function of soil moisture using both observations and a land-surface model in LeMone et al (2007). It is nice to see someone exploring this.
You are right. This was a mistake and has been changed in the text to ' the models reproduce the -1 slope related to an almost constant available energy (cf Supplementary Fig 1) in agreement with LeMone et al (2003).' We now quote the reference to LeMone et al 2003 that inspired us for drawing such graph.

2. When discussing horizontal heterogeneity, terrain plays a big role. It would be good for the

authors to show maps of the terrain used in the three NWP models.

a. This is true, as the authors recognize, because of the presence of mountain-valley circulations of tall types. The different terrains will produce different circulations

b. This is also true for horizontal variability. Although one gets downslope drainage winds event with gentle terrain, more extreme terrain probably has more cold-air pooling. So, there might be less horizontal variability smoothed terrain.

In Table 3, the altitude of each point is indicated, which provides information on the resolved orography in the models for the points used for the intercomparison. According to your comment, we have included below the map of the resolved orography on a small domain around the observation sites for the ECMWF model as well as the AROME model, those two models having the coarsest and the finest horizontal resolution. Indeed, the orography is better resolved in the finer resolution. The impact of the mountain-valley circulation is the subject of an ongoing study.

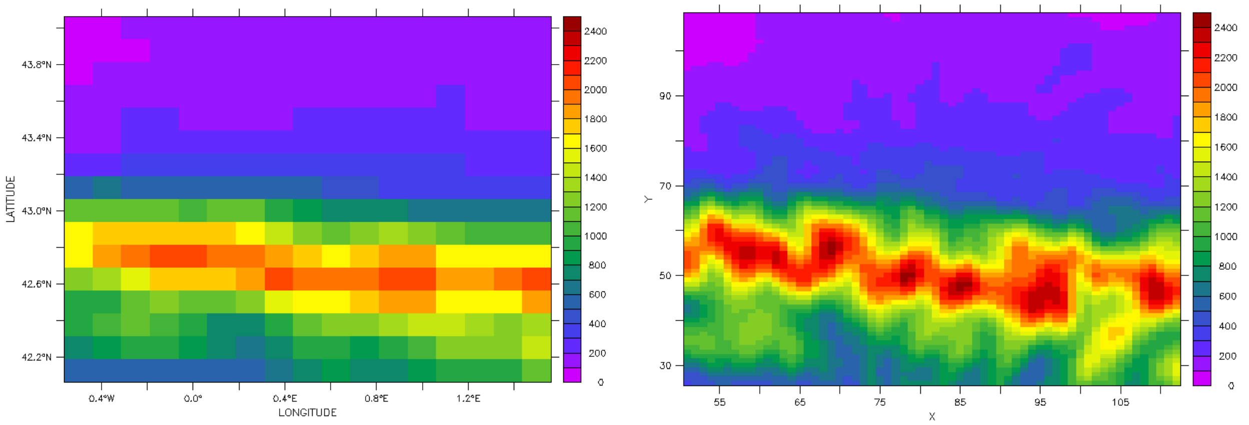


Figure a: orography of the models (in m) in (left) ECMWF operational model and (right) AROME operational model.

3. When discussing TKE, the measurements will inevitably include the impact of large eddies (horizontal wavelength between 1.5 and 3 times the depth of the PBL roughly). These are likely partially resolved in AROME and they tend to grow in models under convective conditions, faster with smaller grid spacing. Comparison to TKE in the other two models is in some sense more realistic from this point of view, since neraly all TKE will be parameterized. For fine grid spacing (< ~4km), the interaction between this resolved eddies and PBL schemes can exaggerate local variations in TKE (see Ching et al. MWR, 2014 and references therein).

Other issues :

a. TKE is mostly horizontal near the surface, especially for eddies extending through the PBL (for which w is very small ; from mass-continuity equation).

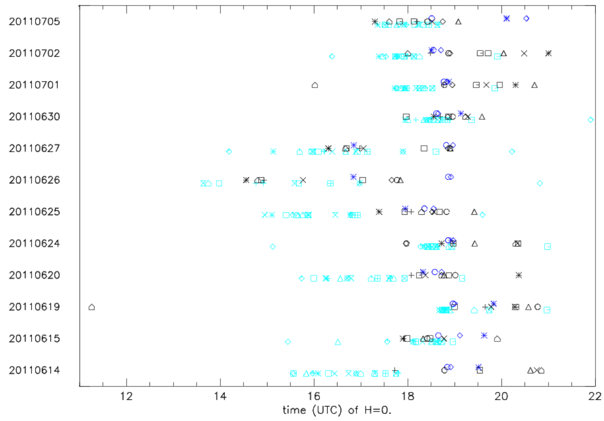
b. Large eddies travel roughly at the mean speed of the wind through their depth (i.e. the boundary layer). Thus if one filters according to scale, the scale should be defined not be the wind at the level of the measurement, but by the mean PBL wind.

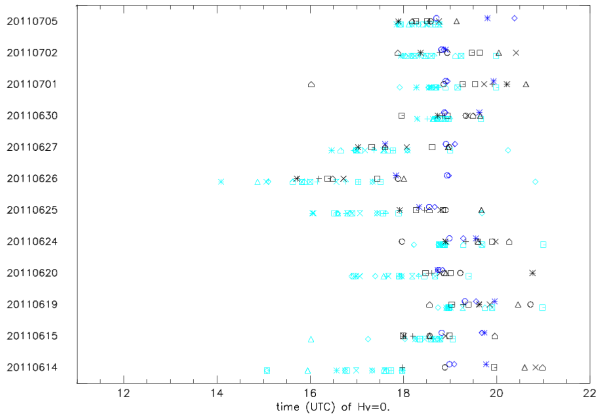
c. A philosophical point (discussed in Ching et al) is the in the 'gray zone' or 'terra incognita' the PBL scheme should account for all the TKE in the PBL, which for fine-grid models mean several grid points horizontally, and there should be no large PBL eddies (convective rolls or cells). (This is the purists' view ; the semi-resolved eddies have been useful in storm initiation or propagation – because large eddies, especially rolls, have been shown to play a role in storm propagation and evolution). One way to look at this is by considering the buoyancy-flux profile. It should be continuous from the surface (where its value is determined by a land-surface model) up through the PBL. If one does time- or space- filtering that is too fine, the fluxes above the surface are too small. I gather from the discussion that the authors were wrestling with this.

Thanks for the reference. The 'gray zone' is indeed an issue for a model at kilometric scale as shown in Honnert et al (2011). However, here, the runs performed with the finest resolution (AROME model) have a horizontal grid of 2.5km and an effective resolution of about 9Dx(~22km, cf Ricard et al, 2013). Therefore this is not an issue for this model and the turbulence is still fully subgrid. We have checked the buoyancy-flux profile as you proposed and also checked in horizontal cross section that no spurious numerical convective rolls or cells occurred. Note that, in AROME, the boundary-layer turbulence is handled by a EDMF scheme with the ED component being represented with a prognostic turbulent kinetic energy scheme and the MF component being handled by a mass-flux scheme which introduces a non-local contribution as advised by Ching et al. Not all the turbulence is handled by the turbulent kinetic energy scheme, part of the turbulence (the non-local thermals) is handled by the mass-flux scheme, this is what we discuss in the paper.

4. The authors should look at the paper by Lindsey Bennett et al. (MWR, 2010) regarding estimates from different instrumentation, in addition to the LeMone et al, and Grimsdell and Angevine work cited.
We have added a reference to this paper in the manuscript relative to the various estimates of boundary-layer depth in page 6 of the new manuscript: *'The comparison of different boundary-layer depths derived from various instruments has been illustrated in Bennett et al (2010).'* and in page 7 of the new manuscript '*The decrease of the boundary-layer depth in the afternoon transition is a delicate process and in practice, its estimation depends on the criteria used to derive the boundary-layer depth as already shown by Angevine and Grimsdell (2002) and Bennett et al (2010).'*

5. Figure 10 : should look at the time when the virtual temperature flux becomes negative in the afternoon ; or, similarly, how this time relates to latent heat flux (and hence vegetation type and soil moisture). If I recall correctly, the sensible heat flux went negative earlier where there was large latent heat flux based on CASES-97 data. Which meant more variability in this time both spatially and from day to day for sensible heat flux than for virtual temperature flux.
Thanks for this comment. We redrew Figure 10 with the virtual temperature flux. Below you can find both figures, in the left hand side, the time when the temperature flux becomes negative and in the right hand side, the time when the virtual temperature flux becomes negative. Indeed, the time when the virtual temperature flux goes negative is later than the time when the temperature flux becomes negative and the scatter is reduced. We have now included in Figure 10 the one computed with the virtual temperature flux.





General editorial comments (more detail later) :
1. The figures are impossible to study in printed form. I am reviewing the paper with the figures enlarged on the screen. The labels on the right hand side need to be larger, and 'range' might be a better label than 'variability'. Also it would be helpful to the reader to label the 'hot' days referred to in the text. And have labels on all the figures to make it easier for the reader.
We have modified the figures accordingly. The labels on the right hand side of Figures 2, 3, 4 are larger and we have replaced 'variability' by 'range'. We have also added labels for the days in all the plots and a label indicating the hot days.

2. It would help in the profile figure (6) to figures to have grid points on the curves for each model –

for one curve for each model. Also, might consider plotting the average profile for each model and time. And finally might consider offsetting the soundings by adding a few degrees for each time interval. The last might not be practical (You could stretch the horizontal axis and only have one altitude label).
We have plotted the average profile for each model and time which is shown with dashed lines and a slightly different color than the profiles for each point. We have also added a 2K and a 2g/kg offset for each time interval in order to better distinguish the different hours.

3. It would be useful to have a table describing the properties of each model (horizontal and vertical grid spacing, PBL scheme, etc.) as well as model-run length and initiation time and data used to initialize the model. Also how the land-surface properties were initialized (often there is a long-spinup).
We have added a Table (now Table 2) describing the properties of each model.

References:

Honnert T, V Masson, F Couvreux, 2011: A diagnostic for evaluating the representation of turbulence in atmospheric models at kilometric scale. J Atmos Sci, 68, 3112-3131
Ricard, D., Lac, C., Riette, S., Legrand, R. and Mary, A. (2013), Kinetic energy spectra characteristics of two convection-permitting limited-area models AROME and Meso-NH. Q.J.R. Meteorol. Soc., 139: 1327–1341. doi: 10.1002/qj.2025

Answer to the reviewer 2 about the manuscript entitled : 'Boundary-layer turbulent processes and mesoscale variability represented by Numerical Weather Prediction models during the BLLAST campaign' by F. Couvreux et al.:

First, we wish to thank the reviewer for her/his review. Below is our response (in blue) to the comments on a point-by-point basis. References to how we plan to modify the text is indicated in italic.

General comments:

After careful reading, my general impression is that the manuscript contains relevant and sound scientific findings as result of a massive analysis work, and deserves publication. It is a pity, though, that the exposure of such a wealth of results is rather poor. The reading is hard and fragmented, with too many inaccuracies and repetitions. The style needs improvement before the paper can be accepted for publication. To my opinion, the figures are simply not to the ACP standard and require a complete rethinking, not only for publication but even for review. I have struggled to get useful information out of the figures in their current format. I leave the editor the decision if they can be accepted as they are. The structure and 'paragraphing' used for the discussion of the results, on the other hand, seems appropriate.

We have tried to improve the layout of the paper by improving the quality of the figures and trying to suppress repetitions and inaccuracies. Figure 1 has been enlarged to be more visible. A new Figure 2 has been added to present the orography in the real world and in the different models. In particular, old figures 2, 3 and 4 have been simplified and now only show the mean curves as you proposed (cf answer to your specific comment below). A figure presenting the overestimation of the sensible heat fluxes in ARPEGE has also been added and is presented with error bars to show the standard deviation in the observations. The colors of Figure 7 have been modified. We suppressed the appendix for clarity. In addition, a native English speaker has checked out the entire manuscript. We hope that now the paper reads more easily.

Specific comments:

First two lines of page 3. I don't understand the meaning of the sentence. Can you please clarify?
Those two lines have been modified to '*Several recent studies also assessed the behaviour of single-column models to represent the entire diurnal cycle by comparison to LES.*'

Second line of page 3. '. . .single-column runs ARE often used as a simplified configuration OF a full 3D simulation ...' . Also, define 'single-column' models. Done, we now define single-column model as *one single column of the atmosphere that integrates the same suite of parameterizations as a full 3D simulation*.

Line 10 page 3. '. . .are quite rare compared to. . .' Done

Define tke the first time you introduce the turbulent kinetic energy and (same for IOP). OK This is now done on page 2 line 7 for the turbulent kinetic energy. IOP and tke are introduced twice, once first in the abstract and a second time in the main text.

Remove 'days' after IOP. Done throughout the text

Page 4, line 7.' . . . all surface stations measuring turbulence. . .' Done

The first two lines of section 2.3 can be removed, or at least, rephrased. 'Due to the coarse grid spacing. . .' Following your advice, we have rephrased the sentence as : « *Due to the coarse grid spacing of each model, real surface heterogeneities, topography and local circulation are not expected to be reproduced by models.*"

Page 6, line 15. '. . .the tke is below . . .. In the observations. . .' Done

Page 6, line 18. ‘. . .usually provides an estimate. . ., based on the vertical gradient of the relative humidity’ Done

Page 7, line 5. ‘. . .at a given hour h correspond. . .’ Done

The paragraph at the beginning of section 3 should be moved to the methodology section Page 11. As you proposed, we moved the paragraph at the beginning of the section 3 to the end of the methodology section (in the part 2.3).
line 12. ‘. . .variables indicating different. . .’ Done

Page 11, line 26. ‘. . . the boundary layer depth estimated by the model with the boundary layer depth estimated by the observations’. Done

Page 11, line 29. Please provide reference We have added in the methodology section some words on the comparison of the boundary-layer height diagnosed from the tke profiles versus boundary-layer height diagnosed from thermodynamical profiles.

Page 11, line 30. ‘. . .the temporal variability in terms of maximum boundary layer depth from on a day to the other. . .’ is not clear. Do you mean the variability diurnal cycle?
We wanted to comment in terms of variability from one day to the other of the maximum boundary-layer depth of the day, so not the diurnal cycle. This has been changed to '*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.*'

Page 11, line 34. ‘. . .the physics of the models respondS . . .’ Done

The end of page 8 is a left-over of some copy-paste?
We are sorry that this happens. In fact, at the last minute, we had to use another template provided by the journal. In fact this was note a left-over of some copy-paste but a foot note. However, due to no changes in the police, neither in its side, it really looks like a copy-paste. We decided to suppress the footnote and included the information into parenthesis in the main text.

The first sentence of section 3.3 is unnecessary (already said a few times) We suppressed this sentence.
In the Appendix the last words sounds strange..’ A=3 would be a value too large’.
Eventually, we decided to suppress the Appendix and therefore the supplementary 2.

Table 2. The roughness length is measured in meters
You are right. We have added the unit in this table.

Figures 2. I would suggest to keep only the mean curves and/or to replace the time series with box and whiskers, four for each IOP (obs plus three models) or three is you prefer to plot the bias (obs – mods). Add the legend to all figures if possible, to help the readers.
As you propose, we decided to only keep the figures showing the mean curves. We also added a figure to show the over-estimation of the sensible heat flux for ARPEGE point (now Figure 3). In this figure, we have use bars to indicate the horizontal standard deviation of observations. Eventually, now a complete caption is present for each figure.

Figure 6. The choice of colours is unfortunate. Why not blue, red and green for example? The graph is anyway difficult to interpret, please try to make clearer (in the caption please use ‘becomes’ in place of ‘goes’).

We have changed the colours. The colours are chosen here to reflect the daytime to nighttime evolution with red for 1200 profile, orange for 1400, dark-green for 1600, purple for 1800 and grey for 2000. We also changed 'goes' to 'becomes'.

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] ~~[3] Eric Nilsson[2,3]~~

[1]CNRM (Météo-France and CNRS), Toulouse, 31057, France
[2]Laboratoire d'Aérologie, University of Toulouse, CNRS, Toulouse, France
[3]Uppsala University, Uppsala~~Stockholm~~, 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~~AROME, ARPEGE and ECMWF~~, 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~~AROME is a 2.5km limited area non-hydrostatic model operated over France, ARPEGE a global model with a 10km grid-size over France and ECMWF a global model with a 16km grid-size. We analyze~~ the representation of the vertical profiles of temperature and humidity and the time evolution of near surface atmospheric variables and ~~as well as~~ the radiative and turbulent fluxes over~~for~~ a total of 12 ~~24h-long~~ Intensive Observing Periods (IOPs) each lasting 24h. Special attention is paid to the evolution of the turbulent kinetic energy (*tke*), which ~~that~~ was sampled by a combination of independent instruments. For the first time, this variable, ~~which is~~ a central one~~variable~~ in the turbulence scheme used in AROME and ARPEGE, is evaluated with observations.

In general, the 24-h ~~24h~~ forecasts succeed in reproducing the variability from one day to another in terms~~the other in term~~ 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~~ and over-predominance of forests~~. AROME shows a moist bias within the daytime boundary layer, which is consistent ~~consistently~~ 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 ~~better~~ resolves the vertical structures better, in particular the strong daytime inversion and the thin evening ~~evening thin~~ stable boundary layer. This model is also able~~capable~~ to capture some specific~~the peculiar~~ observed features, such as the orographically-driven subsidence and a well-defined maximum that arises during the evening of the ~~in~~ water vapor mixing ratio in the upper part of the residual layer ~~that arises during the evening~~ due to fine scale ~~mesoscale~~ advection. The model reproduces ~~mesoscale variability is analyzed and~~ the order of magnitude of spatial variability observed at mesoscale (a few tens of kilometers)~~is also well reproduced in AROME~~. AROME provides a good simulation of the diurnal variability of the turbulent kinetic energy while ARPEGE shows the~~a~~ 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 ~~Due to the increasing resolution, it becomes~~ important to evaluate their capability to reproduce~~in reproducing~~ the low-troposphere vertical profiles of temperature and moisture and ~~as well as~~ their surface turbulent and radiative fluxes as they being increasingly ~~are more and more~~ used for numerous applications such as predictions of black ice on roads ~~road black ices~~ or agro-meteorology~~ for instance~~. Here we present the performance, which has remained largely unexplored so far, of these models in representing ~~of those models on the representation of the~~ near-surface variables and boundary-layer turbulent kinetic energy (*tke*)~~which has been largely unexplored~~.

The evaluation ~~Evaluation~~ and improvement of models is often a motivation for deploying~~to deploy~~ instruments in field campaigns~~campaign~~. However, field campaign observations are less ~~not so~~ often extensively used to evaluate the representation of surface and boundary-layer processes by operational models. Atlaskin and Vihma (2012) used~~use~~ observations from a field campaign to evaluate NWP models. They focused~~focus~~ on the representation of very stable conditions at very low temperatures (<-10°C) in ~~in the~~ northern Europe and showed~~show~~ a systematic positive bias for the 2-m temperature, due to an underestimation of the stratification ~~warm bias~~ 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 ~~various diagnostics of~~ the boundary-layer schemes and their diagnostics based on~~depth applied on simulations of a~~ mesoscale model simulations. In parallel, models have been evaluated ~~evaluation of models has been carried out~~ over permanent observing sites such as the ground-based remote sensing observations from the Swiss plateau (Collaud Coen et al., 2014), ~~from~~ the Atmospheric Radiation Measurement (ARM, Morcrette 2002 or Guichard et al., 2003) or the Cloudnet ~~Cloud-Net~~ sites (Illingworth et al., 2007). In particular, the Cloudnet project has allowed ~~CloudNet project allows~~ a systematic evaluation of clouds in different operational forecast models. For instance, Bouniol et al. (2010) showed~~show~~ that models tended~~tend~~ to overestimate ~~the~~ 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~~size~~ 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 ~~surface types~~ (Lothon et al., 2014). During this campaign, twelve fair-weather days were extensively documented by Intensive Observing~~Observation~~ Periods (IOPs). These days corresponded~~Those days correspond~~ 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. ~~Indeed,~~ Acevedo and Fitzjarrald (2001) used ~~showed with~~ observations complemented by a Large-Eddy Simulation (LES) to show that the spatial variability peaked~~peaks~~ in the evening transition and that land use and orography played~~play~~ 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~~plays~~ 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.

~~In addition to a better understanding of the processes involved in the transition, several recent studies assessed the behaviour of single-column models to represent the entire diurnal cycle by comparison to LES. Single-column runs is often used as a more simplified configuration than a full 3D simulation in order to highlight some deficiencies in the physics parametrization of the model and to test new developments. By comparing 1D model to LES on a case based on observations at Cardington, UK (Beare et al., 2006) that covers the transition from early afternoon to the next morning, Edwards et al. (2006) show that 1D model had difficulties to represent correctly turbulence diffusivity during the afternoon transition which impacts on 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 show a faster decrease of the temperature in the afternoon temperature compared to LES. However, such evaluation has not been carried out for operational NWP models and have not used observations of turbulence in the entire boundary layer. For example, observations of the turbulent kinetic energy are quite rare relatively to mean meteorological profiles, and often punctual (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 first-order scheme (Holt and Raman, 1988), has only been evaluated through comparison to LES (Cuxart et al, 2006 for instance). Here, we will 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~~to predict~~ 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~~for~~ which the boundary-layer parametrization is based on a prognostic evolution of the turbulent kinetic energy, iii/ to evaluate the variation~~evolution~~ of surface thermodynamic parameters for different covers. The observations and the models evaluated ~~Observations and evaluated models~~ are described in Section 2 together with ~~section 2 as well as~~ the methodology used to carry out the comparison. Results are presented in Section 3,~~section 3~~ focusing on the general representation of the entire diurnal cycle~~-~~: we provide separate analyses of ~~separately analyse~~ 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 ~~have been~~ acquired during the BLLAST field campaign and have been described in details by~~in~~ 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 ~~have not been~~ used in the assimilation system and could~~can~~ therefore be used for evaluation purposes~~purpose~~ without ambiguity. Table 1 summarizes all the types of data and measurements used in this study, giving ~~with~~ details on the resolution of the raw data, the estimated parameters and their sampling. In the following, we use~~used~~ the observations from the 12 IOPs~~IOP days~~ of the field campaign (Lothon et al., 2014).

In total, 7 different sites were instrumented with~~by~~ eddy covariance systems~~system~~ 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~~are~~ the two main land types of the area while moor and urban surface types were~~are~~ intermediate and corn, wheat and bare soil were minority covers ~~are in minority~~ (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~~turbulence station measurements~~ and provided surface turbulent and radiative fluxes at 30 minutes'~~a 30min~~ resolution (De Coster and Pietersen, 2012). These observations were~~are~~ used to evaluate the radiative and turbulent fluxes and also ~~as well as~~ 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 ~~has been~~ interpolated to 10 m ~~10m~~ for comparison with~~to~~ 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 ~~we use the radiosoundings (MODEM~~) 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~~) on~~ the north-easternmost site (“main site”~~main site~~ in the following, indicated by large orange dots in Fig. 1b), ii/ radiosondes~~hourly radiosoundings~~ (Vaisala RS92 probes) in~~of~~ the lower troposphere (up to 3 to 4 km, Legain et al., 2013) launched hourly from the southern most launching site (4 km ~~apart~~ 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 ~~from~~ 4 to 10 soundings of the lower troposphere during the afternoons~~afternoon~~ of the IOPs. These measurements provided~~Those measurements provide~~ vertical profiles of temperature, water vapour content and horizontal wind. Boundary-layer depths were~~are~~ derived from these~~those~~ profiles as detailed in section 2.3. Boundary-layer depths derived from UHF and aerosol lidar data were~~are~~ also used.

The combination of various measurements that provided~~provide~~ estimates of the turbulent kinetic energy was a unique aspect ~~specificity~~ of this field campaign. The Doppler lidar (Windcube, manufactured by Leosphere, Gibert et al., 2012), measurements from ground towers, aircraft ~~measurements~~ and the turbulence probe mounted on the tethered balloon (Canut et al., 2016~~, 2015~~) all contributed~~provide~~ 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~~therefore~~ estimates of the turbulent kinetic energy ~~(*tke*)~~.

2.2 Numerical weather prediction models

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

- two NWP models from~~of~~ Météo-France: (i) a global model, ARPEGE (Courtier and~~et~~ Geleyn, 1988) with a stretched horizontal grid of about 10 km x 10 km over France and~~with~~ 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~~PBL~~ scheme, initialization time and forecast period~~run~~, initialization of the land-surface properties) for the three models. ~~Table 4 presents the main physiographic characteristics (altitude, albedo, vegetation fraction and roughness length) of the points extracted from those models for the different IOPs.~~

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 ~~. This allows to provide~~ specific outputs for the 16 grid-~~-~~points surrounding the main site (Fig 1b).

All models employed~~employ~~ a terrain following hybrid sigma-pressure vertical coordinate. However, the ~~The~~ vertical grid differed ~~however differs~~ from one model to another ~~the other~~ (Table 2): ARPEGE had~~has~~ 70 vertical levels with about 11 levels within the first km (first level at 16 m~~16m~~), AROME had~~has~~ 60 vertical levels with about 15 levels within the first km (first level at 10 m~~10m~~), and ECMWF has 91 vertical levels with about 11 levels within the first km (first level at 10 m~~10m~~). The time step varied~~varies~~ from 1 min for the AROME model to about 10 min for ARPEGE and ECMWF. The models also differed~~differ~~ by their ~~different~~ 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~~to represent~~ shallow convection is active (Bechtold et al., 2001). ECMWF uses an~~a~~ 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~~on~~ the shortwave component: ECMWF~~ARPEGE~~ uses the SRTM parametrisation~~RRTM parametrization~~ while AROME and ARPEGE has the Morcrette at~~et~~ al. (2001) code~~parametrization~~. The radiation scheme is called every hour for ARPEGE and every 15 min for AROME. ~~Note that, at the time of the field campaign, in the operational version of ARPEGE the radiation scheme was called every three hours and this induced an abrupt unrealistic decrease of the incoming shortwave radiation during the afternoon transition (not shown) that has now been corrected in the operational model with a hourly call.~~ Concerning the cloud scheme, ARPEGE uses a distribution of relative humidity based on Smith (1990), AROME a distribution of the saturation deficit ~~deficit saturation~~ based on Bougeault (1982) and ECMWF uses a prognostic scheme (Forbes et al., 2011). In ARPEGE, there are 12 different~~various~~ vegetation covers and one grid point can have only one given vegetation cover ~~but a low or high vegetation criterion is affected to each point to rapidly distinguish the points in term of stomatical resistance and roughness length (Table 2)~~ while in AROME each grid is associated with~~to~~ a certain fraction of various vegetation types (crops~~culture~~, land, town, mixtures of crops~~crop~~ and woodland, Landes forest or broad-leaf~~leaves~~ forest).

2.3 Comparison methodology

~~2.3 Methodology of comparison~~

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 km^2, 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.

~~Due to the resolution of each model, real surface heterogeneities, topography and local circulation, a perfect match between observations and model outputs is not expected. In particular, as shown in Lothon et al. (2014), a large variability on the observed surface fluxes exists among the sites (Fig 1) even at scales smaller than 2.5 km, the size of a grid box of AROME (see for example the differences between the moor and the corn sites, or the grass and the wheat sites) that are mainly due to surface cover. However, the variability among observations and the differences between model outputs and observations provide inferences on the main drawbacks of the models. In this section, we detail how the comparison is conducted focusing on the temporal and spatial resolution of the different variables obtained from models and observations.~~

~~The simulated grid-points (and associated columns) surrounding the locations of the measurement sites have been extracted as shown in Fig. 1: 3 neighbouring grid points have been extracted for ARPEGE, 16 neighbouring grid points for AROME (a box of 10x10km² including all sites) and 9 neighbouring grid points for ECMWF.~~ For ECMWF we evaluated~~evaluate~~ both the analysis available every 6 hours and ~~as well as~~ 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~~, was selected~~ to be 24h. The atmospheric variables corresponded~~correspond~~ to instantaneous fields sampled every hour for AROME and ARPEGE and every 6 hours for ECMWF. The diagnostics T2m (temperature at 2 m~~2m~~), rh2m (relative humidity at 2 m~~2m~~) and ws10m (horizontal wind speed at 10 m) were~~10m) are~~ obtained using a vertical~~an~~ 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~~gets~~ below 0.01 m² s^{-2}. In the observations, various diagnostics allowed ~~allow to derive~~ 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~~is~~ an estimate of the inversion height based on the vertical gradient of the relative humidity~~as this criterion detects the level of a humidity vertical gradient~~

ii/ the first level below the height diagnosed through i) ~~previous height~~ where the *tke* dissipation rate becomes~~gets~~ greater than a threshold (10^{-3} m²/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~~on~~ the various vertical profiles from soundings and RPAS (Remotely Piloted Airplane Systems) (Lothon et al., 2014), using ~~either~~ 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~~retained. The~~ comparison of different boundary-layer depths derived from various instruments is presented ~~has been illustrated~~ 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 ~~depends on~~ 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 ~~. This~~ will be given~~detailed~~ in Sect. 3.5. The diagnostic used in the model was ~~has been~~ compared to the criteria iv) applied to ~~applied on~~ the model profiles. These two diagnostics were consistent but in ~~In~~ ARPEGE, the model diagnostic tended~~tends~~ to overestimate the value derived from the profiles by about 200 m while,~~of about 200m while~~ in AROME, there was ~~there is~~ a very good agreement except for 14 June after 1500 UTC and ~~1500UTC,~~ 15 June after 1400 UTC ~~and 26 June~~ due to the presence of clouds (discussed later). In ~~Therefore in~~ the following, we will use the model diagnostic discarding these~~those~~ hours of disagreement as it depicts the turbulent layer, in particular during the afternoon transition.

When comparing observations and modelling, we considered ~~have taken into consideration~~ 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 ~~have been~~ processed every 30min and ~~are~~ then averaged to provide the 1-hr ~~1hr~~ average for the comparison. Furthermore, it should be kept ~~one must keep~~ in mind that the area (footprint of a ~~the reduced surface (~~few hundred metres) of the surface ~~footprint)~~ sampled in the measured surface turbulent fluxes was small relative~~compared~~ 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.

~~Concerning the *tke*, in the observations, it has been estimated for 20 min time windows for the 60m-tower, the Doppler lidar and the tethered balloon, 10 min for the 10m-tower (sensitivity to a computation with 20min did not change the results) and for horizontal legs of 25-30 km for the aircraft measurements (corresponding to 5 min cf Table 1 and Canut et al, 2015 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 5km high-pass filter has been applied only to the aircraft raw data before the calculation of the *tke* to filter out the mesoscale variability. We also tested the *tke* estimates obtained with 2.5km high-pass filter but it was affected by a large time-variability which highlighted that the samples were not large enough.~~

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.

~~In the models, a horizontal resolution of 2.5km and 10km respectively in AROME and ARPEGE is equivalent to 9 and 30 min respectively according to a wind speed around 3-5ms⁻¹ in the boundary layer, which is consistent with the 20 min used to derive the *tke* from surface point observations. The estimation of the *tke* with the Doppler lidar (Gilbert et al, 2012) assumes that the turbulence is isotropic and derives the value from the measured vertical velocity variances. To evaluate this hypothesis, we compute the ratio~~

$$A = 1.5 \frac{\overline{w'^{2}}}{\text{tke}}$$

~~a coefficient from the tower measurements (both from the 60m tower and the 10m 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 null). When A<1, the contribution of horizontal variance is dominant. Both the tower measurements as well as the tethered balloon[1] measurements indicate that above 0.1 to 0.2 zi, zi being the boundary-layer height, and in the middle of the boundary layer, this coefficient is between 1 and 2 suggesting that the variance of the vertical velocity is often the main contributor to the *tke* at that height and the *tke* can be estimated from the~~ $\overline{w'^{2}}$ ~~as~~ $\text{tke} = 1.5\,\overline{w'^{2}}$ ~~. A sensitivity to this ratio for the estimation of the *tke* is indicated in the Appendix. Aircraft measurements indicate that closer to the top of the boundary layer this coefficient decreases again with value between 0.75 and 1. Below 0.1 zi, the variance of horizontal wind is important and this coefficient is mostly below 0.6 (see Canut et al, 2015 for more details). Therefore, in the following, we only use Doppler lidar estimates from altitudes above 100m. More complex computations taking into account the day-to-day and vertical variation of the anisotropy factor derived from tethered balloon or aircraft could be done in a future study. Note also that as we derive the *tke* as 1.5~~ $\overline{w'^{2}}$ ~~we tend to overestimate the observed *tke* most of the time but we may underestimate it on days with more wind, conditions in which horizontal wind fluctuations are expected to be larger.~~

Eventually, in order to characterize the afternoon transition (AT), the time at which the buoyancy flux became negative was~~sensible heat flux gets negative is~~ determined in both observations and models. This was~~is~~ done by finding the 0 cross-~~-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.

~~In the following, we evaluate the representation of the diurnal cycle of the boundary-layer characteristics and surface energy budgets over all IOPs.~~

3 Results

[1] The tethered balloon never reaches height above 500m

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

~~In this section, we compare surface fluxes, meteorological variables, boundary-layer structure, turbulent kinetic energy for the 12 IOP days. As shown in Lothon et al (2014), those days correspond to mainly high-pressure fair-weather conditions with no cloud cover or a small amount for 14, 15, 24 and 30 June. 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 which was due to the presence of a low pressure in the Gulf of Lion (for more details see Nilsson et al, 2015a). In the following, those three days will be referred to as hot days.~~

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.

~~Figure 2 presents series of 24h sequences for the 12 different IOPs (from 14 June to 5 July 2011), of the observed and simulated surface downwelling solar radiation. In Figure 2a, the different model grid points are plotted as well as the dark grey shading that indicates the envelope containing the different surface sites, which quantifies the spatial variability. Figure 2b shows the mean value and the maximum range[2] for a given type (observations or models) averaged for daytime and nighttime respectively as a measure of the spatial variability. The cloudy days are clearly depicted by an increase in the spatial variability of the observed surface downwelling solar radiation (Fig 2a) consistently with Lothon et al (2014). ARPEGE and AROME mostly distinguish between the clear-free days (noted 'C') and the cloudy days indicated by an increase spatial variability (Fig 2b). ECMWF for at least two observed clear days (20 June, 27 June) depicts a decrease of downwelling solar radiation from 1030 to 1330 UTC which suggests the presence of clouds in the model. The 26 June has some clouds from 1400 UTC to 1900 UTC while ECMWF predicts variability in the downwelling solar radiation from 1030 to 1330 UTC. The 27 June has high clouds in ARPEGE throughout the day while observations only registered thin cirrus after 1700 UTC (not shown). The 30 June presents stratocumulus in the morning that clear up through the afternoon with however quite a variable cloud cover in the afternoon while ARPEGE and ECMWF predict a cloud-free atmosphere. The spatial variability is slightly overestimated for 14, 15, 30 June in AROME but otherwise in good agreement with~~

[2] This is computed at each time step by the difference between the maximum and the minimum over all the points of the given type

~~observations. In summary, all models capture in general the spatial and temporal variability in downwelling solar radiation with however a better behaviour for AROME in terms of cloud occurrence and spatial variability.~~

There is more discrepancy in the simulation~~simulations~~ of sensible heat fluxes with biases reaching more than 100 Wm^{-2} (Fig 3b). For instance, ~~3a). First, ARPEGE predicts very large sensible heat fluxes which have similar range as observations above the forest (dashed and dash-dotted black lines in Fig 3a) for two of the three points (ARP1 and ARP3 in Table2 which mainly differ from ARP2 in terms of altitudes and roughness lengths) : those two model grid-points are characterised by high vegetation cover which have lower albedo (0.12 against 0.2); they are also at higher altitude. These simulated sensible heat fluxes are too large values to be representative of a 10km wide grid box over the area which is characterized by much more surface heterogeneities at this size (cf Fig 1). The third point (northernmost, ARP2) is in better agreement with the non-forest sites (indicated by the grey shading).~~ ECMWF overestimates the surface sensible heat fluxes. The variability from one IOP to another ~~the other~~ (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~~overestimate the~~ value in the first night (25 to 26 June). Concerning the spatial variability, ~~one can note~~ 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~~range~~ especially during the first part of the period. The spatial variability among the various ECMWF grid ~~-~~points is much smaller; this ~~which~~ 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).

~~Latent heat fluxes predicted by AROME systematically overestimate the observed values by up to 100 Wm^{-2} (Fig 3c) and this may be related to a too large soil moisture content (however, no observations were available at various sites to evaluate this variable). The two high-vegetation points of ARPEGE do not tend to evaporate more as could have been expected from a larger net radiation (due to a lower albedo). ECMWF correctly reproduce the range of observations. The variability among the various IOPs is also correctly reproduce with higher latent heat fluxes during the hot days (Fig 3d).~~

~~The spatial variability is about the same order of the observed one in AROME, slightly underestimated in ARPEGE and strongly underestimated in ECMWF. Interestingly, when plotting the latent heat fluxes as a function of 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) versus the cloudy days (green and purple symbols) in agreement with Lohou et al (2014). Most of the observations also record a negative relationship (even though with a less steep slope) except the observations at 60m on the tower (grey squares) and observations at 30m over the forest (dots).~~

To sum-up, we ~~In summary, one can~~ note an overestimation of the sensible heat flux by ARPEGE for the two points covered with forest and, to ~~with high vegetation and by ECMWF in~~ 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~~underestimate the observed spatial variability. This underestimation is larger~~ for ECMWF probably because of~~due to~~ 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.

~~Figure 4 presents the same figures as Figure 3 for the 2m temperature, 2m water vapour mixing ratio and the 10m wind speed observed and simulated. AROME and ARPEGE are in very good agreement with the observed close to surface meteorological variables. First, all models reproduce the variability, through the period, of the 2m temperature with in particular a warming period 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 can not be analysed in ECMWF with a 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). Interestingly, the spatial variability in night time temperature among sites is smaller for the hot period; this might be due to larger wind speed during this period. The models do not reproduce this behaviour: during the hot period, the model predicts both an increasing variability of night sensible heat fluxes and 2m temperature. The underestimation of the spatial variability by AROME and ARPEGE during most days is not due to a misrepresentation of the wind as the wind is relatively weak over the whole period and in more or less agreement with observations. ECMWF overestimate the spatial variability which is partly explained by the westerly grid points being warmer (not shown). Also the diurnal cycle of the spatial variability in ECMWF is inverse compared to the observations with higher daily variability than nightly variability. This needs further investigation.~~

Concerning the 2-m~~2m~~ water vapour mixing ratio, the models reproduce the progressive moistening before~~increase that follows~~ a precipitating event (the days with precipitation were not IOPs and thus correspond to an interruption of time in Figure 4, ~~events (~~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 ~~overestimates the range of variations from night to day and the spatial variability. Also ECMWF has~~ a dry bias during daytime especially in the second part of the period. It can be seen ~~One can note~~ that the overestimation of the latent heat fluxes by AROME has no clear consequences in the reproduction of the 2-m~~2m~~ water vapour mixing ratio. Concerning the 10-m~~10m~~ wind speed ARPEGE and~~&~~ AROME reproduce higher~~larger~~ 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, ~~one can note a very good simulation of~~ the surface meteorological variables were well simulated in AROME and ARPEGE but were~~; it is~~ slightly less accurate in ECMWF especially for wind speed and water vapour mixing ratio. In the following sections, we focus only~~only focus~~ on the French models for which we have hourly outputs.

3.3 Vertical~~vertical~~ structure

~~Thanks to the numerous soundings of the atmosphere via various techniques (radiosoundings, low-atmosphere radiosoundings or RPAS profiling), it is possible to extensively evaluate the evolution of the boundary-layer vertical structure predicted by the models.~~

Figure 6~~5~~ presents scatterplots of the simulated versus observed values of the potential temperature and water vapour mixing ratio averaged over the first 500 m ~~500m~~ deep layer. First, there is ~~a~~ good agreement among~~between~~ 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~~5K~~. In ARPEGE, the temperature bias is dependent on the average temperature with less ~~no more~~ bias for temperatures higher than 305 K~~temperature greater than 305K~~. 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 ~~too high~~ latent heat flux being too high, while ARPEGE exhibits a dry bias. The AROME moist and cold biases are~~were~~ not clear in the time evolution of 2-m variables, indicating distinct~~2m variables indicated different~~ reproduction of the surface layer and~~versus~~ the boundary layer.

Figure 7~~6~~ 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 ~~IOP days the~~ 27 June 2011 (one of the hot days) and ~~the~~ 1 July 2011. AROME captures ~~better~~ the strong inversion in potential temperature that occurs at the top of the boundary layer (at 1400 UTC on ~~the~~ 27 June or 1 July) better ~~the 01 July)~~ and this is true for most of the IOPs. This may be due to the finer vertical grid. In both models, there is~~are~~ more spatial variability during the hot period than otherwise and this remains true throughout the day, and~~entire day, this~~ is consistent with the results at the surface (higher variability in terms of surface heat fluxes and 2-m ~~2m-~~meteorological variables) as shown previously. In particular in AROME, on~~the~~ 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 ~~size~~ of ARPEGE. For 1 July, note the ~~One can note for the 01 July the~~ maximum in water vapour mixing ratio in the upper part of the boundary layer simulated by AROME; this maximum ~~, which~~ is also observed in the radiosoundings. Analysis of the moisture budget indicated~~indicates~~ that this maximum was~~is~~ mainly related to fine scale advection not resolved at 10 km (not shown)~~advection (not shown) suggesting that mesoscale circulation has an impact on this peculiar boundary-layer structure~~.

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.

~~To further assess the representation of the vertical structure of the boundary layer, we compare the boundary-layer depths estimated by the model with boundary-layer depths estimated with observations. Figure 7 presents the time evolution of the different boundary-layer depth estimates for all the IOPs. The overestimation of the boundary-layer depth for 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* that depicts the top of the shallow cumulus layer. Both AROME and ARPEGE are able to reproduce the temporal variability in terms of maximum boundary-layer depth from one day to the other 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 highlights that the physics of the models respond 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 the 20, 24, 25, 30 June and 02 July, slow growth the 26 June, 27 June and 05 July and rapid growth the 14, 19 June and 01 July. This distinction is reproduced by the models. Evaluating the decrease of the boundary layer in the afternoon is more complex. The aerosol diagnostic based on the lidar measurement always depicts the top of the inversion layer in the afternoon while the profile diagnostics as well as the reflectivity gradient from the UHF~~

~~indicate either the top of the stable layer or the top of the inversion layer depending on the cases. The model diagnostic depicts the top of the turbulent layer which is also the case of the boundary-layer depth diagnosed from the dissipation rate measured by the UHF. The difference between those diagnostics in the afternoon depicts the existence of a pre-residual layer in-between the top of the turbulent layer and the top of the inversion layer as detailed in Nilsson et al (2015b). Concerning the decrease, ARPEGE most of the time predicts a later decrease than AROME. AROME is in better agreement with the boundary-layer depth diagnosed from the dissipation rate even though AROME tends to be slightly higher, which could also be explained by the fact that the turbulence variable used to diagnose the boundary-layer depth is different: tke versus dissipation. Eventually, one can note the large spatial variability among the model grid-points in particular 26 and 27 June and 02 and 05 July, however the highest boundary layer is not systematically always over the same grid point, so this can not be explained by particular surface characteristics.~~

3.4 Turbulent~~turbulent~~ kinetic energy

A unique feature ~~specificity~~ of this campaign was~~is~~ the existence of various simultaneous measurements of the turbulent kinetic energy at various heights in the atmosphere. We used these~~use those~~ 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.

~~Figure 8 presents the time evolution of the *tke* for all the IOP days 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 modelled *tke* at the first level (at 11m in AROME and 17.5m in ARPEGE). Often, observations show significant *tke* in the morning that is not simulated except for a few days (25, 26 and 27 June for AROME and 24 June for ARPEGE) characterized by a larger wind speed and therefore a stronger shear production (Fig4c). 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 to a minimum of wind speed which is present for most days with weak wind. Note that the maximum measured the evening of the 27 June is associated to convective storms and are reproduced by the models. Those morning and evening *tke* values are related to slope-wind and also potentially effect of nocturnal low-level jet in the early morning. ARPEGE tends to present a Gaussian diurnal cycle of the *tke* most of the days (except 3 days : 24 June, 27 June and 05 July where maximum of *tke* exist 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 the other. This underestimation is in apparent contradiction with a larger sensible heat at least in the end of the period. The higher value in ARPEGE can be explained by a higher model level (17.5m versus 11m, as less turbulence is expected close to the ground), a larger grid size (9km versus 2.5km). Higher in the atmosphere, the modelled *tke* and observed one are in better agreement. Note that the various types of observations agree together in terms of intensity. The temporal variability at those 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 the *tke* during the day (see also Nilsson et al, 2015). At 60m and higher up, AROME has systematically less *tke* than ARPEGE probably for the same reasons as for the low levels.~~

~~Figure 9 illustrates the time evolution of vertical profiles of the turbulent kinetic energy modelled and observed for the 1 July (this is the only day where we have enough observations to retrieve a time-varying vertical profile of the *tke*). AROME has lower *tke* than ARPEGE and it decreases the turbulence earlier (starting at 1400) than ARPEGE (starting at 1500) as also shown in Fig 8. The shape of the vertical profiles is consistent among each model and the observations. The lidar observations (triangles, note that this 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 which is not simulated by the models. However, reminds that the lidar only measures the vertical velocity variances and therefore neglects any fluctuations in the horizontal velocity variances. But the comparison of the squared (tethered balloon) and the triangle (Doppler lidar)symbols of the same colour and at the same altitude provides an estimation of the error obtained from this estimation : A is underestimated during daytime with values more around 1.3-1.8 while A is overestimated in late afternoon (1700 and 1800) with A around 0.4-0.8. This deserves further investigation with more measurements of the vertical profiles. Also the contribution of the shear contribution versus the buoyancy contribution in the creation of *tke* could be further analysed in observations and models and in general the budget of *tke*.~~

3.5 Afternoon transition

In this section, we focus on the afternoon transition period. During this period, the ~~Most of the physical processes, including turbulent ones, are small and on the same order of magnitude during the later part of the transition and 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 ~~IOP days~~ can be separated into~~in~~ the two categories proposed by Grimsdell and Angevine (2002) as defined~~depicted~~ by the behaviour of the UHF reflectivity with 24/06, 30/06, 1~~01~~/07 and 2~~02~~/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~~height~~ 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 ~~in potential temperature~~ and associated with strong shear as shown in Nilsson et al. (2016a). These ~~(2015a) ; those~~ 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~~depicted~~ by the height derived from the dissipation rate (in pink in Fig 8~~6~~). This ~~latter~~ 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 ~~better agrees~~ with the evolution of the height derived from the dissipation rate. The layer between the pink and the red symbols was named the ~~has been denoted as a~~ pre-residual layer by Nilsson et al., (2016b~~, (2015b~~). 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~~10~~ 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~~goes~~ negative, t_Hv0, through the IOPs and the various points. This time varies strongly ~~strongly varies in the observations~~ 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~~5~~ July. The range of t_Hv0 is much larger in AROME (green~~cyan~~ symbols) with a range varying from 2 hours to 6 hours with, however, ~~however~~ no systematic behaviour for a given point (indicated by a given symbol). AROME systematically has an earlier t_Hv0~~H0~~ than ARPEGE, consistently with an earlier decrease of turbulence. Also this ~~time~~ 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 ~~are doing~~ a relatively good job during the afternoon. This could be related to the quasi-stationary~~stationnary~~ behaviour discussed in Darbieu et al. (2015) and Nilsson et al. (2016a),~~ (2015)~~ where no changes~~change~~ in turbulence structure or characteristics are evident after normalization ~~once normalised~~ by the decreasing surface sensible heat fluxes. The difficulties increase ~~are picking up~~ in the very late afternoon. We have also noted~~indicated~~ more difficulties when~~in~~ the models attempt to reproduce the varying characteristics of close-to-~~ to~~ surface variables at night. This highlights the models' difficulties in reproducing ~~highlighting difficulties in the models to reproduce correctly the~~ 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 km^2 in this 10 x 10 km^2 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.

~~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 that enables us to extensively evaluate three numerical weather prediction models. In summary, all models reproduce the temporal variability observed among the different IOPs in terms of variations of the cloud amount (clear versus cloudy conditions), of maximum height of the boundary layer, of variations of temperature. This is also a necessary first step if we want to further use those models to derive large-scale fields such as large-scale advection that are needed for smaller scale modelling studies. For instance, during the hot period, models and observations predict less sensible heat fluxes, larger temperature, larger wind speed, less *tke*. The different types of growth of the boundary layer encountered during the field campaign and detailed in Lothon et al (2014) are correctly distinguished by AROME and ARPEGE. However, systematic biases appear over those 12 IOPs: too large latent heat fluxes in AROME, a too large relative humidity diurnal amplitude at 2m and a dry bias during the day for ECMWF (especially at the end of the period). For two ARPEGE points the surface fluxes are similar to measurements over forest whereas the satellite map does not indicate a homogeneous forest patch over 10x10km^2 in the area. AROME better reproduces the vertical structures as well as the variability among the different IOPs in boundary-layer depth in terms of daily maximum value or growth in the morning. The spatial variability reproduced by AROME is similar from the one derived from the various in-situ surface sites.~~

For the first time, ~~the~~ turbulent kinetic energy, the prognostic variable of the turbulence scheme in AROME and ARPEGE, has been evaluated. Both models reproduced~~reproduce~~ the right order of magnitude. AROME reproduced ~~better reproduces~~ the variation from one day to another ~~the other~~ of its diurnal cycle better while ARPEGE always predicted~~predicts~~ a similar bell shaped~~shape~~ evolution. However, AROME underestimated~~underestimates~~ the value while ARPEGE was~~is~~ in better agreement with the observed intensity. Note that we took ~~This may be due to difference in grid-size but also in physical parametrization. The EDMF scheme used in AROME predicts~~ the contribution of the ~~thermals to the turbulence by a~~ mass-flux scheme to ~~which indirectly feedbacks~~ the *tke* into account here. This may be due to differences not only in grid-size but also in physical parametrization~~via the thermal production term (the mass-flux scheme contributes to the buoyancy flux profile) whereas in ARPEGE, the turbulence is only reproduced by a *tke* prognostic scheme~~. In a future study, we could gain some insight by evaluating the different simulated terms of the near-surface *tke* budget that have~~has~~ also been derived from observations by~~in observations in~~ Nilsson et al. (2016a ~~(2015a~~).

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

~~Appendix~~

~~In order to evaluate the sensitivity of the determination of the tke using the $<w'^2>$ variance, we plotted in Supplementary Figure 2, the time evolution of the tke at two different vertical levels (200 m and 450 m) derived from Doppler lidar measurements using three different values of A. However comparison of this figure and Figure 9 indicates that A=3 is a much too large values.~~

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.~~**REMERCIER LEFE & MEDDE & DEPHY-2**~~

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

~~Canut, G., et al: The eddy-covariance method applied in a tethered-balloon, BLLAST issue, 2015~~

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 ~~Dis~~, 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 ~~Dis~~, 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

~~Instrument~~	~~Used measured parameters~~	~~Derived diagnostics~~	~~Time resolution/range~~	~~Spatial resolution/range~~	~~Location~~
~~Standard radiosoundings (MODEM)~~	~~q, q_v, wind speed~~	~~h_{BL}~~	~~0000, 0600, 1200, 1800 UTC~~	~~~10-15m/0-20km~~	~~Main site~~
~~Low-troposphere soundings~~	~~q, q_v, wind speed~~	~~h_{BL}~~	~~Hourly from 1200 to 2200 UTC in IOP~~	~~~10-15m/0-2km~~	
~~Turbulence station (eddy-covariance~~	~~T2m, q2m, ws10m, sensible & latent heat~~		~~30 min from 20 Hz (except the forest site~~		~~7 stations over wheat, grass,~~

~~system)~~	~~flux, u'^2, v'^2, w'^2~~		~~that has 10 Hz) sampling rates~~		~~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)~~	~~~75m /175m-4000m~~	
~~Doppler lidar~~	~~Vertical velocity~~	~~tke~~	~~4s time resolution; turbulence moments calculated on *20* min~~	~~50m~~	
~~Aerosol lidar~~	~~Aerosol backscatter~~	~~h_{BL}~~	~~4s time resolution but diagnostic derived every 15 min~~	~~15m~~	~~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 total and in the first atmospheric kilometer, first 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 with this model~~
~~ARPEGE~~	~~10 km~~	~~70 / 11 / 16m~~	~~10~~	~~ISBA~~	~~TKE prognostic scheme – mass-flux scheme for cumulus~~	~~00 TU; 36~~	~~From a surface reanalysis with this~~

							~~model~~
~~ECMWF~~	~~16 km~~	~~91 / 11/ 10m~~	~~10~~	~~HTESSEL~~	~~Non-local K profile ; mass-flux for cumulus~~	~~00-06-12-18 TU; 06~~	~~From a surface reanalysis with this model~~

~~Table 3. Surface characteristics of the various points extracted from the models~~

~~Points~~	~~Altitude (m)~~	~~Albedo~~	~~Vegetation fraction~~	~~LAI~~	~~Roughness length~~
~~ARO-1~~	~~535~~	~~0.18~~	~~0.95~~	~~3.4~~	~~0.78~~
~~ARO-2~~	~~611~~	~~0.19~~	~~0.93~~	~~3.5~~	~~0.53~~
~~ARO-3~~	~~595~~	~~0.19~~	~~0.92~~	~~3.2~~	~~0.26~~
~~ARO-4~~	~~558~~	~~0.20~~	~~0.92~~	~~3.4~~	~~0.16~~
~~ARO-5~~	~~552~~	~~0.20~~	~~0.92~~	~~3.5~~	~~0.24~~
~~ARO-6~~	~~605~~	~~0.19~~	~~0.93~~	~~3.4~~	~~0.38~~
~~ARO-7~~	~~609~~	~~0.16~~	~~0.85~~	~~3.3~~	~~0.45~~
~~ARO-8~~	~~593~~	~~0.17~~	~~0.94~~	~~3.2~~	~~0.39~~
~~ARO-9~~	~~532~~	~~0.19~~	~~0.93~~	~~3.5~~	~~0.49~~
~~ARO-10~~	~~567~~	~~0.19~~	~~0.91~~	~~3.7~~	~~0.37~~
~~ARO-11~~	~~579~~	~~0.20~~	~~0.91~~	~~3.3~~	~~0.18~~
~~ARO-12~~	~~575~~	~~0.19~~	~~0.91~~	~~3.5~~	~~0.47~~
~~ARO-13~~	~~505~~	~~0.18~~	~~0.93~~	~~3.8~~	~~0.83~~
~~ARO-14~~	~~521~~	~~0.18~~	~~0.92~~	~~3.7~~	~~0.64~~
~~ARO-15~~	~~529~~	~~0.19~~	~~0.88~~	~~3.2~~	~~0.23~~
~~ARO-16~~	~~527~~	~~0.19~~	~~0.90~~	~~3.5~~	~~0.38~~
~~ARP-1~~	~~701~~	~~0.12~~	~~0.86~~	~~3.7~~	~~1.8~~
~~ARP-2~~	~~477~~	~~0.2~~	~~0.84~~	~~3.2~~	~~0.17~~
~~ARP-3~~	~~778~~	~~0.12~~	~~0.85~~	~~3.6~~	~~1.93~~
~~ECMWF-1~~	~~1068~~	~~0.15~~	~~Not available~~	~~Not available~~	~~6.2~~
~~ECMWF-2~~	~~894~~	~~0.15~~	~~Not available~~	~~Not available~~	~~5.1~~
~~ECMWF-3~~	~~772~~	~~0.15~~	~~Not available~~	~~Not available~~	~~4.8~~
~~ECMWF-4~~	~~510~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.65~~
~~ECMWF-5~~	~~491~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~

~~ECMWF-6~~	~~463~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.88~~
~~ECMWF-7~~	~~282~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.65~~
~~ECMWF-8~~	~~314~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~
~~ECMWF-9~~	~~325~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~

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

~~Info avec caractéristique de végétation:~~

~~Points~~	~~Altitude (m)~~	~~Albedo~~	~~Vegetation fraction~~	~~LAI~~	~~Roughness length (m)~~	~~broad.leaves Forest~~	~~Cultures~~	~~Town~~	~~Land~~	~~Landes forest~~	~~mixtures~~
~~ARO-1~~	~~535~~	~~0.18~~	~~0.95~~	~~3.4~~	~~0.78~~	~~62~~	~~0~~	~~0~~	~~38~~	~~0~~	~~0~~
~~ARO-2~~	~~611~~	~~0.19~~	~~0.93~~	~~3.5~~	~~0.53~~	~~37~~	~~38~~	~~0~~	~~25~~	~~0~~	~~0~~
~~ARO-3~~	~~595~~	~~0.19~~	~~0.92~~	~~3.2~~	~~0.26~~	~~12.5~~	~~25~~	~~25~~	~~38~~	~~0~~	~~0~~
~~ARO-4~~	~~558~~	~~0.20~~	~~0.92~~	~~3.4~~	~~0.16~~	~~0~~	~~67~~	~~0~~	~~33~~	~~0~~	~~0~~
~~ARO-5~~	~~552~~	~~0.20~~	~~0.92~~	~~3.5~~	~~0.24~~	~~8~~	~~67~~	~~0~~	~~25~~	~~0~~	~~0~~
~~ARO-6~~	~~605~~	~~0.19~~	~~0.93~~	~~3.4~~	~~0.38~~	~~17~~	~~42~~	~~0~~	~~33~~	~~8~~	~~0~~
~~ARO-7~~	~~609~~	~~0.16~~	~~0.85~~	~~3.3~~	~~0.45~~	~~0~~	~~0~~	~~25~~	~~42~~	~~33~~	~~0~~
~~ARO-8~~	~~593~~	~~0.17~~	~~0.94~~	~~3.2~~	~~0.39~~	~~0~~	~~11~~	~~0~~	~~56~~	~~33~~	~~0~~
~~ARO-9~~	~~532~~	~~0.19~~	~~0.93~~	~~3.5~~	~~0.49~~	~~33~~	~~42~~	~~0~~	~~25~~	~~0~~	~~0~~
~~ARO-10~~	~~567~~	~~0.19~~	~~0.91~~	~~3.7~~	~~0.37~~	~~17~~	~~83~~	~~0~~	~~0~~	~~0~~	~~0~~
~~ARO-11~~	~~579~~	~~0.20~~	~~0.91~~	~~3.3~~	~~0.18~~	~~0~~	~~60~~	~~20~~	~~20~~	~~0~~	~~0~~
~~ARO-12~~	~~575~~	~~0.19~~	~~0.91~~	~~3.5~~	~~0.47~~	~~18~~	~~35~~	~~10~~	~~10~~	~~0~~	~~27~~
~~ARO-13~~	~~505~~	~~0.18~~	~~0.93~~	~~3.8~~	~~0.83~~	~~58~~	~~42~~	~~0~~	~~0~~	~~0~~	~~0~~
~~ARO-14~~	~~521~~	~~0.18~~	~~0.92~~	~~3.7~~	~~0.64~~	~~42~~	~~58~~	~~0~~	~~0~~	~~0~~	~~0~~
~~ARO-15~~	~~529~~	~~0.19~~	~~0.88~~	~~3.2~~	~~0.23~~	~~0~~	~~78~~	~~0~~	~~0~~	~~0~~	~~22~~
~~ARO-16~~	~~527~~	~~0.19~~	~~0.90~~	~~3.5~~	~~0.38~~	~~17~~	~~75~~	~~0~~	~~0~~	~~8~~	~~0~~
~~ARP-1~~	~~701~~	~~0.12~~	~~0.86~~	~~3.7~~	~~1.8~~						

~~ARP-2~~	~~477~~	~~0.2~~	~~0.84~~	~~3.2~~	~~0.17~~						
~~ARP-3~~	~~778~~	~~0.12~~	~~0.85~~	~~3.6~~	~~1.93~~						
~~ECMWF-1~~	~~1068~~	~~0.15~~	~~Not available~~	~~Not available~~	~~6.2~~						
~~ECMWF-2~~	~~894~~	~~0.15~~	~~Not available~~	~~Not available~~	~~5.1~~						
~~ECMWF-3~~	~~772~~	~~0.15~~	~~Not available~~	~~Not available~~	~~4.8~~						
~~ECMWF-4~~	~~510~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.65~~						
~~ECMWF-5~~	~~491~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~						
~~ECMWF-6~~	~~463~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.88~~						
~~ECMWF-7~~	~~282~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.65~~						
~~ECMWF-8~~	~~314~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~						
~~ECMWF-9~~	~~325~~	~~0.15~~	~~Not available~~	~~Not available~~	~~0.62~~						

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