Author's response

Thank you for reading the manuscript and providing valuable suggestions to improve the paper. Our responses to the reviewer comments are shown below in blue, and changes to the manuscript are indicated in italics. Additionally, all modifications are marked in the revised manuscript.

Reply to the comments of reviewer 1

Major comments

45-47 ". In the second step, FFMC, DMC and DC are used to model the rate of fire spread (ISI) and the potential fuel available for surface fuel consumption (BUI)." The wind speed is missing (ISI = FFMC + WS10), and it would correlate to the POTENTIAL rate of spread (I mean, it is a variable involving the combustion of surface fuel, like dry leaves and such + the wind). Also, the BUI = DMC + DC, so it is not surface fuel consumption, but it involves the potential for a surface fire to burn the deeper fuel (Build Up) and become a much more persistent fire.

Thank you for your comment. The suggestions have been included in the revised manuscript. The description of the FWI calculation reads now as follows:

"The FWI is calculated in two steps. First, the 2-meter temperature, 2-meter relative humidity, 10-meter wind speed, and 24h accumulated precipitation at local noon are used to calculate the moisture content of three separate fuel layers. These fuel layers are characterized by different depths and fuel consistencies, which result in varying water capacities and drying speeds. The Fine Fuel Moisture Code (FFMC) represents the moisture content of litter and other fine cured fuels at a nominal depth of 1.2 cm; the Duff Moisture Code (DMC) indicates the moisture content of loosely compacted layers (nominal depth ~7 cm); and the Drought Code (DC) denotes the moisture content of deep, compacted layers in a depth of around 18 cm (Van Wagner, 1987). DMC and DC respond slower to weather variations compared to the fast-drying fuel represented by the FFMC. Consequently, the effective day length, which determines the amount of drying that can occur during a given day, must be considered and monthly day length adjustment factors for DMC and DC are applied based on latitude (Lawson and Armitage, 2008). The fuel moisture codes are dependent on previous weather conditions; therefore, the preceding day's noon values for FFMC, DMC and DC are necessary for their calculations. In the second step, FFMC and the 10-meter wind speed are used to model the potential rate of fire spread (ISI). DMC and DC are used to calculate the Buildup Index (BUI), a numeric rating of the total amount of fuel available for combustion which comprises the potential of a surface fire to burn deeper fuel layers (build up) and thus evolve into more persistent fires. These fire behaviour indices are then used to calculate the FWI. The FWI values are always non-negative, with low numbers indicating low fire weather danger and high values indicating high fire weather danger. The FWI is often classified into danger classes and values above 50 are considered extreme. However, those levels can vary depending on local conditions (e.g., vegetation types).

Consequently, what is considered a low or extreme FWI in one region may not be equivalent in another. A more comprehensive description of the FWI system can be found in Van Wagner (1987) and Lawson and Armitage (2008). For this study, we implemented the calculation of the Canadian Forest Fire Weather Index (FWI) using Python programming language, following the source code provided by Wang et al. (2015), with modifications to utilize gridded input data."

76-78 "Figure 1b shows the scatter plot of analysis and observations for all stations and every time step. While the FWI derived from the forecasted weather parameters seems to generally underestimate the FWI values compared to the values derived from the observations (slope ∼ 0.63)" AND Figure 1b: Many doubts arise from this scatterplot:

- The plot itself shows too many points to use a scatter. A density plot NEEDS to be used in this case, or two if you want to show separately the data from Finland.

Thank you for your comment. We agree that the scatterplot may not have been the most suitable choice for representing our data and conveying our message. We decided to change the representation to histograms showing the distribution of the FWI calculated from high-resolution forecast (analysis) and from observation (Fig.2 in the revised manuscript). Additionally, we split the Figure into 2 parts to avoid overcrowding the figure. The new Fig.1 contains the map of observation stations (a), with colour representing the linear correlation coefficient, and the time series of three example stations (b,c,d). The new Figure 2 contains the histograms of the frequency distribution of the FWIs (a,b) and the linear correlation coefficient (c).

- The regression which leads to the 0.63 slope seems to be off by looking at the scatterplot, which might be due to the lack of information about the point density. What causes the slope to be 0.63 (and not closer to 1, as the scatterplot would suggest)? Also, please specify the method used for the regression (I assume linear regression).

The regression line in the original Figure 1 appeared off due to the high concentration of values near zero. This becomes now evident considering the distribution shown in histogram Fig.2a. We also clarified in the text that we used linear regression to obtain the coefficient.

- I am aware that it is common practice to use the ECMWF analysis at minimal lead time in place of observations, but once you have found an important underestimate like you did, why did you dismiss it so fast? It seems like a very important matter that can have a huge impact on the paper's reliability. Please explain in depth why you can ignore this bias or what you did to correct it.

We agree that a correction of the analysis based on the observations would be beneficial. We found that the strongest discrepancies between observation and analysis occur in mountainous areas, e.g., Austria or Norway. This is likely caused by difficulties in capturing the small-scale weather phenomena occurring in the complex terrain with relatively coarse resolution of the forecast model. However, in those regions the fire danger is generally rather small because of high precipitation and low temperatures. To further investigate this bias, additional station-specific characteristics such as elevation or land use would be

necessary. In this article, we want to present an easy, straightforward method to calibrate the FWI and we believe that the correlation of analysis and observation is already quite good and justifies the use of the analysis as substitute for observations. However, the use of additional station-specific parameters could be topic of future research.

The section that previously related to the scatterplot has been changed to:

"To determine if the analysis is suitable to be used as observation substitute, we check their agreement with actual observation-based values, which is shown in Fig. observation and 2. We use observations available from the Finnish Meteorological Institute's observation database for the years 2021–2023. Figure 1a shows the stations, for which it is possible to calculate the FWI for more than 200 consecutive days in addition to the reference areas which are introduced in Chapter 3. In total 682 stations can be used. The time series of three example stations are shown in Fig.1a,b,c. The FWI analysis is shown in orange while the FWI calculated from observations is shown by the black dashed line. Overall, there is good correlation between the forecasted and observed FWI values. However, especially for high FWI values, the FWI derived from forecasted weather parameters tends to underestimate the values compared to those derived from observations. This is particularly evident during the summer months of 2022 and 2023 in Meiningen, Germany (Fig.1c) and Chrysopouli, Greece (Fig.1d).

Figure 2a overlapping histograms of the density distribution of the FWI analysis (orange) and FWI calculated from observations (grey) for all stations shown in Fig.1a and every time step. The distributions show a high frequency of low values (FWI< 1), and when plotted on a logarithmic x-scale, a bimodal structure becomes evident with a separation at an FWI value of 1. This bi-modality results from a necessary restriction imposed on the FWI calculation (Eq. 30a,b Wang et al. (2015) or Eq.40 in Van Wagner (1987)). Figure 2b again displays the density distribution of FWI values, but focuses on FWI values greater than 1, as these are generally more relevant for assessing wildfire danger. For higher FWI values, the forecast-derived FWI underestimates those calculated from observations, while for lower FWI values, it tends to overestimate. One contributing factor is the different spatial resolution of the data sources. The analysis uses gridded data with a resolution of 0.1° (approximately 9 km in Central Europe), whereas weather observations are local measurements. However, in general there is a fairly good positive correlation of FWI analysis and FWI calculated from observations. Figure 2c shows the histogram of the linear correlation coefficient of stations that observe FWI values greater than 1, the mean correlation coefficient is 0.72. Low correlation coefficients are mainly caused by differences between analysis and observation in mountainous areas, such as Austria, Switzerland, Romania, and Norway, as indicated by the colored markers in the station map (Fig. 1a). This is due to the difficulty in capturing the complex terrain and its small-scale weather phenomena with relatively coarse resolution (~9km) of the forecast model. Because there is a fairly good correlation of observation and analysis, we will use the forecasted FWI with short lead time (analysis) derived from ECMWF high-resolution forecasts as observation to compare with the FWI forecasts derived from ECMWF medium-range ensemble forecasts."
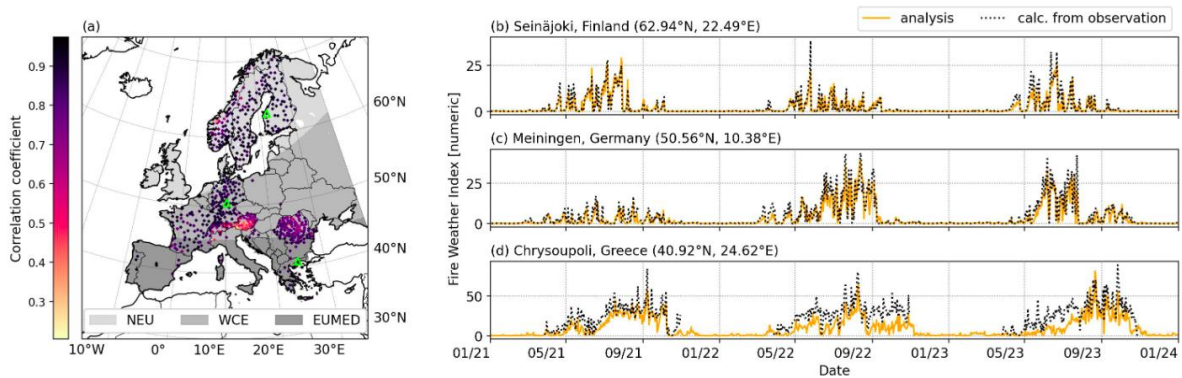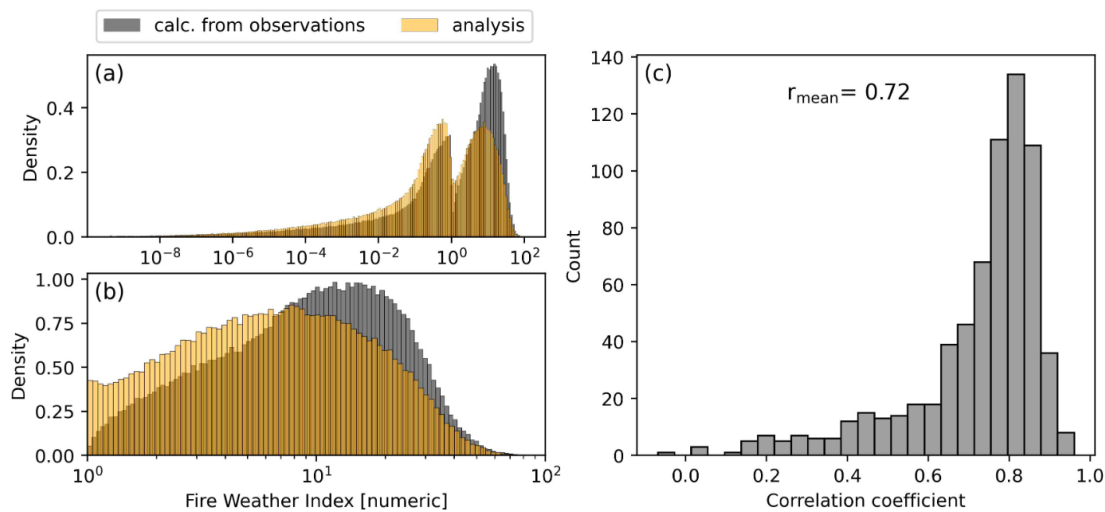
Fig.1

Fig.2

78-79 "a correlation is apparent. This good correlation can also be seen in the time series examples for a station in Finland and Greece": Please provide us with the necessary quantitative information (e.g., correlation coefficients for all the stations) to support this claim, especially coming right after the previous comment. Two sample stations (Figure 1c) are not enough to validate a claim on over 600 others; by this I do not mean that Figure 1c must be removed.

We agree that only showing two examples is not enough to show that there is a good correlation. In the revised manuscript we included an additional figure (Fig.2c) which shows the distribution of the correlation coefficient for all stations. The mean correlation coefficient is 0.72, which indicates a fairly good positive correlation. Additionally, we added a third station in Fig.1, providing an example for each subregion. Changes to the manuscript can be seen in the reply to the previous comment.

90-97 "Furthermore, data from all grid points in the training area is used to estimate a single set of coefficients for the given day (regional EMOS)": some additional explanation is then needed, how do you go from the μ_{kl} to the estimate used (I guess μ_l ?)

We improved this section (now 2.3.1) to make the method better understandable and to clarify what training data is used. Specifically, we changed this sentence to:

"We adopt here a regional approach, pooling training data from all grid points within the training area to derive a single set of calibration coefficients (a_l - d_l) specific to each lead time for the given forecast."

We hope this clarifies that the calibration coefficients are estimated by the calibration method and then used to calibrate the ensemble mean of the given forecast.

Figure 2:

- It can be made clearer if the legend was more explicit (dotted line with triangles: spread (raw) / solid line with dots: RMSE (calibrated) / etc.)

We made the legend more explicit as suggested.

- The legend, which is relative to all the three graphs must be outside the first graph. Consider also putting everything in a single column.

We placed the legend outside as suggested.

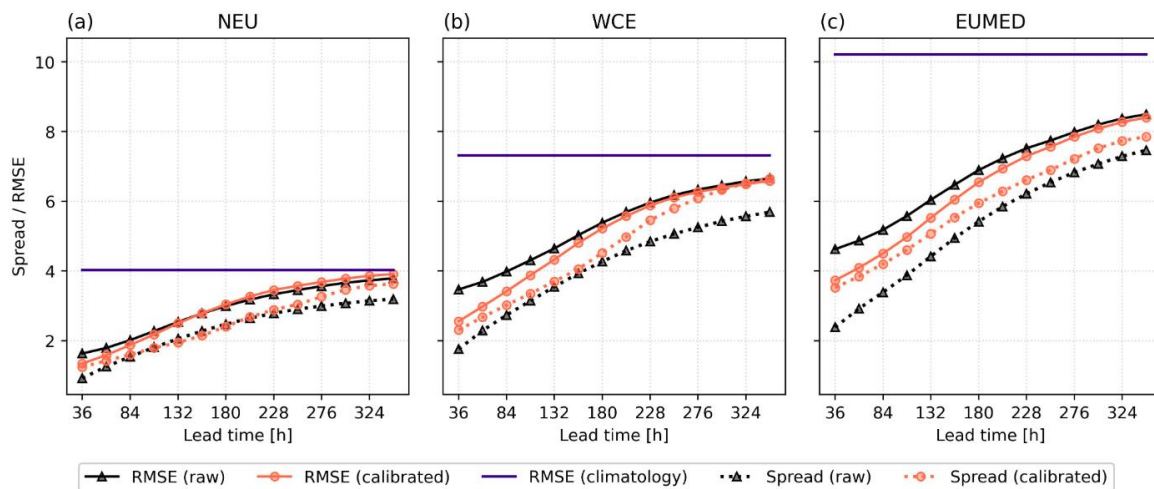The previous Figure 2 is now Figure 3 in the revised manuscript:

Fig.3

152-153 "In Northern Europe, the RMSE of the calibrated forecast is slightly above the RMSE of the raw forecast after 7 days of forecast," : please provide at least a hypothesis as to why this happens. The sentence on the subsequent lines "The regional differences could be explained with the generally higher FWI values in the more southern, fire prone regions compared to Northern Europe where FWI values are often very small" addresses the regional differences, but the difference from uncalibrated and calibrated NEU RMSE is not addressed.

We added a hypothesis trying to explain the slightly decreasing RMSE of the calibrated forecasts in NEU as follows:

"In Northern Europe, the RMSE of the calibrated forecast is slightly above the RMSE of the raw forecast after 7 days (180 hours) of forecast, whereas the skill of raw and calibrated forecast in Central and Mediterranean Europe is similar for forecasts longer than 9 days (228 hours) and 12 days (300 hours), respectively. The FWI forecast in Northern Europe lacks skill for lead times longer than 7 days and calibration based on 30 day rolling window fails to improve and even worsens the forecast. This finding could be explained by the rather small FWI values and large relative uncertainty in NEU, unlike FWI values in WCE and EUMED (not shown here). The applied calibration appears to be more effective for higher FWI values and shows limitations for smaller FWI values close to 0. This also explains the regional differences and the better calibration results in the more southern, fire prone regions with generally higher FWI values compared to Northern Europe, where FWI values are often very small."

Figure 3:

- The legend, which is relative to all the three graphs must be outside the first graph. Consider also putting everything in a single column.

The legend has been placed below the figure to make it clear that it relates to all three subplots. Figure 3 is now Fig.4:
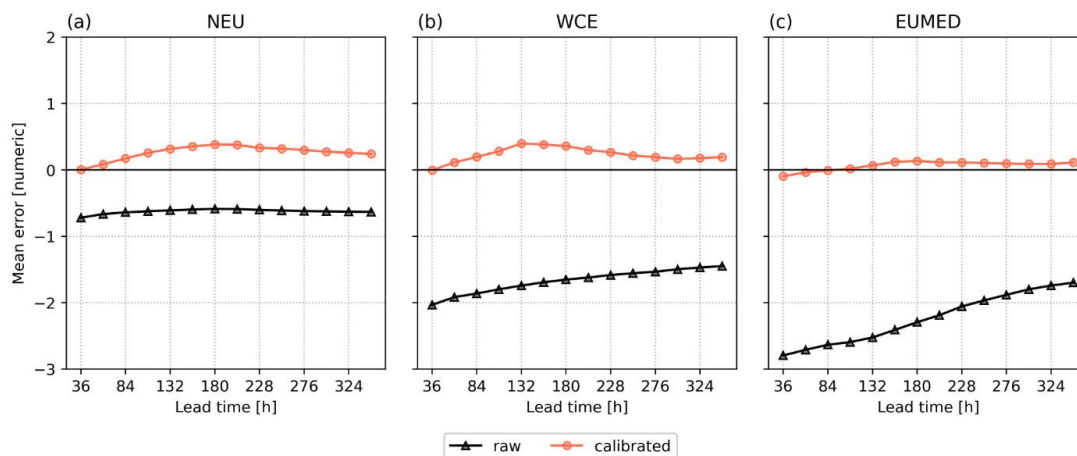

Fig.4

Minor comments (typos and formalities)

13: "prevalent": word choice

We changed the wording to "*frequent*".

16: "But not": cannot start a sentence with "but not"

We rephrased this sentence as follows:

"While the Mediterranean region continues to face the highest occurrence of wildfires, Central and Northern Europe have experienced an increase in extreme temperature events and heatwaves in recent years (Ibebuchi and Abu, 2023; Rousi et al., 2023; Ionita et al., 2017; Barriopedro et al., 2011)."

Furthermore, we added references to support the statement.

18: unnecessary comma after "periods"

The unnecessary comma has been removed.

19 "heatwave 2018": either heatwave of 2018 or 2018 heatwave

We changed it to "the 2018 heatwave".

24 missing (Oxford) comma after "during"

The comma has been added.

24-25: "Accurate and reliable weather forecasts ranging from a couple of days to multiple weeks to identify high wildfire risk areas is an important part of SAFERS": Accurate and reliable weather forecasts, ranging from a couple of days to multiple weeks, are an important part of SAFERS for identifying high wildfire risk areas. (Or equivalent paraphrasis)

We rephrased this sentence and changed the order of the sentence to make it more clear that the weather forecasts are part of SAFERS.

"An important component of SAFERS for identifying high wildfire risk areas are accurate and reliable weather forecasts, ranging from a few days to several weeks."

25 "Here,": "in this paper,"

Wording has been changed to "in this paper"

26 "short FWI": (FWI)

This has been changed.

26: "Wagner, 1987": Author's last name is Van Wagner, throughout the paper

Thank you for pointing this out. This was a mistake in the used Bibtex reference and has been fixed.

29: "One widely used": A widely used

A paragraph mentioning other calibration methods has been added to the introduction following the suggestion of reviewer 2 (comment 2). The wording has therefore changed:

"Various methods have been developed for calibrating probabilistic ensemble forecasts. Commonly used calibration methods include Bayesian model averaging (Raftery et al., 2005), non-homogeneous Gaussian regression (Gneiting et al., 2005), logistic regression (Hamill et al., 2004) and non-parametric ensemble post-processing methods such as rank histogram techniques (Hamill and Colucci, 1997), quantile regression (Bremnes, 2004) and ensemble dressing approaches (Roulston and Smith, 2003). Non-homogeneous Gaussian regression (NGR) is one of the most commonly used calibration methods and adjusts both ensemble mean and spread, while still be efficient and easy to implement. It has been proved effective for various weather variables like temperature (Hagedorn et al., 2008), precipitation (Hamill et al., 2008) or wind-speed (Thorarinsdottir and Johnson, 2012) and can be applied using a truncated or censored distribution to account for a constraint to non-negative values."

36-39 "Although originally developed for Canadian weather and vegetation, it is used in many other regions, e.g., by the European Forest Fire Information System (EFFIS) to provide information on wildfires in the EU and neighboring counties (Giuseppe et al., 2020)": sentence needs to be more orderly and written better; also, author's last name is "Di Giuseppe"

We improved this sentence in the revised manuscript and corrected the author's name.

"Although originally developed for Canadian weather and vegetation, the FWI system is now used in various regions. For instance, the European Forest Fire Information System (EFFIS) employs the FWI to provide information on wildfires in the EU and neighboring countries (Di Giuseppe et al., 2020). "

39-40 "One advantage of using FWI is the relatively simple calculation only requiring four weather parameters in addition to information of the season (time of year) and geographical location": Rephrase, e.g. "The main advantage of using FWI is its relatively simple computation, only requiring four weather parameters and information about the season (time of year) and geographical location"

We rephrased this sentence to:

"The main advantage of the FWI system is its relatively straightforward computation, requiring only four weather parameters and information about the season (time of year) and geographical location as input parameters."

42-43 "the moisture content of three separate fuel layers of different depth and diameter": this is one interpretation of the three parameters (of course, the main one), meaning that -more or less- they contribute to the fire danger with the same time scale of a certain fuel layer. For example, the DC can also be an index of the lack of precipitation for a long time. I tend to be more cautious when interpreting these indices, but it is a relatively small issue.

We rephrased this section to add the depth of the moisture levels (see comment 15, reviewer 2). We also added a comment about the response of the moisture codes to weather variations, please see reply to the first comment.

48 "Often the FWI is classified": the FWI is often classified

We changed the order of this sentence accordingly.

50 "e.g. vegetation types": since this is the second time it appears, I have to point out that you cannot put "e.g." in the middle of a sentence without it being in parentheses or in a parenthetical expression (between commas). This is not formal enough for a paper, in my opinion.

We corrected this throughout the paper. Thank you for this comment!

52-53 "Fuel moisture codes (FFMC, DMC, DC) and consequently FWI values are dependent on preceding conditions. Thus, the preceding days noon values are used for FWI calculations and the calculations need to be initialized": To be put above, together with the input variables, and to be written more clearly.

We moved the sentence about the initial values to the paragraph about the moisture code calculation (see reply to first comment). The sentence about the dataset used for initial values was moved to the Forecast and observation data section (2.2).

62 "The resolution of the used TIGGE data": The resolution of the TIGGE data used in this paper

We rephrased this sentence accordingly.

65 "can not": cannot

This typo has been fixed.

Figure 1 (caption): please provide a reference (IPCC) for the AR5 regions.

We wrongfully called the regions here AR5 when it was supposed to be the AR6-WGI regions. We corrected that and added the reference for the AR6 regions.

90 Formula (2): shouldn't it be log(σ_{kl})?

It should indeed be log(σ_{kl}). We apologize for this mistake and correct the formula.

91-92 "The logarithmic link log(sd)": define sd (ensemble standard deviation?). Besides, isn't it sd_{kl} ?

sd is defined as the standard deviation of the 50 ensemble members in the previous sentence. We agree with the reviewer that it should be sd_{kl} and corrected this mistake.

119 "The bias of the forecast can be accessed by simply evaluating the difference between the average forecast and average observation, which is defined as the mean error (ME)":

The bias of the forecast can be accessed by simply evaluating the difference between the average forecast F_i and average observation O_i, which is defined as the mean error (ME)

We added the symbols F_i and O_i to the sentence.

137 "defined by the 6th IPCC Assessment Report (AR6 (Iturbide et al., 2020))": did you not show the AR5 regions before (Figure 1a)? It needs to be coherent, at least in the name in the caption of Figure 1a, if the regions did not change.

Thank you for making us aware of this mistake. The reference regions did not change and it is supposed to be AR6. We changed it to be coherent throughout the manuscript.

140 "Other regions can be selected as well, e.g. the calibration can also be done country-wise or at even smaller level.": colloquial, rephrase like "the calibration can also be performed over smaller areas (e.g., single countries)"

We rephrased this sentence to:

"The calibration can also be performed over smaller geographical areas, e.g., individual countries."

158 "the forecasted FWI is too low compared to observations": rephrase in a more formal "the forecasted FWI underestimates the observations[…]" or similar.

We rephrased the sentence to:

"Uncalibrated forecasts have a negative bias for all lead times, indicating that the forecasted FWI values are consistently lower than the observed values."

Reply to the comments of reviewer 2

Major comments

1. Study workflow is not very clear, e. g. which datasets are used in which step. This is due to missing research questions and the structure of the manuscript. This should be addressed by:

   a. adding research questions, e. g. can NGR be used to calibrate FWI derived from mid-range weather forecasts, how well does NGR improve the FWI derived from mid-range weather forecast. The research questions should be placed at the end of the introduction, as they help the reader to know what to expect from the following chapters.

Thank you for this comment. We added a sentence stating the research question at the end of the introduction, as follows:

"In this article, we investigate whether non-homogeneous Gaussian regression (NGR) can be used to calibrate the Fire Weather Index (FWI) derived from medium-range ensemble weather forecasts and assess the extent to which NGR improves the accuracy of the FWI predictions."

   b. Restructuring the manuscript by summarizing chapters 2 to 4 to a data and methods section with subchapters (the following is a suggestion):
      i. 1. Introduction, providing research questions at the end of the chapter
      ii. 2. Data and Methods,
         1. 2.1 FWI and FWI calculation
         2. 2.2 Forecast and observation data
         3. 2.3. Validation and Calibration methods
            a. 2.3.1 NGR
            b. 2.3.2 Verification methods
      iii. 3. Results
      iv. 4. Discussion (missing, see other comment.)
      v. 5. Conclusion

We agree that the restructuring of the manuscript makes the workflow better understandable and adjusted the structure of the manuscript. However, we decided to integrate the Discussion into the conclusion as "4. Discussion and conclusion". The new structure of the revised manuscript is as follows:

1. Introduction
2. Data and methods
   2.1. Fire weather index calculation
   2.2. Forecast and observation data
   2.3. Calibration and verification methods
      2.3.1. Non-homogeneous Gaussian regression

2. Introduction: Please add a paragraph why you chose the NGR method over other methods, e.g. other variations of EMOS calibrations or bias-correction methods (see Whan et al. 2021, https://doi.org/10.1016/j.wace.2021.100310).

We added a short paragraph at the end of the introduction mentioning other calibration methods and why we chose NGR.

"Various methods have been developed for calibrating probabilistic ensemble forecasts. Commonly used calibration methods include Bayesian model averaging (Raftery et al., 2005), non-homogeneous Gaussian regression (Gneiting et al., 2005), logistic regression (Hamill et al., 2004) and non-parametric ensemble post-processing methods such as rank histogram techniques (Hamill and Colucci, 1997), quantile regression (Bremnes, 2004) and ensemble dressing approaches (Roulston and Smith,2003). Non-homogeneous Gaussian regression (NGR) is one of the most commonly used calibration methods and adjusts both ensemble mean and spread, while still be efficient and easy to implement. It has been proved effective for various weather variables like temperature (Hagedorn et al., 2008), precipitation (Hamill et al., 2008) or wind-speed (Thorarinsdottir and Johnson, 2012) and can be applied using a truncated or censored distribution to account for a constraint to non-negative values."

3. Fire weather index calculation: It is not clear how the FWI is derived. Please clarify in section 2 how you derive the FWI, i.e. which R-package or python package you are using, as there are differences between cffrds and the ECMWF fire products derived from the ECMWF GEFF- Model (see Vitolo et al. 2019, https://doi.org/10.1038/sdata.2019.32)

We agree that it was not very clear how we calculate the FWI. We added a sentence in the end of section 2.1 stating that we have implemented the FWI calculation use Python programming language and following the source code provided by Wang et al. 2015, (https://publications.gc.ca/site/eng/9.805607/publication.html), as follows:

"For this study, we implemented the calculation of the Canadian Forest Fire Weather Index (FWI) using Python programming language, following the source code provided by Wang et al. (2015), with modifications to utilize gridded input data."

To check if our implementation provides consistent values with other FWI algorithms, we tested our method using the test values given in Vitolo et al. 2019 and default initial values for FFMC of 85, DMC of 6 and DC of 15 (instead of the ERA5 climatology values used in the presented paper). The outputs closely match those of ECMWF fire products and cffrds,

either exactly matching cffrds outputs or falling between the values of ECMWF fire products and cffrds

Code Source	FFMC	DMC	DC	ISI	BUI	FWI
R package cffdrs *	87.69	7.29	17.76	10.85	7.28	9.46
ECMWF *	87.70	8.54	19.01	10.80	8.49	10.10
This study	87.69	7.59	18.76	10.85	7.59	9.62

* values taken from Vitolo et al. 2019

4. Forecast and observation data:

   a. Lines 57 – 65: You state that you derive the FWI from the ECMWFs operational forecast system (ENS). Later you state that you use the TIGGE dataset. Can you clarify if you used the ENS dataset, the TIGGE dataset or both datasets later in your analysis? I understand that the TIGGE dataset has a higher temporal resolution than the ENS dataset, however, the spatial resolution is coarser (0.5° vs 0.2°). Later in your results section you show the earliest results for 36h lead time. Can you add a sentence why you are more interested in the increased temporal resolution of the TIGGE dataset over the spatial resolution of the ENS dataset in your manuscript?

The mentioning of ENS and TIGGE dataset is indeed confusing. In this study, we use the ECMWF ensemble forecasts retrieved from the TIGGE database (spatial resolution of 0.5°), although the method is intended for post-processing the operational ECMWF forecast (spatial resolution of 0.2°) to provide FWI forecasts for the SAFERS project. However, we don't have access to multiple years of ECMWF ENS forecast data with higher spatial resolution. To still perform a statistical analysis, we decided to use TIGGE data instead.

To avoid confusion, we decided not to mention ECMWF operational forecasts. Instead, we introduce in the beginning of the chapter 2.2 "Forecast and observation data" all datasets that are used in this study, i.e., the TIGGE ECMWF forecasts, ERA5 reanalysis and High-resolution forecasts.

"2.2 Forecast and observation data

In this study, we use ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) to calculate ensemble forecasts of the FWI. ECMWF medium-range ensemble forecasts consists of 50-members, initialized twice a day at 0000 and 1200 UTC and provide forecasts up to 360 hours (15 days). The forecasts are derived from the TIGGE archive (Bougeault et al., 2010), which provides operational medium-range ensemble forecast data for non-commercial research purposes. The data is accessible through the ECMWF API1. The temporal resolution of the ensemble forecast data used in this paper is 6h for all lead times and the spatial resolution is 0.5° (~50 km). We use only the forecasts

initialized at 0000 UTC. Although available forecasts cover the whole globe, our focus is on the European region from 25°N to 72°N and 25°W to 39.80°E and we specifically use forecasts from the years 2021 to 2023.

For the FWI calculation, we derive initial values for FFMC, DMC, and DC from ERA5 reanalysis data (Hersbach et al., 2020) to account for preceding conditions at forecast initialization. The initial values are determined using the climatological from 40 years of historical data (1980-2019) for each day of the year, using a 15-day rolling mean centered on each day. ERA5 reanalysis data can be retrieved from the C3S Climate Data Store (CDS) (Hersbach et al., 2017).

For calibration and verification purposes, we use ECMWF high-resolution forecasts initialized at 0000 UTC. 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. The FWI values calculated using ECMWF high-resolution forecasts with the shortest lead time to the local noon with corresponding 24h precipitation act as substitute for actual observations and are hereafter called analysis. Ideally, FWI forecasts would be verified using FWI values calculated using surface observations of the relevant weather parameters as the FWI cannot be observed directly. 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 needs to be available rapidly."

    b.   Lines 65 – 68: These sentences should be moved to the paragraph where you discuss how you verify the FWI from ECMWF data to observation data.

We moved these sentences to fit better in the context of the verification of FWI analysis and observation data. Please see changes to Section 2.2 given in the previous comment.

    c.   Line 68: "We therefore use the FWI calculated ECMWF high-resolution forecasts …". You did not introduce what the ECMWF high-resolution forecasts are yet. You can optimize this by merging this sentence with the next sentence (i. e. "ECMWF high-resolution forecasts have …"), but it remains now unclear why you introduced the ENS and the TIGGE dataset before. Please clarify in this section on which datasets you derive the FWI from and for which later steps you use which datasets (NGR regression and verification).

This section (2.2) has been completely rewritten to make it more clear which dataset is used for which step and to accommodate the changes to Figure 1 suggested by reviewer 1. Please see reply to comment 4a and second major comment of reviewer 1.

5. Results / Figures:

a. All figures with subplots: add labels for subplots, i. e., (a), (b), (c) as suggested in the NHESS publication guidelines https://www.natural-hazards-and-earth-system-sciences.net/submission.html#figurestables)

Labels have been added to all subplots to be in accordance with the guidelines.

b. In your written text you relate to lead times as days while in the figures you show lead time in hours on the x-axis. Can you synchronize this information? In the current state the reader has to transform between written "7 days" to 7*24h in the x-axis of the respective plot. You could write the hour also in brackets next to the days in the text.

We believe the article is easier to follow using lead times in days instead of hours but agree that using both units throughout the manuscript is confusing. We applied the suggested changes and added the lead time in hours after every mention of the lead time in days. Please note that the lead times in hours do not exactly correspond to the typical conversion of days to hours (e.g., 7 days is here 180 hours instead of 168 hours). This is because FWI forecast are issued for 12 UTC while they are initialized at 00 UTC. The first day's forecast has therefore a lead time of 36 hours and not 24 hours.

6. Terminology "short lead times": You state multiple times that your results work best for short lead times. Can you clarify in your manuscript how you define short lead times (e. g. 72h or 132h) or be more specific which lead time you still find good performing (e. g. rephrasing to something like: "for short lead times up to 84h").

Thank you for this comment. We agree that it is not very clear what is meant with short lead times. We clarified in the manuscript that an improvement is visible for all regions up to a lead time of 84 hours, and we added this information to every instance where we mention that the method works best for short lead times.

7. The discussion is missing. In my opinion the discussion is an integral part of a research paper and as the of the study presents a novel way of calibrating mid-range weather forecasts, it would be good to critically reflect on the results:

a. Is the FWI a suitable predictor for fire events in all three regions? What are the challenges regarding wildfire hazard in the three regions? For example, you could address why the postprocessing works particularly well in the MED and summer months of WEU and why not for NEU? Further, you could reflect on how low FWI values, as present in NEU, affect your method.

b. How does your method (NGR) compare to other post-processing methods? Select a two to three different studies, with a similar research question and set your results

in a broader context. For example, you could discuss why you are correcting the FWI instead of the input variables of the FWI, why you chose the NGR method and not a bias correction method or machine learning based method (see Whan et al. 2021 https://www.sciencedirect.com/science/article/pii/S2212094721000086) and Worsnop et al. 2021 (https://journals.ametsoc.org/view/journals/wefo/36/6/WAF-D-21-0075.1.xml?alreadyAuthRedirecting)

 c. What can stakeholders take away from your study. You illustrated quite nicely in the introduction that post-processing helps to make accurate forecasts helping first responders. What do you wish this target group takes away from your results, e. g. will more firefighting resources be placed at locations with elevated FWI?

Thanks for this comment. We decided to incorporate the discussion in the conclusion, which reads now as follows:

"We investigated whether non-homogeneous Gaussian regression (NGR) can be used to calibrate fire weather index (FWI) forecasts based on medium range ensemble weather forecasts by the European Centre for Medium-Range Weather Forecasts ensemble forecasts (ECMWF). To estimate the calibration coefficients of the NGR, we employ a truncated Gaussian distribution with cut-off at zero and use forecast and observation pairs of the last 30 days preceding the forecast. We used ECMWF high-resolution weather forecasts with the shortest possible lead time to calculate the FWI analysis and used these as substitute for observations. Although the FWI analysis seems to underestimate the FWI slightly, a good correlation is observed.

FWI forecasts using medium range ensemble weather forecasts perform generally quite well when compared to the analysis. However, calibration improves the forecasts especially at short lead times up to 84 hours. In the Mediterranean region and Central Europe an improvement of FWI forecast with respect to the FWI analysis is also apparent for longer lead times up to 8 to 10 days, respectively. This is likely caused by the generally higher values in those regions and is supported by the monthly averaged metrics in the appendix, which show a stronger improvement caused by the calibration in the months with high FWI values. Hence, it can be concluded that the calibration performs better for higher FWI values, as indicated by the better performance during the summer months in West and Central Europe and in the European Mediterranean. However, the calibration method shows limitations for low FWI values, which could be observed in the NEU reference region and especially for longer lead times. Although it would be ideal for the method to perform effectively across the entire range of FWI values, it is generally more critical that it demonstrates a good performance at higher FWI values, where the potential fire danger is more significant. While long-range forecasts of potential fire danger are valuable, short-range forecasts, especially for the first 1-3 days, are usually more critical for firefighting resource management. Reliable

and accurate forecasts, particularly when the fire risk is high and over the short-term time frame, are crucial for decision-makers and emergency responders to effectively coordinate resources. 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, e.g., bias correction (Cannon, 2018) or the correction of the input parameters instead of the FWI as in Worsnop et al. (2021). When correcting the individual input parameters, different models need to be applied for the individual variables, which require careful verification and access to quality controlled observation over the whole study region. To further improve the presented calibration method for fire weather index forecasts, it could be tested if calibration of individual components of the FWI system e.g. FFMC, DMC and DC would improve overall skill of the forecast. Furthermore, more advanced models using additional predictors, e.g. elevation or land-use, could improve the calibration but were not tested here."

The take-away message of our study is that NGR can improve FWI forecasts and therefore provide more accurate predictions of potential fire danger. In addition to being illustrated in the introduction, we again mention this now in the conclusion, also. We refrain from giving suggestions for possible actions based on the provided results, e.g. making suggestions for firefighting resources, as this is not the purpose of this article. However, supporting firefighting resources is the objective of the SAFERS EU project, which is mentioned in the introduction.

Minor comments

1. Line 6 & Line9: you use the terms post-processing and calibration interchangeable, please choose one term.

   Thanks for pointing this out. We decided to use the term calibration and changed this throughput the manuscript.

2. Lines 9 – 11: Be more specific about what you mean by short lead times (e. g. 84h?) and regions with elevated FWI (e. g. MEU)

   We included more specific information about the lead time and region as follows:

   "The calibration improves FWI forecast particularly at shorter lead times up to 84 hours and in regions with elevated FWI values, i.e., areas with a higher wildfire risk such as Central and Mediterranean Europe. "

3. Lines 13: I would drop the word "recent" and replace "wildfire in Greece 2023" by "wildfire season of 2023".

We changed the sentence according to the suggestion.

4. Line 15: Drop "Also" at beginning of sentence.

We rephrased this sentence to avoid starting with "Also".

5. Line 16: Drop "But" or make this sentence sound more formal.

We rephrased the whole sentence in order to sound more formal and to accommodate the added references (comment 6).

"While the Mediterranean region continues to face the highest occurrence of wildfires, Central and Northern Europe have experienced an increase in extreme temperature events and heatwaves in recent years (Ibebuchi and Abu, 2023; Rousi et al., 2023; Ionita et al., 2017; Barriopedro et al., 2011). Extended warm and dry periods raise the fire danger and may cause wildfires in regions that were previously not considered wildfire hotspots (San-Miguel-Ayanz et al., 2021; De Rigo et al., 2017). Examples are the 2018 heatwave, which caused wildfires in Sweden (San-Miguel-Ayanz et al., 2019) and across the United Kingdom (Sibley, 2019) or the wildfire outbreaks in Germany and the Czech Republic during the summer of 2022 (Skacel et al., 2022)."

6. Line 17: Provide references for your statement.

References for the increased extreme temperatures and heatwaves have been added to the statement (see comment 5).

7. Line 19: Provide references for your statement.

References are added (see comment 5).

8. Line 20: Here it would be great if more than one example could be provided, e. g. one for each subregion.

We added two more examples, one of which is for the WCE subregion (see comment 5).

9. Lines 22 – 25: Switch the order of the sentence to stress more clearly that weather forecast is part of SAFERS or drop mentioning the project.

We changed the order of the sentence to make it more clear that the weather forecasts are part of SAFERS.

10. Line 26: Add Di Giuseppe et al. 2016 (https://doi.org/10.1175/JAMC-D-15-0297.1) as a reference.

The suggested reference has been added to this sentence.

11. Line 26: Watch out that your citation tool takes the names correctly. It is van Wagner and Di Giuseppe not Wagner and Giuseppe.

Thank you for this comment. This was a mistake in the used Bibtex reference. We checked and corrected all author names with surname prefixes.

12. End of line 27: Here I am missing a short explanation of what is the difference between deterministic and probabilistic weather forecast. A short explanation would be helpful to emphasize that the topic of the manuscript is postprocessing probabilistic forecasts.

We added a short explanation of the difference between deterministic and probabilistic 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."

13. Line 28: drop the word "may" or provide a clear statement whether post-processing is needed. Consider also my first comment on your interchangeable usage of post-processing and calibration. This confuses the reader.

We dropped the word "may" and changed post-processing to calibration.

14. Line 39: Chose a more scientific formulation than "is a relative simple calculation" for the FWI.

We changed the formulation of this sentence to "*straightforward computation*".

15. Line 43: Add the depth of the moisture levels.

The depth and type of fuel layer have been added. All changes to the FWI calculation section (2.1) can be seen in reviewer 1, comment 1.

16. Line 48: Rephrase the sentence starting with often, e. g. "The FWI can be classified"

The sentence has been rephrased to:

"The FWI is often classified into danger classes and values above 50 are considered extreme."

17. Lines 52 – 54: I would change the order of the sentences to make the statement at the beginning of the paragraph clearer, e. g. "we use climatological mean values ..., to account for preceding conditions at the initialization".

    We follow you suggestion and changed the order of sentences.

18. Lines 59, 63 and 70: Please provide an approximation of the grid resolution in km in brackets?

    An approximation of the grid resolution in km has been added in those locations.

19. Line 60: it should be "derived from the TIGGE archive".

    The article has been added.

20. Line 61 and following: Please add dataset after TIGGE, i. e., "The TIGGE dataset ...".

    We completely rewrote the forecast and observation data section (2.2) and don't refer to the TIGGE dataset in those sentences anymore.

21. Line 62: Please add "the" to ECMWF API.

    We added the article "the".

22. Line 62: Please rephrase sentence to "the temporal resolution of the TIGGE dataset ...".

    We rephrased this sentence to:

    "The temporal resolution of the ensemble forecast data used in this paper is 6h for all lead times and the spatial resolution is 0.5° (~50 km)"

23. Lines 65 – 78: Please rephrase this paragraph. Keep the statement that the FWI has multiple input variables. Explain for which later steps you use which dataset to calculate the FWI. Mentioning the station data here is confusing.

    We rephrased the whole section 2.2 (see major comment 4) and hope it now clear which dataset is used for which step. The station data is now mentioned in the second part of this section after introducing all forecast datasets.

24. Line 71: I am not sure which dataset you are meaning by "those", please clarify.

25. Line 75: How many of the 682 stations are in Finland and how many are outside of Finland. Can you provide values.

26. Fig 1 / Line 74 (first mentioned):
    o Add letters for subfigures.
    o Fig 1a
        ▪ Use a different projection, e. g. Lambert Conformal Conic, as the northern latitudes are strongly distorted.
        ▪ It would be very nice to have the stations shown in Fig 1c on the map of Fig 1a as well.
        ▪ Place the region legend (i.e., NEU, WCE, EUMED) inside Fig 1a.
        ▪ Place the legend of the countries (i.e., Finland and others) at a position, where it is clear the legend belongs to both Figures (i. e. Fig 1a and Fig1b)
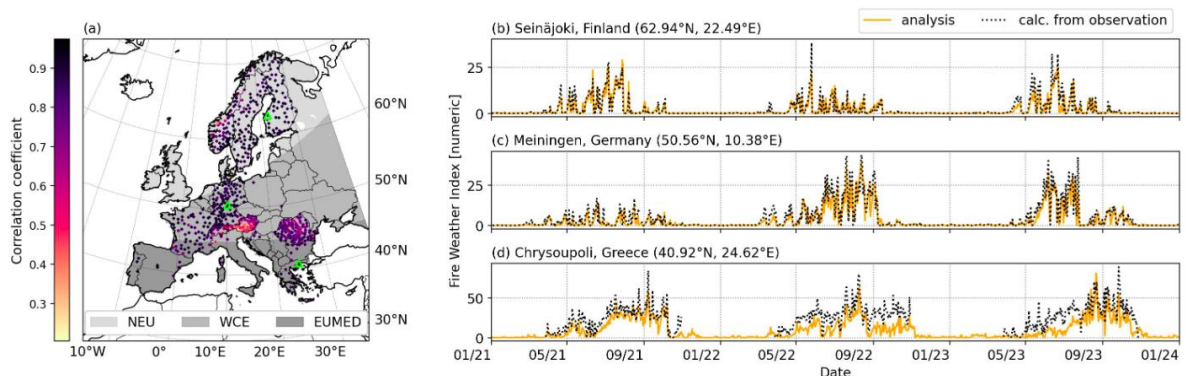
Fig.1

    o Fig 1b:
        ▪ Provide a legend for the regression line. Is this the line for all stations, or for only "other" stations (outside of Finland) or for only Finland stations?
        ▪ Can you provide a line for Finland as well?

There were 2 regression lines in Fig 1b, one for Finland and one for the other stations. Those regression lines were basically identical and therefore not very well visible. Figure 1b has completely been changed according to suggestions of reviewer 1 (major comment) and we now show a separate Figure (Fig.2) with histograms of the FWI values and correlation coefficients.
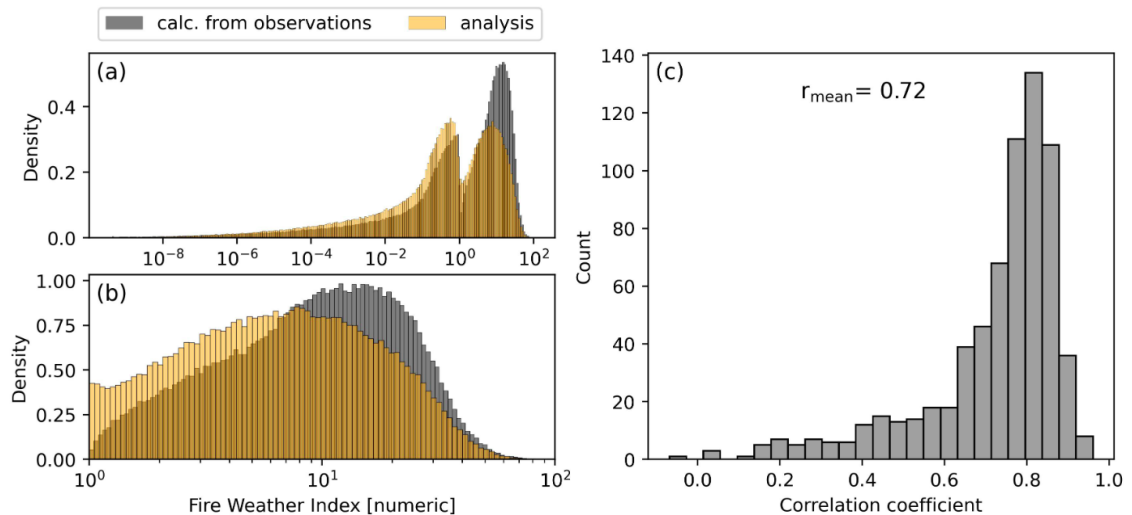

Fig.2

- o Fig 1c:
    - ▪ Please show a 3rd station for WCE.
    - ▪ Add the location of the stations in Fig 1a.
    - ▪ Why are there missing values in the Greece station in the winter of 2022 and 2023? You previously stated that all your selected stations have a sufficient data coverage. Please clarify that your consecutive 200 days refer to the summer half (?) in the Figure caption.

We added a third station (Meiningen, Germany) as an example for WCE and marked its location on Fig. 1a. Accessing data of stations outside Finland with year-round continuous data proved challenging. To expand our analysis to stations outside of Finland, we included stations with some missing data, as long as they had at least 200 continuous days of data. This period isn't restricted to summer months. We chose this 200-day threshold to avoid too frequent reinitialization of the FWI calculation, which could affect the comparison of observation and analysis.

27. Line 77: Provide a reference to Fig 1b as your statements originate from the figure.

    Fig 1b has been updated and mentioned scatterplot is no longer shown in the updated manuscript.

28. Line 77 and Fig 1b: Is your, I assume linearly derived correlation, mainly driven by the large number of low FWI values? Fig 1c shows quite apparent that for high FWI

values the underestimation is much stronger pronounced than for low FWI values. This would be also a good point to be discussed in the discussion.

The correlation is indeed mainly driven by the large amount of low FWI values, which is one reason we decided not to present the scatterplot and linear correlation in the revised manuscript. Instead, we show the distribution of FWI values calculated from observation and analysis using histograms (Fig. 2a,b) and the distribution of the linear correlation coefficient in Fig.2c. Furthermore, we point out that especially for high FWI values the analysis is underestimating the FWI with regard to the observation when referring to Fig.1 b,c,d.

29. Line 79: Provide a reference to Fig 1c.

Fig. 1c has been updated to Fig.1 b,c,d. References to the respective figure are added in the revised manuscript.

30. Line 80: Please clarify which datasets you use for longer forecasts.

We use the medium-range ECMWF ensemble forecast and clarified this in the revised manuscript.

31. Line 83: Please clarify which dataset(s) you mean by the FWI forecasts.

We added "*medium-range*" to the FWI forecasts to clarify which forecasts we mean. With the changes made in Section 2.2, it should now be clearer which dataset we refer to.

32. Line 86: Please specify what you are calibrating, e. g. the parameters of the NGR or the whole post-processing pipeline.

We added that we are calibrating the FWI forecasts to this sentence.

33. Line 86: It should be: "the FWI".

We added the article.

34. Line 91: Please clarify from which dataset the 51 ensemble members come, e. g. by adding (ENS) in brackets. You describe the statistical part very clearly, but it is not clear to which datasets you are applying the formulas.

The ECMWF ENS forecast is no longer mentioned in the revised manuscript. With the changes made to Sect.2.3.1, it should be evident that the calibration is done for the FWI calculated from the ECMWF forecasts (derived from TIGGE database).

35. Line 96: Please specify what the training area is.

In this study the AR6 WGI reference regions are used as training areas. We agree that the training areas need to be introduced already when explaining the NGR method and therefore moved the introduction of the AR6 WGI reference regions from the results section to Sect. 2.3.1.

"The training data is defined using a specific training area and a 30-day rolling window preceding the forecast as training period. As training areas, we use here climatic reference regions, defined by the 6th IPCC Assessment Report (AR6-WGI (Iturbide et al., 2020)), which divide the European domain into Northern Europe (NEU), West and Central Europe (WCE), and the Mediterranean (MED). In this study, however, we use only the European part north of the Mediterranean Sea, referred to hereafter as the European Mediterranean (EUMED). The reference regions are marked grey in Fig.1(a). The calibration can also be performed over smaller geographical areas, e.g., individual countries."

36. Line 97: Why do you switch terminology from NGR to EMOS, I understand that this is the approach, but it would be good to decide for one name.

    We removed the reference to EMOS to avoid confusion.

37. Lines 98 -100: Can you provide results, e. g. a table or a plot, for these findings. This could be part of your supplementary material.

    We added an additional figure in the Supplementary material Fig. S1. This figure shows the average CRPS for a test time period (June-July 2021) using different training areas.

38. Line 100: Here it becomes apparent, that it is not clear what the training area and hence smaller geographical training areas should be. Please clarify and provide results in the supplementary material.

    In the revised manuscript, we now introduce the concept of training areas prior to this statement (see comment 35) and clarify that smaller geographical regions can also be used for calibration purposes. We hope this revision makes the meaning of training areas clear.

39. Line 114: Please add a note that you call RMSE later spread and skill metric (i. e. line 122 and line 144).

    This has been clarified in the text.

40. Line 142: please clarify that you compare fire season length of Northern Europe to Southern Europe.

We added the note that we compare Northern and Southern Europe here.

"The main fire season in Europe is typically from May until October but varies strongly in length and intensity across regions, e.g. the fire season starts later and is shorter in Northern Europe compared to Southern Europe (San-Miguel-Ayanz et al., 2012)."

41. Line 145: the grid points "within" rather than "of" the three study areas.

    We changed the wording accordingly.

42. Line 146: I can't follow how you derived the RSME of the climatology. Can you describe this briefly.

    We added a brief explanation how the climatology RMSE is calculated as follows:

    "The RMSE of the climatology, shown by the solid blue line, is calculated similarly to the forecast RMSE. However, it uses the climatology derived from 40 years of ERA5 reanalysis data (see Sec.2.2) for each day of the year instead of using the forecast."

43. Line 151, Line 163: clarify that your calibration is done as a post-processing.

    This is now clarified in the first sentence of the results section:

    "In this section, we present results of post-processing FWI forecasts of the years 2021 to 2023 by applying the introduced calibration method."

44. Line 152: Specify what you mean by short lead times, e. g. 132h?

    In this specific case, we decided to drop the part of the sentence mentioning short lead times. Short lead times refer only to the NEU region (~132h) while the RMSE of the calibrated forecast is still smaller than the RMSE of the raw forecasts up to 228h for WCE and 300h for EUMED. These lead times are mentioned in the subsequent sentences.

45. Line 154: Provide lead time in hours in brackets after "7 days", i. e. 7 days (168 h).

    We included the lead times in hours in brackets after every following mention of the lead time in days. It has to be noted that the lead time in hours doesn't exactly match the time in days because the FWI forecast is done for 12 UTC using the forecast initiated at 00 UTC. The first day's forecast has therefore a lead time of 36 hours and not 24 hours.

46. Figure 2:
    o Add labels (letters a, b, c) to subplots.

- o Place legend outside of NEU to make it clear it belongs to all three subplots.
- o Adjust the y-axis label (Spread/ RMSE) to your figure caption, which is currently "spread and skill". I would expect them to be the same, e. g. spread and RSME in the figure caption or spread and skill in the y-axis label.

All suggestions have been applied to Fig.2 (now Fig.3). We adjusted it to be Spread/RMSE on caption and y-axis label.
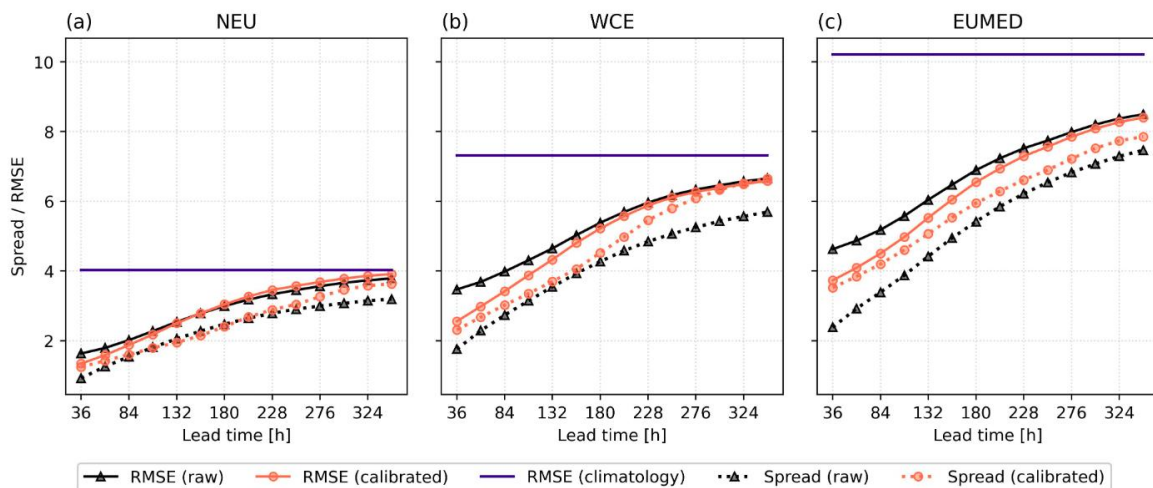

Fig.3

47. Line 157: Rather the three subregions than the respective area.

   The sentence has been rephrased accordingly.

48. Line 158: Rephrase "too low" to something like "lower than observations".

   We rephrased the sentence to:

   "Uncalibrated forecasts have a negative bias for all lead times, indicating that the forecasted FWI values are consistently lower than the observed values."

49. Line 160: "The improvement" instead of "This improvement".

   We changed the article accordingly.

50. Line 164: Provide numbers for what you think is slightly positive.

   We provide numbers of the positive mean error now.

   "In Northern and Central Europe, the bias is slightly positive after calibration especially for longer lead times, ranging from less than 0.1 to 0.4."

51. Line 167: specify short lead times.

    We specified here that lead times up to 84 hours are considered short lead times.

52. Line 166 – 169: This finding would be a good point to discuss in the discussion. For example, I would be interested what these findings imply for the application of your suggested post-processing technique.

    This has been added to the discussion in "4. Discussion and conclusion", see major comment 7.

53. Figure 3: please add letters to subregions.

    The abbreviation of the subregions has been added to the caption.

54. Figure 4: Please add letters to subregions.

    The abbreviation of the subregions has been added to the caption.

55. Figure 4 caption: drop "the grid of".

    "the grid of" has been dropped.

56. Lines 171 – 172: Can you discuss this in your discussion section? Does your approach perform well for higher FWI values, suggested by the better performance in WEU in July and August and MED? Does your approach need to perform well on or low no-fire danger days?

    This has been added to the discussion in "4. Discussion and conclusion", see major comment 7.

57. Figure 5:
    o Add labels to the subplots (i. e., a, b, c, d)
    o Plot the land-sea boundary to improve the visualization.
    o Drop the large white space in the south and west of the plot in such a manner that the plot is filled with results.

    All suggestions to improve the visualization of Figure 5 (now Figure 6) have been implemented, additionally we changed the map projection to be in accordance with the map in Figure 1.
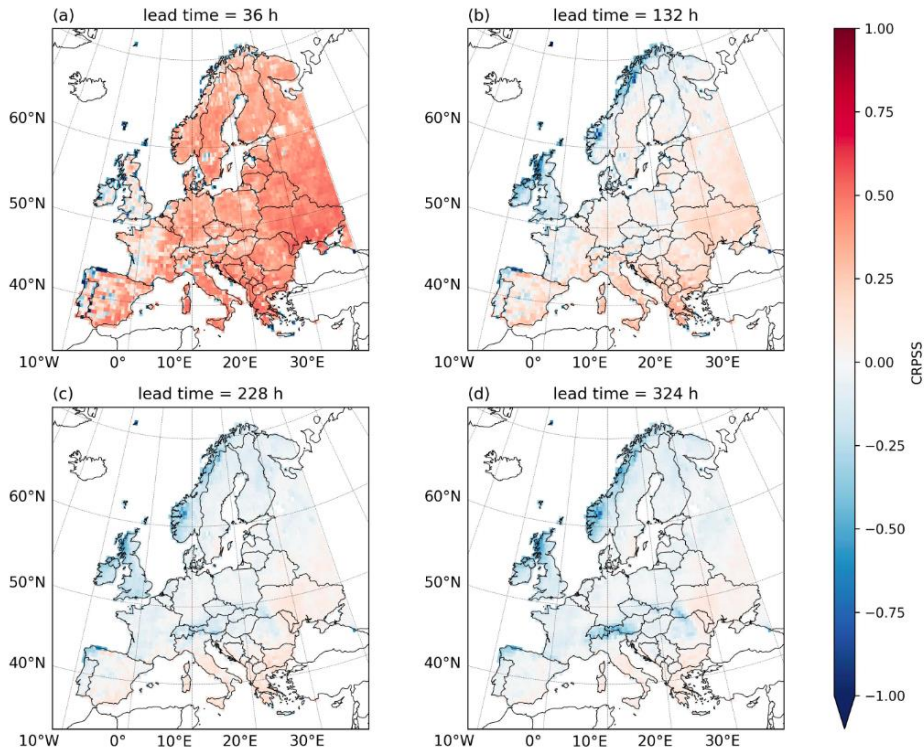
Fig.6

58. Line 176: Here you mention the first time that you calibrate the coefficients of the NGR, this is not clear in your previous description of the post-processing method. Please clarify this in the method section. In Line 176 add "of the NGR" after "to estimate the calibration coefficients".

   We added "of the NGR" to the sentence. The method section has been improved and is hopefully clearer now.

59. Line 177 - 179: This sentence belongs to the discussion section and not the conclusion session. Also, I suggest adding more meaning to this sentence, e. g. what are the implications of sparsely available data?

   We removed this sentence from the conclusion.

60. Line 180: Drop "Thus".

   "Thus" has been dropped.

61. Line 180: which dataset do you mean by high-resolution weather forecast with short lead time? Is this the third dataset you introduced in the data section?

Yes, the high-resolution forecasts are the third dataset that is introduced. We restructured the Forecast and observation data section (now Sec. 2.2.) and hope it is now clear what we mean with high-resolution data throughout the manuscript.

62. Line 181: Make clear that you mean the dataset "analysis" and not the analysis of the FWI.

We added a reference to the previous sentence to make clear that we mean the FWI calculated by the forecast when referring to analysis.

63. Line 189: At the end of your conclusion, I would expect a last sentence coming back to your initial statement that this is important for fire resource management and the SAFERS project. Please add such a sentence.

We added a sentence to the conclusion that highlights the significance of our findings for fire resource management as follows:

"Reliable and accurate forecasts, particularly when the fire risk is high and over the short-term time frame, are crucial for decision-makers and emergency responders to effectively coordinate resources. The improvement of FWI forecasts using the presented calibration method improves the ability to anticipate fire danger, ultimately supporting better response management."

64. Line 209: Add "Di" to "Di Giuseppe".

We fixed this mistake in the reference.

65. Line 249: Add "Van" to "Van Wagner".

We fixed this mistake in the reference.