Reply to reviewer comments

Thank you once again for providing valuable comments to further improve the paper. Please find our replies to your comments below and the changes to the manuscript marked in the manuscript.

General minor comments

1. Terminology

    a. Be consistent in how you refer to figures in your text (see comment 22): You use Figure 2a (line 107), while in line 99 you use Fig. 1 and 2. Also make sure the blank space is consistent between figure and figure number as well.

    We use the abbreviation "Fig." whenever we refer to a figure in the running text and "Figure" when the reference to the figure appears in the beginning of a sentence, which is in accordance with the NHRESS manuscript guidelines. We checked and corrected the blank spaces between the "Fig." and the numbers.

    b. Name your metrics consistently, e. g. RSME and skill are used interchangeably. pointed it out in comments 35,39 and 41 but there might be locations I oversaw. Please check this carefully before resubmission.

    We agree that the use of RMSE and skill should be consistent. We replaced "skill" with RMSE whenever discussing the calculated RMSE, but we continue to refer to "skill" when referring to the performance of the method, as suggested.

2. Definitions:

    a. You use relative statements throughout your paper. Make clear what are low / high FWI values or what is a good correlation. Add numbers to these relative terms and state why you use these numbers as thresholds to define a FWI to be high or low, or a correlation to be good (see comments 20 and 24)

    Thanks for the comment. We improved the definition of high/low FWI values by adding in the Fire weather index calculation (2.1) the definition of low FWI values, additionally we added figures to the supplement showing the mean and 95th percentile of the FWI for each month for all regions. Furthermore, we

checked throughout the paper that it is clear what we mean with high and low FWI values.
We corrected the statements about the correlation coefficient, please see comment 20 and 24.

b. Same for short- and long-term forecasts. Define these terms in your text, e. g. short term (up to 3 days) and long term (more than 3 days).

We now avoid using the term "long-term forecasts" as it can be confused with the definition of long-term forecasts by ECMWF, which is multiple months. Instead, we give the time range we are referring to (1-2 weeks). Furthermore, we now define short-term forecasts as forecasts up to 3 days.

3. General comment on verification and Fig. 2b:

You state several times that your approach does not work well for small/ low FWI values. What are the thresholds for you to consider FWI to be low? From your study I read that this is below 1, but from the EFFIS danger classes (https://forest-fire.emergency.copernicus.eu/about-effis/technical-background/fire-danger-forecast) low FWI values are considered to be below 11.2. Your underestimation of the FWI in the analysis mainly affects this relevant scale (10 -100) (see Fig 2b).

Thanks for this comment. It is indeed not clear what we mean with low values. In general, we mean low values of the FWI below 10. However, in the description of Fig. 2 we especially look at very low values (below 1). We tried to clarify this by adding in the description of the FWI classes, the definition of low FWI by EFFIS and changed the description of Fig.2 to "very low values".

- Can you mention this in the description of the figure (i.e. line 110)
  We added a comment that the underestimation effects the most relevant scale for wildfire risk assessment, as:
  "The analysis tends to overestimate FWI values below 6 and underestimate values above 6. Notably, these higher values (above 10) are the most relevant for assessing wildfire danger."

- Discuss and state stronger that your calibration does not address this discrepancy since you calibrate on the (orange) analysis and not on the (grey) observations density?

  We improved the discussion in chapter 2.1 and hope this is now more clear.

- What are the impactions of your decision to use the analysis as observations in the calibration process. Don't you systematically underestimate large/relevant FWI (FWI > 10) values by calibrating to the analysis over observations? Please add your elaboration on this to the discussion and conclusion.

  It is possible that we underestimate large FWI, however it is difficult to quantify the discrepancy and correct for it because the available stations are not homogenously distributed throughout the study area. Over half of the stations that were available are in Austria, Switzerland, Romania, and Norway (as can be seen in the high density of stations in Fig.1a.). Those regions have a complex topography which can cause discrepancies between model forecasts and observations. Furthermore, there are not many available stations in southern Europe were the FWI is generally high. This also shows the difficulties when using observation data for verification and calibration, the availability and representatitivity of observations for large areas needs to be considered and the quality of the observations, i.e. wind and precipitation, which can strongly vary, needs to be assessed. We added a discussion of this issue to chapter 2.2 and the Discussion and conclusion chapter.

Detailed minor comments

1. Line 2: "makes accurate wildfire risk estimation crucial" add a target group, e. g. decision makers, emergency responders, etc.

   We added target groups by changing the sentence to:

   "Wildfires are increasing in frequency and severity across Europe, which makes accurate wildfire risk estimation crucial for decision makers and emergency responders."

2. Line 6: You already introduced the abbreviation FWI for the Canadian Fire Weather Index. I suggest to not spell it out here and just use FWI.

   We applied this change and do not spell out the FWI here.

3. Add a strong last sentence to your abstract, in which you point out what the target group and the impacts of your study should be.

   We added a concluding sentence to the abstract.

4. Lines 13 - 21: You added good references for your description of the more northern parts in the second part of your paragraph (line 16ff.). Could you underline your statements in the first part as well?

   We added two more references (Turco et al., 2019; Rodrigues et al., 2023) for the wildfires in 2017 and 2022 in South Europe.

5. Line 27: In your abstract you say you look into forecasts up to 15 days which is 2 weeks, while here you state you are interested into "forecasts ranging from a few days to several weeks". Do you refer to your study or the outline of the SAFERS project here and what are several weeks?

   We here refer to the SAFERS project which provides weather forecasts up to 45 days (~6 weeks), however in this study we are looking only into forecasts up to 2 weeks. We clarified this by replacing "several weeks" with "six weeks" and adding to the next sentence that we are looking into forecasts up to 2 weeks in this paper.

   "An important component of SAFERS for identifying high wildfire risk areas are accurate and reliable weather forecasts, ranging from a few days up to six weeks. In this paper, we use the Canadian Forest Fire Weather Index (FWI, Van Wagner (1987); Di Giuseppe et al. (2016)), a widely recognized numerical indicator for forest fire risk, to derive fire risk from weather forecasts with a lead time up to two weeks."

6. Line 28: Rephrase the sentence and add what you use the FWI for, e. g. "in this paper, we use the FWI … for deriving fire risk from weather forecasts".

   We rephrased the sentence to:
   "In this paper, we use the Canadian Forest Fire Weather Index (FWI, Van Wagner (1987); Di Giuseppe et al. (2016)), a widely recognized numerical indicator for forest fire risk, to derive fire risk from weather forecasts."

7. Line 28ff: Here you could improve the reading flow by two things:
   - You should shortly mention the parameters of the FWI, i.e. naming the four weather parameters in it (temperature, precipitation, wind speed and rel. humidity).
   - Add a transition to weather forecasts by adding a statement that these parameters are available in deterministic and probabilistic forecasts.
   - Then you can continue with explaining the pros and cons of deterministic vs. probabilistic forecasts.

We changed this section according to your suggestion as follows:

"The calculation of the FWI only requires four weather parameters: temperature, relative humidity, wind speed, and 24-hour accumulated precipitation, which are often available from deterministic or probabilistic weather forecasts. While deterministic forecasts provide a single forecast based on a given set of initial conditions, probabilistic ensemble forecasts offer a range of possible outcomes by using slightly perturbed initial conditions, giving a more comprehensive picture of potential weather conditions, and providing an estimate of the forecast uncertainty."

8. Line 35ff: Thank you for adding this paragraph about other methods and adding your purpose of the study. This section is much clearer now.

9. Line 49: Add a reference after various regions, e. g. Di Giuseppe et al. (2016)

   We added the suggested reference and an additional reference referencing the adaptation of the FWI in Southeast Asia.

10. Line 51: Please rephrase "relatively straightforward computation" to a more scientific language. I disagree that the many empirical formulations of the FWI and its subindices are straightforward to develop, though they are straightforward to apply once a source code is available.

    We dropped this part of the sentence and now only mention what is necessary to calculate the FWI.

11. Line 55: Add ", i.e. Drought Code (DC), Drought Moisture Code (DMC), and Fine Fuel Moisture Code (FFMC)".

    We added the moisture codes as suggested.

12. Line 65: Add ", i.e. Initial Spread Index (ISI)."

    We changed the sentence to:
    "In the second step, FFMC and the 10-meter wind speed are used to model the potential rate of fire spread, i.e. Initial Spread Index (ISI)."

13. Line 70: Add reference to EFFIS fire danger classes.

We added a reference.

14. Line 70-71: I like that you reflect on the fire danger level thresholds here, but you do not show results for FWI values later in your results section and you do not group your FWI values into these classes. Therefore, you can remove these two sentences about the fire danger levels.

    We would like to keep these sentences as they give some background information about the usual magnitude the FWI. Furthermore, we added what is considered a low FWI according to EFFIS in order to make it better understandable what we mean with low FWI later in the paper.

15. Line 74: Drop “with modifications to utilize gridded input data” or provide the code in a GitHub repository, because I am curious to see/ use it now.

    We dropped the sentence because we might not be able to provide the GitHub repository link in time before the publishing of the article.

16. Line 86: You missed a word (“mean”?) after climate.

    Thank you for finding this mistake. We added “mean” after climatological.

17. Line 87: Rephrase to “using a centered 15-day rolling mean on each day”.

    We rephrased the sentence according to your suggestion.

18. Line 89: Here, I do not understand where the ECMWF high-resolution forecast comes from. Is this part of the TIGGE archive or from a different archive? Please clarify.

    The high-resolution forecasts are retrieved from ECMWF's Meteorological Archival and Retrieval System (MARS) and not part of the TIGGE archive. We clarified this by adding the reference to MARS.

19. Line 93ff: This sentence is confusing. Aren’t you showing weather station data in Fig. 1a)? Please restructure this in accordance with the beginning of your new paragraph and make this a uniform/ clear statement.

    Yes, we are showing the FWI calculated from weather station data in Fig.1, but this observation data is not used for the verification. The purpose of showing observation data in Fig.1 is to compare analysis and observation and to justify

that we are using the analysis instead of actual observations. However, we restructured this section in an effort to make it clearer:

"For calibration and verification purposes, we use ECMWF high-resolution deterministic forecasts initialized at 0000 UTC, which are available from ECMWF's Meteorological Archival and Retrieval System (MARS2). ECMWF high-resolution forecasts have a spatial resolution of 0.1° (~9 km) and a temporal resolution of 1 hour and can therefore give a more accurate picture of the actual weather conditions than medium-range ensemble forecasts with a coarser resolution. Ideally, the FWI forecasts would be verified using FWI values calculated from surface observations of the relevant weather parameters, since the FWI cannot be directly observed. However, measurement stations that provide continuous observations of all necessary weather parameters are sparse and only yield point-wise verification. Furthermore, for an operational calibration of the FWI, observation data would need to be rapidly available. We therefore use the FWI calculated using ECMWF high-resolution forecasts with the shortest lead time to the local noon with corresponding 24h precipitation as substitute for actual observations. These FWI values are hereafter called analysis."

20. Line 103: Please put your correlation coefficient (r=0.72) in brackets here. The definition of a good correlation is varying in different science domains, and I would rather read that you found a correlation coefficient of 0.72 and why you think this is sufficient for your analysis.

    In this section, we do not refer yet to the correlation coefficient but rather to the agreement of analysis and observation in the time series plots. The correlation coefficient of forecasted and observation FWI and the mean value using all the stations is only given later in this section when referring to Fig.2. We therefore don't think it would be correct to provide the correlation coefficient at this point. However, we replaced "correlation" with "agreement" to make it more clear that we talk about the time series plots and not the correlation coefficient.

21. Line 107: Missing word after Figure 2a.

    We added the word "shows" to the sentence.

22. Line 107 and wherever you refer to a figure, here you use Figure 2a, while in line 99 you use Fig. 1 and 2. Please make this uniform across your manuscript.

We use the abbreviation “Fig.” whenever we refer to a figure in the running text and “Figure” when the reference to the figure appears in the beginning of a sentence, which is in accordance with the NHRESS manuscript guidelines.

23. Line 112: Do you mean analysis by “forecast-derived”? Please stick to one term across your manuscript.

We replaced “forecast-derived” with “analysis”.

24. Line 115: Why is it here only a “fairly good” correlation while it was a “good correlation” in Line103?

We removed “fairly”.

25. Line 119: Make the statements about the markers more specific, e. g. “the bright red colored markers”.

We changed this statement to “as indicated by the yellow and red coloured markers”.

26. Fig. 1b-c legend + Fig 2 legend: shorten “calc. from observation” to just “observations”.

We changed this in the figure legends to “observation” as suggested.

27. Fig. 1 caption:
    - “See Fig. 2” to “see Fig. 2c”.
    - Add “(b – d)” after example stations.
    - Move “(b)” in front of “in NEU”.

We applied all your suggested changes.

28. Fig 2c: Can you add the rmean as a vertical line to the plot?

We added the r_mean as vertical line in Fig. 2c.

29. Line 143 – 146. These two sentences need to be restructured. Mention the length of your training periods in the beginning, e. g. “training periods between 15 to 40 days were tested”, then add your benefits and disadvantages of different training lengths.

We rearranged the sentences and now mention the tested training periods in the beginning of these sentence as suggested.

30. Line 147 – 148: What training period are you using in your analysis? You say you found 30 days appropriate for small geographical areas and then that you adopt a regional approach, which is in my understanding a large geographical area. Please clarify which training period you stick with here explicitly.

With the regional approach, we mean that we use all the grid point in the area to derive the calibration coefficient, as opposed to the local approach where coefficients are derived for every single grid point. So, the regional approach is used for small as well as for large geographical areas.

We added a sentence mentioning that we use a 30-day window.

31. Line 158: Please replace several with the verification metrics you are using.

We added the verification metrics that are used to the sentence.

32. Line 160: Drop introductory sentence.

We believe this sentence is necessary to understand why we use RMSE (as measure for the skill) and the spread. However, we rephrased the sentence and added a reference to make it better understandable that we are referring to the spread-skill relationship here.

33. Line 164/ 165: I don't understand how you derive the ensemble spread. Is the ensemble spread the standard deviation (square root of the ensemble variance) of the ensemble? What is the average ensemble variance? The mean of the variance of all ensemble members? Please rewrite this.

The ensemble spread is calculated as the square root of the average ensemble variance, where the average ensemble variance is the mean of the variances of all ensemble members. According to the given reference (Fortin et al., 2014), the square root of the average ensemble variance needs to be used instead of the average of ensemble standard deviation, because smaller values would be achieved unless the spread is the same for all time steps. We changed the sentence to:

"The ensemble spread is calculated as the square root of the average ensemble variance, where the average ensemble variance is the mean of the variances of all ensemble members (Fortin et al. , 2014)."

34. Lines 166/ 167: Please be clear in your terminology. Is skill in line 167 the same as RMSE in line 166? Please be consistent with these terms throughout your manuscript and figures, i.e. Fig. 3.

We changed skill to RMSE in this location because RMSE is here the same as skill.

35. Line 185: Replace “introduced calibration method” with “the NGR method”.

We replaced this part of the sentence accordingly.

36. Line 186: Name the regions again, i.e. NEU, WCE and EUMED.

We added here the region names again:

“We use here the AR6-WGI reference regions, NEU, WCE and EUMED, introduced in Sec. 2.3.1 and shown in Fig. 1a.”

37. Line 190: What is a considerably high FWI? Please find a reference using this fire seasons to refer to your Fig 1b-d.

We added the reference to Fig. 1b-d, which shows the higher FWI during the months May-October. Furthermore, we changed the sentence to make it more clear what we mean with the higher FWI values during these months:

“For the calibration verification, we therefore only focus on forecasts during the months May to October, when the FWI in all regions is substantially higher compared to off-season months (see Fig. 1b-d)”

38. Line 197-199: Here it becomes apparent that you should be consistent with naming it either skill or RMSE.

We changed skill here to RMSE.

39. Line 200: Move the statement that your method improves the forecast to the beginning of the sentence to make your statement stronger.

We changed the sentence as suggested.

40. Line 202/ 203: Consistent naming of RMSE and skill becomes apparent again. From the reading flow it might make more sense to stick to RMSE when you talk about the measured RMSE (line 202: “skill of raw and calibrated”) and use skill when you talk about the performance of your calibration method (line 203: “the forecast lacks skill”).

We agree and changed it like suggested to RMSE whenever we talked about the calculated RMSE and skill when referring to the performance of the method.

41. Line 205: What do you mean by large uncertainty? If you mean that the signal-to-noise ratio in NEU is higher than in other regions, because the FWI values are lower and more unlikely to exceed the variability range, then bring this statement more to the point. You could underline this statement by whatever you mean with "(not shown here)".

Yes, this is what we mean. We changed the statement and added a reference to the figure showing the FWI values in the supplement.

42. Line 206: What are you not showing here and where do you show the relative uncertainty in different regions?! Drop this statement, refer to supplementary material or refer to a Figure.

We dropped this statement.

43. Line 214: Just add in brackets the reference to the figures in your supplementary material.

We changed the reference to the figures like suggested.

44. Line 222: What is your reference in Fig. 5 for stating that the skill worsens in NEU but not in WCE? In Fig 5. the median in WCE goes below 0 at lead time between 180 and 220 and therefore, I think that the skill worsens here as well. In EUMED the skill is not worsening if you refer to the median of the boxplots.

You are right, we corrected these statements and now refer to the median of the CRPSS.

45. Line 223/ 224: I don't understand the context of these two statements: Is this worsening of skill caused by small FWI values or by longer lead times? Didn't you say earlier you excluded FWI values below 1? Can you refer to a figure why you think low FWI values are the cause of the worsening? Throughout your manuscript you bring up this argument frequently, but you never show how high/ low the FWI values in the different regions are, and why you come to this conclusion.

We believe it is caused by the low values and the lower variability of values. In NEU, the FWI values are usually quite small (FWI < 10) and the calibration has less effect, in contrast to EUMED where the range of FWI values is larger. To illustrate what high and low values are in the different regions, we included a figure in the supplement showing the mean and 95th percentile of the FWI for all regions and each month.

46. Line 225: Merge this with the paragraph above. Shorten this by dropping the first sentence and adding the second sentence to the paragraph above by referring to the figure in brackets: “With increasing lead time the skill worsens especially in mountainous areas in Scandinavia and the Alps, which are also the regions with generally low values throughout the fire season (see. Figure 6).”

We applied this change.

47. Figure 6 comment: You claim that your method does not perform well over mountain regions. Which is true for lead times of 228h and 324h. But at lead time 132h, I see that the Atlantic influenced regions all do not perform very good. Has this something to do with the ocean/land interface? Could you investigate this and add a statement about this to your manuscript?

The method indeed does not seem to work well in the Atlantic influenced region, e.g., the west coast of the UK and the northwest tip of Spain. In these regions, the FWI is (similar to the mountainous regions) quite low, which can be seen in the newly added Figures of the FWI mean and 95th percentile in the supplement (new S2 and S3). This can be attributed to the increased rainfall caused by the ocean’s influence in these regions. We added a comment about this in the description of Fig.6.

48. Line 234: please shorty reintroduce the term FWI analysis, e. g. FWI derived from the ensemble forecast (analysis). For people skimming your paper it might appear as you analyze the FWI, which you don’t.

We agree that this would be beneficial for readers only skimming the paper and reintroduced the FWI analysis term as follows:

“We used ECMWF high-resolution weather forecasts with the shortest possible lead time to calculate the FWI. The FWI values from these forecasts are referred to as the FWI analysis and are used as a substitute for observations. Although the FWI analysis seems to slightly underestimate the FWI, a good correlation is observed.”

49. Lines 236 -239: Please be consistent when using lead time units. It is hard to transfer from 84 hours to 8 to 10 days. You can state the lead time in hours and add the days in brackets after the hours as you stated in your response to the reviewers.

We added the days in brackets after the hours. It should now be easy to compare the short lead times to the 8 to 10 days mentioned later.

50. Line 240: this statement is weak and a reference to appendix (e.g. “see Fig S5-7”) is missing. From these figures you can only read that the mean error is smaller in the summer months, but not how high or low the FWI is in the specific subregions. Therefore, you cannot state here that your findings are likely caused by low FWI values.

We added a reference to the appendix and also to the new supplement figures showing the mean and 95th percentile of FWI (S2,S3) to illustrate the range of FWI values in the subregions. From these figures, it becomes clear that in NEU the FWI is generally lower than in EUMED.

51. Lines 240: It is not clear what you consider high and low FWI values, please add values or define low or high FWI somewhere, e. g. “high FWI values (FWI > 1)”.

We added a clarification what we mean with high and low FWI values.

52. Line 244: Rephrase “Although it would be ideal” to more scientific language.

We rephrased to “While it would be preferable” and hope it sounds more scientific.

53. Line 246: Define what you mean by “long-range” and “short-range” forecasts.

We changed the sentence referring to:

“While forecasts of potential fire danger for extended periods (1-2 weeks) are valuable, short-term forecasts for the first 1-3 days, are usually more critical for firefighting resource management.”

In this way, we avoid calling it “long-range forecast” which is defined by the ECMWF as subseasonal (multiple months). However, in this context, we mean longer than short-term forecasts, e.g., 1-2 weeks. Short-term forecasts are forecasts up to 3 days which is stated in the following sub-clause.

54. Line 251: Move everything after “compared to more complex approaches” to the beginning of the discussion and conclusion chapter, e. g. to line 233 before you start with the sentence “We used…”, to end your paper with this very good and strong statement: The improvement of FWI forecasts using the presented calibration method improves the ability to anticipate fire danger, ultimately supporting better response management and shows that a relatively simple method can provide good results compared to more complex approaches. In line 233 you can discuss that NGR is a good method in comparison to other methods (e.g. bias correction) for your research purpose.

We applied you suggested changes to the Discussion and conclusion chapter.