Answers to the Referees regarding the paper named "Air quality forecast at kilometer scale grid over Spanish complex terrains" by M.T. Pay et al.

Response to Interactive Comment by Anonymous Referee #1

Authors: We would like to thanks the Referee #1 for his/her constructive remarks and suggestions. All his/her comments have been implemented and commented accordingly in the reviewed version of the manuscript.

Please, find in the next paragraphs answers to Referee #1.

Referee #1: I would need to know if model outputs and stations observations are instantaneous every hour, or if they are integrated in time in some way. If an inconsistency exists in the temporal sampling, one could argue that 4km is a more sensible horizontal scale than 1km, therefore the 1km model outputs should be degraded somehow to reach the spatial and temporal representativity of the station.

Authors: In our comparison both modelled and measured concentrations are hourly averaged. In the case of the CMAQ, the model provides an output file (named ACON*) with hourly averaged concentrations. Concerning observations, which are received in near-real time, the measurements come from automatic monitoring networks, which are hourly averaged by the people that manage those networks.

We have included a comment on that in section 2.4 as follows:

> *"Representativeness continue to be a challenge when comparing gridded simulations to observational data at a point in time and space, as modeled concentrations represent a volumetric average over an entire grid cell. Furthermore, the stochastic compound embedded in the observations is not accounted for. Concerning temporal representativeness, in the present comparison both modeled and measured concentrations are hourly averaged. […]"*

Referee #1: On a similar topic, the discussion in Section 3 on spatial representativeness is interesting overall, but the reader keeps wondering what support the statements on how realistic are 1km and 4km maps given that we do not have such high resolution data to compare with.

Authors: The realism of the 1 km and 4 km simulations is difficult to evaluate because there are no 2D observations at 1-4 km resolutions. However, the comparison of 1 km and 4 km concentration maps shows that roads are easier identified and better shaped at 1 km than at 4 km. In this sense, we have replaced "better textured"/"significantly better textured" by "more easily identified"/"more textured".

Furthermore, in order to quantify the spatial representativeness of the concentration maps at both resolutions, we have calculated spatial correlations between modelled (1 km and 4 km) and observed concentrations at available air quality stations. The results indicate an increase of NO_2/O_3 spatial correlation coefficients from 0.79/0.69 (4 km) to 0.81/0.73 (1 km).

Referee #1: It is not clear why the evaluation period is so short. If the forecasting system is operational since 2009 for two of the selected areas, one could have expected a more comprehensive validation.

Authors: Although CALIOPE has been forecasting air quality at 1 km resolution over Madrid and Barcelona since 2012, forecasts over Andalucia domain started in 2013. With the aim of evaluating the resolution effect over the most populated areas with complex terrains in Spain (Barcelona, Madrid and Andalucia domains), we selected the most interesting period available by the time we started the present study, which was April 2013 (one month). From the climatological point of view, April is usually effected by transitional synoptic circulations (Valverde et al., 2014), but several exceedances of European limit values for O_3 and NO_2 in April 2013 justify its interest.

As the Referee #1 points out, a more comprehensive evaluation could cover for instant a full year. In this sense, an annual evaluation (September 1st 2011-September 1st 2012) for the Barcelona domain has been already discussed in Baldasano et al. (2013) and the results are in accordance with the present work. Anyways, in a future analysis we will expand the period of the analysis to a full year over the three domains.

Baldasano, J. M., Arévalo, G., Pay, M.T., and Gassó, S.: Influence of horizontal grid resolution on air quality modelling systems in Barcelona Metropolitan Area (Spain), in: 15th HARMO, Madrid, Spain, 6-9 May 2013, 2013.

Valverde, V. V., Pay, M. T., and Baldasano, J. M.: Climatic synoptic classification over the Iberian Peninsula oriented to air quality dynamic characterization, Int. J. Climatol., submitted, 2014.

Referee #1: P2295 L21: The author may consider relevant to add a couple of sentences on the need to reach high resolution in order to improve covariance between population and pollution for health impact assessment, e.g. as done in Thompson, T. M., Saari, R. K., and Selin, N. E.: Air quality resolution for health impacts assessment: influence of regional characteristics, Atmos. Chem. Phys. Discuss., 13, 14141-14161, doi:10.5194/acpd-13-14141-2013, 2013.

Authors: We appreciate the contribution from Referee #1. A comment about the benefits of the resolution increase for the health impact studies will be included in the revised manuscript as follows:

> *"Nowadays, fine horizontal resolution is a persistent challenge when assessing of health impact and population exposure studies (Thompson et al., 2013)."*

Referee #1: P2299 L12&14 : the use of "such as" in this context is surprising.

Authors: We have replaced this sentence from:

> *"AND includes one of the five biggest cities in Spain such as Seville (~ 700 000 inhabitants) and important industrial areas devoted to industrial processes, electric generation and maritime traffic such as Strait of Gibraltar."*

by

"AND includes one of the five biggest cities in Spain, Seville (~ 700 000 inhabitants) which host industrial areas and electric generation activities around the Algeciras bay, and it is affected by dense maritime traffic through the Strait of Gibraltar."

Referee #1: P2303 L 3-6 : in an evaluation paper, it is acceptable and relevant to spend a few lines to introduce the evaluation metrics rather than using references.

Authors: We agree with the Referee #1. We have created an Appendix A with two tables which include the description of the statistics used in this paper, both discrete (Table A1) and categorical statistics (Table A2). The Appendix A has been referred accordingly along the manuscript.

Referee #1: P2304 L16 : "desert"

Authors: Following the reviewer suggestion, the word "dessert" has been replaced by "desert".

Referee #1: P2308 L12: PM10 composition data is probably not available over the domains of interest. A reference to other studies having validated the CALIOPE system for individual PM compounds would be interesting. In particular, the abundance of SOA seems small, does it comply with the average load in Spain?

Authors: Measurements of PM_{10} components for 2013 are not available for the study domains. However, Pay et al. (2012) have already evaluated the PM components at some Spanish urban and rural background stations using the CALIOPE-AQFS based on CMAQv4.5. They showed that the model underestimated the secondary inorganic aerosol by a factor 2-3. The highest underestimation was found for fine carbonaceous aerosols (factor of 4) in part related to the state-of-the-science concerning secondary organic aerosol formation pathways. Based on these results, we can say the SOA in the present work could be underestimated. However, the CMAQv5.0.1 used in the present work includes substantial scientific improvements over the version 4.5, especially devoted to improving SOA formation and dynamic interactions of fine and coarse aerosol.

According to the Referee #1's suggestion, some comments about the CALIOPE-AQFS performance for PM components over Spain have been included in section 4.2 (PM10 components) as follows:

"Pay et al. (2012) already evaluated the PM components at some Spanish urban and rural background stations using the CALIOPE-AQFS based on CMAQv4.5, and they showed that the model underestimated the secondary inorganic aerosols by a factor of 2-3. The highest underestimation was found for fine carbonaceous aerosols (a factor of 4), in part related to the state-of-the-science concerning secondary organic aerosol formation pathways. The updated version of CMAQ, v5.0.1, includes scientific improvements concerning SOA formation and aerosol dynamics which could improve the modeled PM10 performance for its components."

Referee #1: P2308 L 17&18: replace "in" for "by".

Authors: The correction has been amended.

Referee #1: P2308 L26: a word is missing between “wind speed” and “relative humidity”

Authors: The correction has been amended.

Referee #1: P2308 L27 “not shown”

Authors: The correction has been amended.

Referee #1: P2309 L 3: what is the reason for the change in primary PM load with resolution? One can expect increases in horizontal gradients reported later in the same paragraph but the change in total abundance is more surprising.

Authors: The increase of primary PM concentrations when increasing resolution is due to the fact that the 1 km simulations allocate emission in a lower grid cell, which leads to a reduced effect of artificially dilution of emissions, so near high emission sources the concentration gradients could be stronger than at 4 km simulation.

However, as the Reviewer #1 points out the PM10 concentration increase when increasing the resolution is not in the same proportion as for primary pollutants. This is a result of a bias compensation of PM10 components, mainly controlled by the PPM and the EC concentration increase and the SS concentration decrease when increasing resolution. This has been discussed in the manuscript as follows:

> *“For primary PM components (EC and PPM) increasing resolution presents the highest increase in concentration (by 10 and ~12%, respectively). As for NO_2, the 1-km simulation leads to a reduced effect of artificial dilution of emission in a grid cell, so concentration gradients are stronger than in the 4-km simulation.”*
>
> *“The grid effect is less pronounced for PM_{10} than for NO_2 and O_3. When the resolution increases, the low increment of PM_{10} mean (<0.1 μgm^{-3}) is the result of compensating biases of PM_{10} components, which is controlled mainly by the PPM and EC increase as well as and the SS decrease.”*

Referee #1: P2309 L 15 : “increase on daily cycles”

Authors: The correction has been amended.

Referee #1: P2309 L19: please clarify what is referred to as “lamination” of the PBL.

Authors: This concept of the “lamination of the PBL growth by the Mediterranean sea breeze” makes reference to the entrance of the on-shore flow that leads to a reduced mixing height (Perez et al., 2004; Millan et al. 1997). Millán et al. (1997) have already documented the first rapid rise of the mixing height during the morning followed by the sinking of its capping inversion during the afternoon in the Mediterranean coastal area. Sicard et al. (2006) and Perez et al. (2004) also measured this phenomenon in Barcelona area using LIDAR.

Millán, M., Salvador, R., Mantilla, E., and Kallos, G.: Photooxidant dynamics in the Mediterranean basin in summer: results from European research projects, J. Geophys. Res., 102, 8811– 8823, 1997

Pérez, C., Nickovic, S., Baldasano, J.M., Sicard, M., Rocadenbosch, F., and Cachorro, V.E.: A long Saharan dust event over the western mediterranean: Lidar, sun photometer observations, and regional dust modeling, J. Geophys. Res. 111 (D15214), 1-16, 2006.

Sicard M., C. Pérez, F. Rocadenbosch, J. M. Baldasano, D. García-Vizcaino: Mixed-Layer Depth Determination in the Barcelona Coastal Area From Regular Lidar Measurements: Methods, Results and Limitations. Boundary-Layer Meteorology, 119, 1, 2006.

Referee #1: P2310 L8: what is the dynamical process leading to a lower PBL in the high resolution simulation?

Authors: The PBL height diurnal cycle has not been evaluated because there are not available measurements. However, comparison of PBL at both resolutions has been performed in order to find potential reasons of pollutant concentration differences between resolutions.

The 1 km resolution displays a lower PBL height than 4 km simulation in the morning after the sunrise and in the evening after the sunset. The reason of this behavior could be a result of features depending on topography like temperature, wind field and mesoscale sea-breeze and mountain-valley circulations. In this sense, some meteorological fields such as wind speed at 10 m (U10), wind direction (WD10) and temperature at 2 m (T2M) have been evaluated when increasing resolution from 4 km to 1 km (Sect. S1, http://www.geosci-model-dev-discuss.net/7/2293/2014/gmdd-7-2293-2014-supplement.pdf). Overall, comparison with METAR reveals that the resolution increase slightly improves T2M (bias in 0.1ºC), U10 (bias in 0.1 ms^{-1} and r in 0.1) and WD10 (error in 52º and r in 0.1). However, it slightly decreases WD10 bias (in 2º).

According to Fay and Neunhäuserer (2006) high resolutions (ranging from 1 to 5 km) are essential to reproduce mesoscale phenomena, e.g. those controlling O_3 transport along the mountainous northeastern Mediterranean coast where features depending on topography like temperature, wind speeds, channelling, convergence/divergence lines and mesoscale circulations are better described.

Referee #1: P2310 L 19-24: which additional measurement or modelling experiment could lead to a better understanding of the reason for this diurnal cycle in the model bias?

Authors: Our proposal to go more in detail with the PM10 underestimation during the daily cycle is evaluate the modelled PM10 components with hourly measurements in order to identify if the underestimation come from primary or secondary aerosol. Additionaly, it could be desirable evaluate the PBL height on an hourly basis to check if the model is reproducing the mixing height properly. For instant, high temporal resolution of PBL high from LIDAR measurements can be useful to evaluate modelled PBL.

Answers to the Referees regarding the paper named "Air quality forecast at kilometer scale grid over Spanish complex terrains" by M.T. Pay et al.

Response to Interactive Comment by Anonymous Referee #2

We would like to thanks the Referee #2 for his/her comments which have contributed to increase the quality of the present work. On the one hand, the manuscript has been improved in terms of writing and grammar after a review from a native speaker. On the other hand, all the specific remarks and discussion from Referee #2 have been implemented in the reviewed manuscript.

Please, find in the next paragraphs answers to Referee #2.

Referee #2: While I do agree with the other reviewer's comment regarding the length of the analysis being presented (one month), given the amount of effort required to perform a thorough analysis of the data for multiple domains and grid resolutions, the short duration does not, in my opinion, significantly harm the analysis presented. However, it does make it impossible to make any general, sweeping conclusions regarding the performance of 4km vs 1km grid resolutions, since model performance can change significantly throughout the year (and from year to year as well). I don't believe the authors make any of these types of generalized conclusions, so that is not an issue. Perhaps in the future the analysis could be extended to a longer time period (perhaps cutting down on the number of domains analyzed).

Authors: As mentioned in the answers to Referee #1, the reason why we selected the present period to study the resolution grid effect is based on the availability, by the time we started the present study, of CALIOPE-AQFS simulations at 1 km resolution during an interesting period in terms of air quality over the three study domains (AND, BCN and MAD).

As the Referee #2 points out, a more comprehensive evaluation could cover for instant a full year. In this sense, an annual evaluation (September 1st 2011-September 1st 2012) for the Barcelona domain has been already discussed in Baldasano et al. (2013) and the results are in accordance with the present work. Anyways, in a future analysis we will expand the period of the analysis to a full year over the three domains.

Baldasano, J. M., Arévalo, G., Pay, M.T., and Gassó, S.: Influence of horizontal grid resolution on air quality modelling systems in Barcelona Metropolitan Area (Spain), in: 15th HARMO, Madrid, Spain, 6-9 May 2013, 2013.

Referee #2: And incommensurability between observations and model values will always be an issue, and should probably always be noted, as the comparisons being made are between point observations and grid volume concentrations. But noting whether the measurements are instantaneous values or hourly average values would be useful (same goes for the model values).

Authors: We agree that representativeness challenges continue to be present whenever gridded simulation are compared to observed data at a point in time and space as modelled concentrations represent a volumetric average over an entire grid cell, meanwhile the stochastic compound embedded in the observations is not accounted for. Measurements have

their own uncertainty due to biases and artifacts related to sampling and laboratory analysis methods. The European legislation (2008/50/EC) requires that the uncertainty of measurements meet the air quality objective of 25% for PM10 and PM2.5 and 15% for O_3, NO_2, and SO_2.

As mentioned in the answers to Referee #2 concerning temporal representativeness, in the present comparison both modelled and measured concentrations are hourly averaged. In the case of the CMAQ, the model provides an output file (named ACON*) with hourly averaged concentrations. Concerning observations, which are received in near-real time, the measurements come from automatic monitoring networks, which are hourly averaged by the people who manage those networks.

We have included a comment on that in section 2.4 as follows:

> *"Representativeness continue to be a challenge when comparing gridded simulations to observational data at a point in time and space, as modeled concentrations represent a volumetric average over an entire grid cell. Furthermore, the stochastic compound embedded in the observations is not accounted for. Concerning temporal representativeness, in the present comparison both modeled and measured concentrations are hourly averaged. […]""*

Referee #2: P2294L4: Define "main pollutants" here.

Authors: The suggestion has been included as follow:

> *"It provides a 48-h forecast of the main pollutants (NO_2, O_3, SO_2, PM_{10}, $PM_{2.5}$, CO, and C_6H_6) found at a 4-km horizontal resolution over all of Spain […]"*

Referee #2: P2294L12: Replace "in" with "by". This change applies to the entire article.

Authors: The correction has been amended in the whole manuscript.

Referee #2: P2295L23: Define CAMx and CMAQ here. I don't believe they have been defined yet.

Authors: CAMx and CMAQ have been defined in the revised version of the manuscript as follows:

> *"Several studies have evaluated the impact of increasing horizontal resolution on different scales over the eastern and southeastern USA using the Community Multiscale Air Quality (CMAQ) model and the Comprehensive Air Quality Model with Extensions (CAMx), which range from 32 km – 12 km – 4 km [...]"*

Referee #2: P2296L28: What is meant by "larger spatial concentration"?

Authors: The original sentence:

> *"[…] bottom-up emission inventories provide better performance and larger spatial concentration than down-scaled inventories."*

Has been replaced by:

> *"[…] the predicted concentrations and corresponding gradients which are more consistent with observed concentrations when provided by bottom-up emission inventories rather than those from down-scaled inventories."*

Referee #2: P2297L16: Define OPANA.

Authors: OPANA means OPerational Atmospheric Numerical model for urban and regional Areas. The definition has been included in the revised manuscript as follows:

> *"The lowest resolution system is the Technical University of Madrid's OPANA (OPerational Atmospheric Numerical model for urban and regional Areas), running at 27 km x 27 km […]"*

Referee #2: P2298L23: I assume the numbers provided in parentheses are the length of the mountain ranges and not height. That needs to be made clear in the text.

Authors: The numbers between parentheses are the height of the mountain ranges. I have clarified it the text as follows:

> *"BCN is a coastal area characterized by several valleys perpendicular to the coastline and two main mountain ranges, one coastal (500 m height) and one pre-coastal (1000-1700 m height). These features induce mesoscale phenomena such as sea-breeze and mountain-valley winds."*

Referee #2: P2298L25: What is meant by "Central System"?

Authors: The Central System is one of the main mountain ranges in the Iberian Peninsula (2.592 m height). It has been defined in the manuscript accordingly as follows:

> *"MAD is a continental region with a much simpler topography that includes the Tajo valley in the southern of MAD and the mountain range of the Central System located in the northwestern MAD, with summits reaching 2500 m height. These features brings different locally-driven flows."*

Referee #2: Figure 1: In Figure 1 the domains are labeled d1-d5, but here they are named. They should be made consistent.

Authors: In Figure 1 the D-domains make reference to the nested sequence of the simulated domains starting from the mother domain (D1, Europe). However, the acronyms AND, MAD and BCN correspond to the study domains (Andalucia, Madrid and Barcelona, respectively). For instant, the impact of horizontal resolution increase in terms of air quality concentrations over the AND domain is analyzed using the simulations from D2 (IP - 4 km) and D3 (AND - 1 km), which in the rest of the paper is referred as 4 km and 1 km simulations, respectively.

This fact has been clarified in the caption of Figure 1 as follows:

> *"CALIOPE-AQFS nesting strategy (D-domains) and study domains (Andalucia, AND; Madrid, MAD; and Barcelona, BCN)."*

Referee #2: P2299L6: What is meant by “logistic”?

Authors: In this sense, “logistic” means “commercial activities”. This error has been amended.

Referee #2: Section 2.2: What land-use data is used in the WRF simulation?

Authors: As indicated in the Section 5 (Conclusion), the WRF uses the land-use data from the U.S. Geological Survey (USGS) which is based on the year 1993. This fact has been mentioned in the Section 2.2 as follows:

> *“The Noah land-surface model (NoahLSM), based on the U.S. Geological Survey’s (USGS) land-use data is used by default in the present WRF configuration.”*

Referee #2: P2300L25: After collapsing, how many CMAQ layers are in the PBL?

Authors: Six CMAQ sigma layers cover the PBL. It has been implemented in the manuscript as follows:

> *“[…] meanwhile CMAQ vertical levels are obtained by collapsing from the 38 WRF levels to a total of 15 layers that steadily increase from the surface up to 50 hPa. Six layers are within the PBL, and the first layer depth is 39 m.”*

Referee #2: P2301L4: Why is the reference here to the previous version of CMAQ stated as v4.5? The previous version of CMAQ before 5.0 is 4.7 (and before that 4.6).

Authors: The comparison between CMAQv4.5 and CMAQv5.0 is based on the CALIOPE-AQFS progress. It has been clarified in the revised manuscript as follows:

> *“Based on the evaluation results from the previous CMAQ version within CALIOPE-AQFS (4.5 vs. 5.0) (Pay et al. 2012b), CMAQ has been updated to Version 5.0.1, using the CB05 chemical mechanism (Yarwood et al., 2005), the AERO5 for aerosol modeling, and the in-line photolysis calculation.”*

Referee #2: P2301L8: Why use AERO5 and not AERO6?

Authors: The science devoted to the speciation of PM aerosols has significantly improved in AERO6 compared to AERO5 including, for instance, speciation of PM fraction (including Fe, K, Mg, Ca and Ti), primary organic aerosol aging, and some updates in the sulfur chemistry. However, we have decided not to use the AERO6 module for two reasons:

1- The computation cost and the size of the input files (in the emissions) and output files significantly increase in AERO6 because of the increase of the number of species. These are critical issues when working with high resolutions in a forecast mode.
2- There are no specific emission profiles to speciate PM fine emission to new ion species in Spain.

Referee #2: P2301L24: What is meant by “300 min”?

Authors: This 300 min is the computational time used to simulate 48 h of meteorology, emission and air quality.

Referee #2: P2301L27: What is meant by “soft reservation”? Section 2.3: I’m not sure how much value this section adds to the manuscript. Every group uses a different computer configuration for their modeling efforts, so these numbers are really unique to your modeling exercise. While some readers may find the information interesting, I think most readers will not find the information overly useful. If a strong argument can be made for keeping the section, then fine. Perhaps it could be consolidated into a single paragraph however.

Authors: In this case, a soft reservation means a special book over the whole supercomputational resources to be sure that the forecast will run with a sufficient number of CPUs.

I agree with the Referee #2 that this computational setup is unique according to our resources and objective. However, the increase of computational resources and horizontal resolution in forecast issues requires of this kind of setups. In this sense, I have kept this section, but it has been synthetized as follows:

> *“Running CALIOPE-AQFS at 4 and 1 km is a technical challenge. The simulations are run on MareNostrum supercomputer (Intel Xeon E5-2670, 16 CPUs and 64 GB RAM memory per node) at BSC-CNS. Table 1 depicts the computational requirements for forecasting air quality at 48 h for each domain. The number of CPUs was chosen to maximize CPU efficiency. Thanks to the parallelization of meteorological and air quality models, MareNostrum uses up to 256 CPUs. Due to the variable nature and complex dependencies, the computational time for forecasting 48 h of air quality fields for the 4 domains is 8-9 hours. The most computational demanding domain is the AND, at 1-km resolution (366x358 cells, 256 CPU max., and 300 min). For the April 2013 simulation, times add up to 2880 CPU hours/day, or 86400 CPU hours in one CPU (9.86 years). The storage for the April 2013 output files was 6.13 TB (~200 GB/day).”*

Referee #2: P2302L12: I assume the 1 ug/m3 is a MINIMUM cutoff. That should be made clear.

Authors: Yes, 1ug/m3 is a minimum cutoff. It has been clarified in the manuscript.

Referee #2: P2302L16: Define METAR.

Authors: METAR means METeorological Aerodrome Report. It has been included in the revised manuscript.

Referee #2: P2303L14: What is meant by “considering the 75% of the values”? It’s not really clear here.

Authors: It has been clarified as follows:

> *“Note that mean and maximum concentrations are calculated by considering at least 75% of the data in the corresponding time base”.*

Referee #2: P2303L25: What is meant by “maps are conserved”?

Authors: The sentence means that when the resolution increases from 4 km to 1 km, the O_3 concentration pattern are similar. However, slightly differences appear along areas with high

NO_x emission where titration processes are very active. This senesce has been rewritten in the revised manuscript as follows:

> *"Consequently, when the resolution increases the monthly mean O_3 concentration maps are almost identical, although the NO_x titration effect on O_3 is significant along highways and major point sources"*

Referee #2: P2304L8: Not sure the language "significantly better textured" is appropriate here. I think the authors are just trying to indicate that the roadways are more easily identified and better defined in the 1km simulation than the 4km, but that doesn't necessarily mean they are "better".

Authors: We agree with the Referee #2, we cannot say they are "better" because they have not been compared with 2D observations yet. We are trying to say that 1 km simulation allows to more easily identified roadways and even mountains. In this sense, we have replaced "better textured"/"significantly better textured" by "more easily identified"/"more textured".

Referee #2: P2305L25: The authors need to be consistent with their language when describing the results. Here, the authors state "monthly r slightly decreases when resolution increases from 0.67 to 0.58". That's a difference of 0.09. However, just above the authors state "slopes significantly improve with resolution increase from 0.72 to 0.77 for NO2 and from 0.50 to 0.54". Both of those increases are much less than the decrease for PM10. The authors need to be fair here and use consistent language instead of highlighting the improvement as "significant " and the degradation as "slightly decreases".

Authors: We totally agree with the reviewer comment. We have commented the results from an objective point of view.

Referee #2: P2305L26: Surely the value here should not be 0.4 (I assume it should be 0.04).

Authors: The reviewer is right. It has been amended in the revised manuscript.

Referee #2: P2306L14: A lot of these values lack units. Units need to be added for all values where appropriate.

Authors: The units have been added accordingly.

Referee #2: P2307L11: I authors say "bias" but the values are in percent, so it must actually be some kind of normalized bias being presented.

Authors: The bias (B) for categorical evaluation is not exactly a real normalized bias. But it is expressed in %. Following the reviewer 1 suggestion a description of categorical statistics has been added in the revised supplementary material.

Referee #2: P2307L16: A number of times the incorrect abbreviation CIS is used instead of CSI.

Authors: It has been corrected.

Referee #2: P2310L3: A reference should be included here regarding the model performance for morning and evening transitions.

Authors: We agree with the reviewer, and some references supporting this issue have been provided. The sentence has been rewritten as follows:

> *"Simulations by photochemical modeling systems are known not to reproduce faithfully the morning hours after sunrise and the evening hours after sunset when the mixing height experiences rapid changes."*

by

> *"Several works indicates that WRF does not faithfully reproduce the morning and evening transition over urban environment, possibly because it does not model the heat retention in cities (Makar et al., 2006; Appel et al., 2013)."*
>
> *Makar, P. A., Gravel, S., Chirkov, V., Strawbridge, K. B., Froude, F., Arnold, J., and Brook, J.: Heat flux, urban properties, and regional weather, Atmos. Environ., 40, 2750–2766, 2006."*

Referee #2: P2314L10: Change "better captured" to "more evident". Also, the NO2 measurements are likely not made right on the roadway, so it's probably not possible to attribute the improvement in NO2 performance at finer resolution to only the roadways. If the NO2 measurements are made right at the roadways, it would be good to state this earlier in the text regarding the proximity of the NO2 measurements to the major roadways.

Authors: The change "better captured" to "more evident" has been implemented.

Concerning the NO_2 measurements, as it shown in Figure 2, most of the stations in the urban domains of BCN and MAD are classified according to Garber et al. (2002) as traffic stations, which means they are located at building up areas under the direct influence of traffic emissions. These stations can be located either at the roads or nearby the road. In the case of BCN and MAD where more than 60% of NOx emissions come from the on-road traffic, the improvement of the NO_2 performance at finer resolution could be attributed to a more comprehensive modeling of the emission and chemistry near traffic emissions.

Referee #2: P2316L20: This detail should be included earlier in the text. Also, why was such an old land-use data set employed? Using a more up-to-date land-use data set could improve the model results significantly.

Authors: We agree with the Referee #2. A description of the land-use data implemented in WRF has been included in the section 2.2.

We used the USGS land-used data because the WRF works by default with this kind of categories. For next improvement of CALIOPE-AQFS, we have implemented the land-use data from a high resolution and updated data base called CORINE land cover following the methodology of Pineda et al. (2004) to do the assignation between categories in USGS and CORINE.

Referee #2: P2316L30: What is the CORINE data set? A very brief description of the data would be nice here.

Authors: The CORINE (Coordination of Information on the Environment) Land Cover (CLC) is the database about the coverage and use of the land in the European Union managed by the European Environmental Agency. The CLC has a resolution of 100 m and includes 44 land cover classes. The first inventory was based on 1990 (CLC1990), it has been updated to the year 2000 (CLC2000), and recently to 2006 (CLC2006).

We have included a comment on that in the revised manuscript as follows:

> *"Furthermore, in order to gain any benefits from increasing resolution, the meteorological modeling should include an improved description of the land instead of relying on USGS data from the year 1993. To this end, the Coordination of Information on the Environment (CORINE) provides a high resolution (100 m) land use database, which was developed by the European Environmental Agency and updated to the year 2006 (CLC2006) (EEA, 2007). This could be implemented in the WRF model following the methodology described in Pineda et al. (2004). "*
>
> *"EEA: CLC2006 technical guidelines. EEA Technical Report 17/2007. ISBN 978-90-9167-968-3. doi 10.2800/12134, 2007."*

Answers to the Referees regarding the paper named "Air quality forecast at kilometer scale grid over Spanish complex terrains" by M.T. Pay et al.

Response to Interactive Comment by Anonymous Referee #3

We would like to thanks the Referee #3 for his/her constructive remarks and comments. All his/her comments have been implemented and discussed accordingly in the reviewed version of the manuscript.

Please, find in the next paragraphs answers to Referee #3.

Referee #3: I would prefer to add a short paragraph on the impact of the (resolutionof the) meteorology on the model results. From the paper it is not quit clear at what resolution the meteorological input is used. If in particular the resolution goes down to 1 km, the local meteorological phenomenae become important.

Authors: Following the same nesting strategy as for the air quality simulations (see Figure 1), the meteorological fields from WRF are first simulated at 12 km x 12 km over Europe (the mother domain) using the GFS global data for boundary and initial condition. By means of a one-way nesting, the WRF simulates meteorology at 4 km x 4 km horizontal resolution over the Iberian Peninsula and at 1 km x 1 km over the study domains (Andalucia, AND; Barcelona, BCN and Madrid, MAD). The WRF model configuration and set-up is further described in Section 2.2.

The impact of the resolution increase is also analyzed in terms of meteorological parameters. Indeed, the meteorological fields are evaluated for wind speed at 10 m (U10), wind direction (WD10) and temperature at 2 m (T2M) at 10 METAR stations located at airports (6/2/2 stations in AND/BCN/MAD. However, due to the long extension of the paper we have decided to move this discussion to the supplementary material (Sect. S1, http://www.geosci-model-dev-discuss.net/7/2293/2014/gmdd-7-2293-2014-supplement.pdf). Along the discussion some comments and link the meteorological performance are included.

Overall, comparison with METAR reveals that the resolution increase slightly improves T2M (bias in 0.1ºC), U10 (bias in 0.1 ms^{-1} and r in 0.1) and WD10 (error in 52º and r in 0.1). However, it slightly decreases WD10 bias (in 2º). High resolutions (ranging from 1 to 5 km) are essential to reproduce mesoscale phenomena, e.g. those controlling O_3 transport along the mountainous northeastern Mediterranean coast where features depending on topography like temperature, wind speeds, channelling, convergence/divergence lines and mesoscale circulations are better described (Fay and Neunhäuserer, 2006).

Referee #3: Related to the first point, and also mentioned by Referee# 1, is the issue of the spatial representativeness of the observations. Some clarification is needed in the paper.

Authors: Comments about spatial representativeness have been already discussed in answers to the Referee #1 and Referee #2.

Referee #3: For typographical errors I refer to the other referees.

Authors: Typographical errors mentioned by the Referees #1 and #2 have been amended accordingly.