First of all, we would like to thank again the reviewers for their time spent on reviewing our manuscript and their comments helping us improving it. Please see below our point-by-point replies (reviewer's comments are displayed in black, our replies in blue font).

#Reviewer 1

The authors have generally addressed some concerns raised during the first review and have conveniently neglected to mention other concerns […] The authors may choose to not revise the manuscript if they disagree with reviewer suggestions, but they should at least say why.

The paper could be greatly strengthened by looking at why the model results improve, or at least by providing additional information to help readers gain context. This could be done by looking at the energy budgets.

A: We fully agree with the reviewer. This work relies on a collaborative group of simulations. For taking a look at the energy balance, we would need a number of variables that were not (unfortunately) stored by the contributing groups to this simulations.

Another way the paper could be improved is by looking beyond daily values. Only looking at daily values hides a lot of model deficiencies. Comparing the models against hourly temperature data, as well as moisture and PM2.5 amounts, would provide much more detail for understanding why the models change when including aerosol feedbacks. This would also bring the plume behavior of local aerosol sources more into play.

A: The reviewer is right. E-OBS database included only daily data. Further discussion on the observational database is presented below. As we really think the reviewer comment is very appropriate, we have included hourly evaluation in a manuscript that was about to be submitted to ACP (discussion of AOD, PM2.5 evaluation, by using hourly databases), whose submission will wait for the inclusion of hourly data.

The concern was raised that the authors compare the models to one gridded temperature dataset, and differences between datasets could be bigger than the differences shown due to the aerosol impacts within the models […] If the authors do not wish to add a comparison with a second dataset to take observation uncertainty into account, they should at least add text to the manuscript that puts the observations into context and note the limitation of the current study due to the single dataset.

A: The reviewer is right. We have clarified that in the manuscript. An effort has also been made to change the text just to highlight the fact that we have used only one dataset and its limitations. Section "Observational database" has been changed as follows:

The comparison of regional models with gridded datasets has to be carefully taken into account given the differences between available databases. For instance, Gómez-Navarro et al. (2012) showed that even in areas covered by dense monitoring networks, uncertainties in the observations are comparable to the uncertainties within state-of-the-art regional climate models, at least when they are driven by nominally perfect boundary conditions like reanalysis.

This work uses the E-OBS (Haylock et al., 2012) version 11.0 gridded observational database for maximum, mean and minimum temperature. E-OBS is a high-resolution European land-only daily gridded data set covering the period 1950-2014. The E-OBS 0.25 degrees regular latitude-longitude grid has been used as the reference for validation. Thus, data from all model runs have been bilinearly interpolated onto the E-OBS grid. Since the resolution of the models is similar to that of E-OBS, the interpolation procedure is not expected to alter significantly our results.

The election of this gridded dataset is based on the abundant scientific literature using E-OBS for the evaluation of regional climate models (e.g. Costa et al., 2012, Jiménez-Guerrero et al., 2013; Turco et al ., 2013; Ceglar et al., 2014), among many others). However, several authors highlight the E-OBS limitations. In this sense, Kysely and Plavcova (2010} compare E-OBS and a data set gridded onto the same grid from a high-density network of stations in the Czech Republic (GriSt), finding that large differences existed between the two gridded data sets, particularly for minimum temperatures and diurnal temperature range. The errors tended to be larger in tails of the distributions. Therefore, when evaluating regional models against one gridded dataset, results have been to be carefully taken into account.

The authors attempted to clarify the issue of what it means to turn aerosol-cloud interactions on and off within a model. Text has been added on p. 5 [..] However, as phrased this is a bit confusing. Please clarify. The sentence talks about aerosol assumptions and then provides a cloud droplet number concentration. [..] The authors should double check with each modeling group using WRF to identify how they chose to not have aerosol-cloud interactions. This can either be done by not compiling in "chemistry mode" and then one gets the 250 cm-3 droplet number concentration for Morrison microphysics. Or, one can compile with the chemistry mode turned on but not use an aerosol module. The latter sets a constant aerosol

number concentration (naer in the namelist) instead of a cloud droplet concentration. This is important because the physical processes related to activation and cloud formation change depending on the mode used.

A: The reviewer is right. We have checked with all WRF-Chem groups that the number we have given is done by not compiling the chemistry mode and getting 250 cm^{-3} droplet number concentration for microphysics in C1X and C2X. We have rephrased our sentence to clarify that in the manuscript. So the sentence remains as:

Although the NRF case does not consider the aerosol effects and feedbacks, this configuration considers an assumption of 250 cm^{-3} used by WRF-Chem in the absence of ACI for estimating cloud droplet number. This number is used in the corresponding microphysics parameterization (Morrison or Lin).

p. 13, l. 16–19: The authors claim the following sentence has been corrected, but it still does not make sense: "In general, coefficients of determination are highest for mean temperature (0.60 to 0.78) and lowest for minimum temperature (0.50 to 0.56), presenting the ensemble always maximum values for ρ2 (0.75, 0.79 and 0.61, respectively for maximum, mean and minimum temperature)." It is unclear what is meant by "presenting the ensemble always maximum values."

A: Here we meant that the coefficients of determination for TEMP are higher than those found for TMAX or TMIN. The lowest coefficients are estimated for TMIN when compared to the other variables. Moreover, the coefficients of determination for the ensemble are always higher when compared to the ρ^2 of the individual models. This is found for the 3 studied variables. We have rewritten the sentence in this sense:

In general, coefficients of determination are highest for mean temperature (ranging from 0.60 to 0.78 depending on the individual model) with respect to minimum and maximum temperature. The variable with the lowest ρ^2 is minimum temperature (varying from 0.50 to 0.56 depending on the model). Moreover, the coefficient of determination for the ensemble is always higher than that of each individual models for the three studied variables (0.75, 0.79 and 0.61, respectively for maximum, mean and minimum temperature).

Figure 1 is blurred and unreadable.

A: The reviewer is absolutely right, but this is related to the quality of the images has been reduced because of the size of the original pdf file containing high-

quality figures (over 200 MB). When final submission is done, the original pdf file for Figure 1 will be uploaded.

p. 4, l. 19: on-line coupled models simulations

A: Changed as suggested.

p. 5, l. 20: Although the NRF case

A: Changed as suggested.

p. 10, l. 6: have a notion of the aerosol loading (it would be better to reword to not use the colloquial phrase "have a notion" and replace it with "have an understanding of")

A: Changed as suggested.

p. 10, l 8ff: The sentence starting with "Despite the work of Palacios-Pena et al." is phrased poorly. It would be better to refer to the other work for full details and to say the current article provide brief details for context.

A: Changed as suggested.

#Reviewer 2

Figure 1 shows that the dust case has low values of AOD. Why is this a good case for the objective of the paper which is to show the impact on surface temperature?

A: The reviewer is right. The dust episode has not as high AOD levels at 550 nm as the Russian Forest Fires, but the aerosol loads are comparable in magnitude. Dust extinction is more noticeable at higher longwaves because of the size distribution of the dust particles. However, it does not mean that the aerosol load is low. We would have the same issue whatever dust episode we would select. The dust case was selected since it was also a humid episode with some rain over the Mediterranean and the dust plume effects extended over Europe.

The spatial panels in Figures 1 - 10 are quite difficult to read, are they really necessary, all of them?

A: The reviewer is right, in this sense we have added information with additional tables (Table 2, 3, 4 and 5) as well as information on the text. We considered that it was also interesting to see the Base case, E-OBS levels as well as differences between models.

The authors have chosen not to show specific areas where the clouds and/or plume impact surface temperatures. This would enhance the value of the article and make the point more clearly.

A: In order to follow the reviewers's suggestion, two new tables (Table 3 and Table 5) have been added to the final version of the manuscript. These tables summarize the results of those specific areas where the plume would have higher impacts on surface temperatures. For that, we have calculated the bias and the coefficient of determination by masking those timesteps and areas when 1-hr AOD550 > 1.0 for the fires episodes and AOD550 > 0.5 for the dust case. The results obtained are very similar and do not modify the previous discussion presented in the manuscript. However, because of its interest, the information has been introduced in the revised version of the manuscript.

#Reviewer 3

Page 2, Line 22: "…which are dependent…" to "… which is dependent…"

A: Changed as suggested.

Page 10, Line 09: "…fully compiled the AOD evaluation against diverse satellite observations of the ensemble considered in this work…" to "…fully compiled the evaluation of AOD of the ensemble considered in this work against diverse satellite observations …"

A: This sentence here has been changed following the Reviewer#1 suggestion.

Need to include a reference to MODIS AOD product (e.g. Levy et al., 2013)

A: The following reference has been added.

Levy, R., Mattoo, S., Munchak, L., Remer, L., Sayer, A., Patadia, F., and Hsu, N.: The Collection 6 MODIS aerosol products over land and ocean, Atmospheric Measurement Techniques, 6, 2989-3034, 2013.

Page 10, Line 18: "a low overestimation …" to "a lower overestimation …"

A: Changed as suggested