Dear Dr. Nicholas Henry Savage,

We appreciate valuable comments, which have helped improve the paper. We revised the text according to the suggested corrections and would like to thank you for the thorough reading of the paper. Below we provide our point-by-point replies, where for clarity the comments are displayed in bold italics.

Title. As per the instruction of GMD, please include the version of WRF-Chem in the title.

Thank you for this reminder, model version is now included in the title: »Evaluation of the high resolution WRF-Chem (v3.4.1) air quality forecast and its comparison with statistical ozone predictions«.

Abstract. Please specify the resolution of the model configuration applied (high resolution is rather a relative term)

We included the information about model resolution in the first sentence of the abstract, which is now: »An integrated modelling system based on the regional on-line coupled meteorology-atmospheric chemistry WRF-Chem model configured with two nested domains with horizontal resolution 11.1 km and 3.7 km has been applied for numerical weather prediction and for air quality forecast in Slovenia.«

2.1 WRF-Chem forecast system. Please state the height of the model top.

The height of the model top is 50 hPa, this information is now included in the paper in the following sentence: » The vertical structure of the atmosphere is resolved with 42 vertical levels extending up to 50 hPa, with the highest resolution of ~25 m near the ground.«

Please provide a reference (even if it is only a report) for the emissions inventory.

We added the reference to the project presentation at Slovenian Environment Agency (report is not yet available).

2.2 Statistical ozone daily maximum forecast. Please provide references for the statistical model.

We added the reference to the final report about statistical model (also available online).

2.3 Evaluation methodology. What is the height of the lowest model level, and how does that compare to a typical inlet height?

We added this information to the paper the following way: »In the case of air pollutants, the instantaneous lowest model level mixing ratios (with grid point center about 12 m above model orography - an exception is KRV station as explained below) are compared to the hourly averaged concentrations measured at monitoring stations (which have a typical inlet height of 3 m) from the national network and some other environmental information systems in Slovenia. Figure 3 shows locations of these AQ monitoring stations, and Tab. 1 lists the basic characteristics, including comparison of the station altitude, the height of model orography, model analysis height, and pollutants with higher than 75% availability of valid data during the analyzed time period for each of the AQ monitoring site«

Have you considered using data from above level 1 - in very mountainous terrain, an observation site can be well above the model orography at the relevant grid point and it is more appropriate to use data from level 2 or above.

Thank you for this question. In the case of AQ variables we usually use results from a higher model level for the KRV station. The altitude of this station is well above model topography (model height: 1272 m, model grid point at the lowest level: 1284 m, station altitude: 1740 m). In the present paper we originally included results for all stations (also KRV) at the lowest model level, because the correlation coefficient at the lowest model level is highest (CORR decreases with increasing the model level), showing that in spite of the negative bias due to too low model topography, the near surface processes still play an important role in ozone dynamics. In the review process we reconsidered this and decided to use model data from the 5th model level for KRV (model grid point center: 1414 m), but stay with the lowest model level for all other stations. For KRV the 5th model level is still well below the station altitude, but this reduces the bias for KRV from -12 μgm^{-3} to -2 μgm^{-3} for ozone hourly values, and from -16 μgm^{-3} to -7 μgm^{-3} for ozone daily maxima (which lowers the impact of KRV bias on overall model performance). Unfortunately also CORR then decreases from 0.76 to 0.74 for ozone daily maxima (which has a negligible impact on overall model performance). For other stations the differences between model height and station altitude are smaller. Also for some of the stations model height is too low (e.g. VNA, model height: 468 m, station altitude: 630 m), but for other stations the model height is too high (e.g. HRA, model height: 540 m, station altitude: 290 m), related to very complex topography in sub-alpine region of Slovenia. Consequently, by increasing the model levels we could reduce the negative bias for stations of the first group (with too low model orography), but cannot decrease the positive bias for the stations of the second group with too high model orography. This makes an approach of using higher model levels for stations with too low model orography questionable, also in the light that also CORR decreases with increasing model levels. We thus support the approach of using the data on the lowest model level and make a posterior bias correction, which does not impact the ozone dynamics and can be applied for all stations. We only made an exception for KRV station, for which the height in the model was significantly underestimated, as well as the station is known to be influenced by the conditions of the free troposphere (except during hot summer daytime conditions), which is not the case for other stations.

For meteorological variables we did not explore the impact of using results from higher model levels. This would be far beyond the scope of this study, focused on ozone prediction, also because the impact of using the higher layer data depends on meteorological variable, as well as the set of meteorological stations is not the same as in the case of AQ stations.

In the paper due to using results for KRV on the 5th level we corrected all of the AQ statistics and also the text throughout the paper accordingly. We included the following text:

»In the case of the elevated alpine KRV station, AQ variables are evaluated for the 5^{th} model layer instead of the first model layer. We made this exception for KRV, since the height of the model topography was significantly underestimated there (Tab. 1), as well as the station is known to be strongly influenced by the conditions of the free troposphere. The selection of the 5^{th} model layer for KRV station is based on analyses performed for different model layers (results not shown) and was found to reduce the negative bias for O_3 due to too low WRF-Chem topography at this location.

Although even for this model layer the location of the grid point representing KRV station (1414 m) is still well below the true station altitude (1740 m), the O_3 bias for KRV station is significantly smaller than for the first layer, while the correlation coefficient between the measured and simulated O_3 levels remains similar in both cases (the 5th or the lowest model layer). Taking results from higher model layers would further decrease the negative model bias, but would also worsen the correlation coefficient for O_3 at this station due to decreased impact of surface processes.«

Later in text also:

Instead of: »The elevated alpine KRV station is the only one with negative bias (-12 μgm^{-3}) in forecasted 1-hour O_3 concentrations, which can be explained by the too low altitude of the KRV station in model topography, since the mean O_3 concentration increases with height.«

We added: » In Fig. 4a the elevated alpine KRV station is the only one with high negative bias (-12 μgm^{-3}) in forecasted 1-hour O_3 concentrations at the lowest model layer, which can be explained by the too low altitude of the KRV station in model topography. The high negative bias for hourly O_3 concentrations at KRV station is reduced to a value of only -2 μgm^{-3} by using the 5th model layer concentrations as explained in chapter 2.3. The 5th model level predictions will be used for KRV in all analyses that follow.

We added also: » For sites with highest positive bias in 1-hour O_3 concentrations (TRB, ZAG, HRA and ISK, with bias of 36 μgm^{-3}, 31 μgm^{-3}, 26 μgm^{-3} and 32 μgm^{-3}, respectively), this can also be partly explained by too high altitude of the stations in model orography (Tab. 1), since the mean O_3 concentration increases with height.«

Later in text we deleted: »or Alpine stations (KRV)« and added: » Here we recall that high negative bias in WRF-Chem forecast for alpine KRV site due to too low altitude of the station in model topography was compensated by taking prediction from the 5th model level.«

Also the values of statistics in text and figures are changed throughout the paper.

3.1 Evaluation of meteorological variables. There is a large decrease in the precipitation bias from day 1 to day 2 - is this a model spin up issue? If so would a different initialisation improve this error?

We agree. Additional circumstance here is also that in the 3.4.1 model version it was not possible to include the information about hydrometeors at the boundaries of the nested domain (in the applied 1-way nesting procedure). Since the intensity of (relatively rare) summertime precipitation events was expected to have a less significant impact on ozone concentrations, we considered this issue less problematic (in our study focused on ozone). We added the following text: “It must also be taken into account that the 3.4.1 model version does not allow to include the information about hydrometeors at the boundaries of the nested domain (in the applied 1-way nesting procedure), which contributes to the negative simulated bias of precipitation. A large decrease in the precipitation bias from day 1 to day 2 suggests that different initialization methodology (e.g. using 1 day spin-up for meteorology) could improve the prediction of precipitation events.”

Please provide some evidence for the statement "the main precipitation events were well predicted and simulated" or remove this statement.

Although we performed analyses and produced some plots we think that including additional material here is beyond the scope of the paper. We thus decided to remove this statement.

> ***3.3 Evaluation and comparison of different methods for O3 daily maximum predictions. Please correct the statement "ideal forecast would lie in the right-bottom corner". It fact the ideal model would have correlation coefficient of 1 and a standard deviation equal to the observations, i.e. it would be co-located with the black dot which indicates the model. The black dot is not always in the bottom right corner on these plots.***

Thank you, we corrected this statement. The statement that is now included is: » The ideal model would have a correlation coefficient of 1 and a standard deviation equal to the observations, which means that it would be co-located with the black dot on the diagram. «

> ***In the section on the evaluation of the model's ability to predict episodes, too much weight is given to accuracy. For example, the statement "Accuracy ... increases with threshold level" is misleading. A model which always forecasts "no event" will have an increasing accuracy as the number of events decreases. To compare skill at different thresholds you need to use a differnt metric e.g. Critical Success Index or Equitable Threat Score. These would be better choices in general than accuracy in this section. There is no harm in including accuracy in the tables, but it should not be the primary criterion for judging forecast skill.***

In the revised paper we replaced Accuracy (A) measure by Equitable Thread score (ETS), we also changed the order of categorical statistics in Tab. 5, so that ETS is shown in the first column, followed by CSI, B, FAR and POD. We corrected the text, to give most weight to the ETS and briefly mention the rest of them. The text that we now have in the paper regarding the categorical evaluations is the following: »Equitable Threat Score (ETS) measures the fraction of observed and/or correctly predicted events, adjusted for the frequency of hits that would be expected to occur by random chance. Although this score takes into account the climatology it is not truly equitable. It ranges from -1/3 to 1, where the minimum value depends on climatology (it is near 0 for rare events). Looking at Tab. 5 ETS shows equal skill for WRF-Chem and statistical forecast, higher than persistence for the 120 μgm^{-3} threshold (1-day and 2-day forecast). ETS decreases with increasing the threshold for both WRF-Chem and statistical forecast, indicating the challenge that both models have to accurately predict the extremes. In the case of 140 μgm^{-3} threshold, WRF-Chem has the same ETS as persistence, higher than the statistical model for 1-day forecast, while for 2-day forecast WRF-Chem outperforms the statistical model, followed by persistence. In the case of 160 μgm^{-3} threshold persistance has the highest ETS for a 1-day forecast, followed by statistical model and WRF-Chem, while in the case of 2-day predictions, statistical model shows the highest skill and WRF-Chem the lowest. Another measure, the critical success index (CSI), is similar to ETS, except that it does not take into account the climatology of the events and thus gives poorer scores for rarer events. It measures the percentage of cases that are correctly forecasted out of those either forecasted or observed, and ranges from 0 to 1 (1 indicating the perfect forecast). Similar as ETS, CSI gives higher scores for persistence in the case of 1-day forecast for the higher two thresholds, while on the second day WRF-Chem or the statistical model already performs better. Bias (B) determines whether the same fraction of events are both forecasted and observed. A tendency of the statistical model and of WRF-Chem to under-predict O_3 threshold exceedances shows as a B below 1 for these two

models. The false alarm ratio (FAR) that measures the percentage of forecast high O_3 events that turn out to be false alarms, gives highest skill for WRF-Chem, followed by statistical model and persistence. The probability of detection (POD) is a measure of how often a high threshold occurrence is actually predicted to occur, and is relatively low for WRF-Chem with respect to other models. «

Also why were these specific three thesholds chosen?

There was no specific reason for these certain three thresholds. We also performed the calculations for different thresholds, e.g. 130 μgm^{-3} or 150 μgm^{-3}, distinguishing between higher and lower ozone maxima, and the conclusions were similar. We included some thresholds which present an elevated ozone levels and pose a greater risk to human health, and decided to exclude the statistics for a higher threshold (180 μgm^{-3}, a legislation limit value) due to a very low number of exceedances for this threshold. In the paper we extended the following sentence: »Table 5 summarizes the categorical evaluation results for three different thresholds (120, 140, 160 μgm^{-3}) of elevated ozone levels, which pose a greater risk to human health.«

Grammatical and other minor corrections.
p1030 line 22, "The first RT-AQF systems.."
p1030 line 25, delete "existing"
p1032 line 13, "during summertime conditions"
p1032 line 21, "a one-way"
p1032 line 22, "evaluated a forecast"
p1033 line 2, "based on WRF-Chem are implemented worldwide"
p1033 line 4, "over the topographically complex"
p1033 line 6, "with a statistical model"
p1033 line 6, "at the Slovenian"
p1036 line 19, "a southwestern"
p1036 line 24, "shows a mean O3 daily mean"
p1037 line 27, "is a mountainous station"
p1037 line 27, "As well as the elevated station KRV, the ISK, OTL and VNA stations area are also influenced by regional transport of pollutants.
p1038 line 7, "information about the AQ forecast can also be gained by the evaluation of meteorological forecasts for these stations."
p1038 line 16, "index of agreement"
p1041 line 3, "with a range of 0.64 to 0.90 for 1 day forecasts"
p1041 line 7, "On average"
p1042 line 8, "3 month accumulations by"
p1042 line 3, "has problems simulating the"
p1043 line 1, "the model over-predicts"
p1043 line 5, "explained by model error in"
p1043, line 16, "poorly reproduced meteorological"
p1043, line 26, "Also interesting to discuss are the results"
p1045, line 3, "In this section we want to answer the question: 'how accurate is the 1 h O3 daily maximum WRF-Chem forecast in comparison to the statistical model prediction or to persistence?'."

p1045, line 8 "which is, along with their computational efficiency, "
p1045, line 9 "Among the strengths of the deterministic models are that they give"
p1045 line 12, "Furthermore, they also allow forecasts for"
p1045 line 14, "descriptions of"
p1045 line 27, "because a statistical"
p1046 line 1, "with an available"
p1046 line 5, "already beats persistence"
p1046 line 12, "than the statistical forecast"
p1046 line 25, "MNBE in Fig. 8 has very similar results to ME."
p1047 line 13, "also contingency-table-based statistics are an important metric of"
p1047 line 15, "It is important to take into account"
p1048 line 9, "were to be applied to"
p1049 line 7, "local emissions result in model underestimations of NO2"
p1049 line 12, "show good WRF-Chem model performance"

We revised the text according to the suggested corrections and would like to thank again for the thorough reading of the paper.