Referee #2, anonymous

*We would like to thank the Referee #2 for the in-depth review and detailed comments. His/Her comments are in **bold** and our replies to the comments are in normal font. We included here and replied only the critical comments.*

…

Specific Comments

Page 1, Line 17: There are many instances of short but 'heavy' freezing rain.

We suggest removing 'short-lived'.

Page 2, Line 12: A recent climatology over parts of northern Eurasia has been completed:

- **Groisman, P.Ya., O. N. Bulygina, X. Yin, R. S. Vose, S. K. Gulev, I. Hanssen-Bauer, and E. Førland 2016: Recent changes in the frequency of freezing precipitation in North America and Northern Eurasia. Environ. Res. Lett. 11, 045007.**

We suggest familiarizing ourselves with this work and, if found relevant, including it as a reference.

Page 3, Line 27: 3-hourly reports are probably insufficient. Most freezing rain events occur at shorter time scales. Was any attempt made to at least 'estimate' how many events were uncounted by using hourly information as well?

No such attempt was made since we did not have hourly observations available. Without good observations, too many assumptions are required for the estimation of the distribution of the short-term events. This would be an interesting exercise to do, but it is out of the scope of the present article, and indeed requires 1-hourly (or maybe even denser) observations.

Page 3, Line 27: I may have missed this but how did you treat combinations of precipitation types? It is common for freezing rain to occur with ice pellets for example.

According to WMO standards, only the weather with the largest code number is reported in the SYNOP present weather code. Therefore, if e.g. ice pellets (79) occur together with freezing rain (66, 67), only ice pellets are reported, and no information of the simultaneous FZRA is recorded at all. We calibrated the FMICLIM method to identify reported FZRA, and for that reason selected SYNOP codes 24 (freezing rain within past hour but not at observation time), 66 (light freezing rain) and 67 (moderate to heavy freezing rain) as target classes.

Page 4, Line 3: A threshold of 80% is quite low. Why didn't you show the fraction of missing data during the cold season?

We suggest showing the fraction of missing data during the cold season to help readers to estimate the number of missing FZRA observations in the total numbers of FZRA events.

Page 4, Line 15: What fraction of observations was beyond these thresholds? Were any of the high valued temperatures associated with very low relative humidities that would lead to much lower wet bulb temperatures?

We did not understand if the Referee #2 is asking what fraction of FZRA observations, or what fraction of all observations of surface air temperature is beyond the thresholds. Here, we assume that he/she means FZRA observations. In that case we suggest (1) testing how much FZRA data is filtered out by those thresholds, and (2) testing how much FZRA data is filtered out using the stricter interval proposed by Referee #1 (Dr. Nikolov). If a large proportion is filtered out in (2), we suggest (3) testing how much validation scores are altered. If major enhancements in validation scores are found in (3), we should also consider rerunning the calibration procedure.

Page 5, Line 7: This scale is very large for freezing rain. It is quite common for these regions to be less.

We are not quite sure if we understood this comment correctly.

If the Referee #2 is referring to the coarse spatial resolution of the ERA-Interim reanalysis data, we agree that we would prefer to use spatially denser reanalyses. As said in the text (Page 13, Line 28), this is considered to be one part of the future work.

Page 6, Line 3: Ice-initiated precipitation is not initially generated in the inversion aloft. What is the implication of assuming it is?

The implication is that the algorithm does not identify those FZRA events. Those non-identified events are compensated in the calibration by adjustments of the temperature thresholds in different layers. These adjustments then lead to generation of erroneous FZRA events to some other time steps (i.e. false alarms), and to deterioration of the validation results.

We suggest adding a sentence about that in the revised paper.

Page 11, Line 31: This paragraph is poorly worded and hard to follow.

We suggest to rewrite the paragraph, for example as follows:

“In validation, the gridded meteorological dataset is compared with the point-like surface observations. Each grid cell represents spatial means in the 0.7° resolution, while weather stations represent more local variability of the atmosphere. It is possible that in some cases FZRA has not been observed at a station although it has occurred rather nearby. Although hypothetical, this suggests that our estimates, derived from ERA-Interim, might at least occasionally represent the occurrence of FZRA inside the 0.7 grid cells better than the stations do.”

We hope this suggestion is better worded and easier to follow.

Page 12, Line 5: There are standard observing practices to identify freezing rain. Why is this so hard to do?

As said in the text, simply “because the phenomenon can be easily confused with ordinary, non-freezing rain”: For an observer, freezing rain looks the same by eye as non-freezing rain, and accumulation of ice might not be visible (1) in short-term events or (2) when snow covers objects on the ground. Besides, the observed 2m temperature, which is most commonly used by the observers to distinguish FZRA from non-freezing rain at stations, might not represent the temperature of the thick (maybe several hundreds meters) near-surface cold layer well enough in all cases, and knowing if the rain actually is supercooled or not is not so straightforward.

Page 12, Line 8: You are associating ‘minor’ with short duration. On what basis? There can be severe impacts with durations smaller than 6 h and precipitation rates can be high as well..

We suggest replacing that sentence with these:

“Additionally, short-term events, which are more common than longer ones (Ressler et al., 2012; Cortinas, 2000), might not be recorded in the 6-hourly observations. Short-term events are difficult to predict using spatially and temporally smoothed 6-hourly reanalysis data.”

We also suggest correcting references to ‘minor FZRA’ in other parts of the text.

Page 12, Line 11: Are the errors ‘random’?

They are ‘random’ in the sense that we try to predict station level variability using grid cell -level information. Also they are ‘random’ in the sense that we assume the observational errors to be random: equal amounts of (1) false identifications of FZRA events and (2) false rejections of FZRA events happen.

We suggest completing the sentence:

“To maximize the reliability of the observations, a large number of SYNOP stations was used, which is believed to average out random errors in calculation of spatially and/or temporally aggregated results, such as mean annual numbers of events in subgroups (Fig. 3).”

Page 12 Line 25: Given the enormous smoothing at 70 km, maybe the authors should only consider analyses over ‘flat regions’?

This is a good comment. In addition, there is actually even a more important aspect to justify exclusion of the highest elevations from the analysis: the FMICLIM algorithm can not detect FZRA reliably at high altitudes (perhaps >2000m or maybe >1750m), because there the pressure levels used (925, 850, 700 hPa and 2-meter levels) are too few to represent the cold layer -- melting layer structure. We suggest replotting the maps (Figs. 8 and 9) using a mask which hides the suspicious high altitude results. Also, analyses including the high altitude results should be recalculated (e.g. Fig. 7) by excluding the high-altitude data.

Page 12, Line 33: “Occasional misclassification”? How often did this occur?

It happens sometimes, as shown in the Section 3.2.1 (SYNOP weather code classification) and in Fig. 5.

Page 13, Line 3: To me, this section is too long and wordy. This is a long shopping list. What are the most important and feasible next steps? From my perspective, some of these should be done within this article.

We agree to order the list based on the expected importance of the future steps. However, we prefer to present all the suggested steps, because we think that they are all necessary to get more accurate results.

As well, a recent article (Liu et al., 2016) pointed out that precipitation at the surface (including freezing rain) is calculated directly from the model's microphysical package without needing the approach used here. Isn't that the best way forward?

- **Liu et al., 2016: Continental scale convection permitting modeling of the current and future climate of North America. Climate Change, DOI 10.1007/s00382-016-3327-9**

This is of course the optimal solution. However, the precipitation type is not included in the output variables of ERA-Interim.

Page 14, Line 12: I do not think that it is 'sophisticated enough...'. Melting rates of particles aloft, for example, depend on the features of the particles themselves as well as temperature and moisture conditions.

We suggest removing 'sophisticated enough'.

Page 14, Line 15: Why did you not examine sounding information taken during freezing rain events? You could then more quantitatively assess how well the approach is handling particular instances. The lack of such validation is a major drawback in this article.

We agree that the division of uncertainty to method-dependent and ERA-Interim-dependent components would be extremely informative. Examining sounding information is suggested to be given the highest priority in the list of the future steps (Sec. 4.2).

However, the 'perfect data' approach would only give a sort of upper limit estimation of the performance of the algorithm, as, with our gridded data, the algorithm has to operate with the limited number of pressure levels for example. We were interested in the overall capability of the method *and* the ERA-Interim reanalysis in describing the FZRA climatology, and found that together they give a rather good estimate, when aggregated results are evaluated (e.g. Fig. 3). We also got and presented information about the total uncertainty consisting of both uncertainty components.

Page 14, around Line 25: Why not compare against previous studies on the climatological features?

We suggest adding some sentences to this part of the text, even though climatological studies of FZRA in Europe are almost inexistent.

Page 14, Line 30: Given the limitations of the dataset as you have mentioned, how confident are you that you can 'reliably' address such questions?

In that sentence, we (implicitly) refer to Fig. 3, and there we can be quite confident because our results are backed by the observations. There are other results that we are less confident and which are a motivation for further development of the FZRA detection methodology.

Page 14, Line 31: Clarify what is meant by 'station scale analysis'.

We suggest replacing 'station scale' with 'station level' in all occurrences (i.e. in the Conclusions and the Abstract sections).

Station level analysis means comparisons of raw modelled and observed time series of FZRA in all individual stations. No aggregation of data temporally or spatially is applied prior to analysis.

Technical Corrections

Page 2, Lines 5 and 7: The word 'where' is not correct in referring to an event 'in time'. This error was done in other places as well.

We suggest correcting this.

Page 2, Line 22: Another incorrect use of 'where'.

We suggest correcting this.

Page 13, Line 28: 'criteria'

We suggest correcting this.