The authors would like to thank both reviewers for their useful comments which helped to improve the manuscript.

In addition to the changes asked by the reviewers, we have made improvements also in Figure 1. The ROC curves (Figure 1a) and Areas under the ROC curves (Figure 1b) where re-computed, using a bootstrap process (with 1000 repetitions). The envelopes presented in Figure 1b depict confidence intervals and correspond to the 10^{th} and 90^{th} percentiles.

Considering that Figure 8 was split into two news figures (New Figure 7 and 8) and to keep a manageable number of figures, the authors combined in Figure 4 both previous Figures 4 and 5 (case study) without losing its readability.

The revised figures and corresponding figure captions are shown below after the reply to the reviewer comments.

In addition, the manuscript with track changes is also included at the end.

Reviewer #1. Maximiliano Viale

General comment

This paper evaluates the ECMWF ensemble forecasts up to 15 days for AR that make landfall on western Iberian Peninsula. The paper is straightforward, reads well, their results have important relevancies for weather forecasting in the region. My criticism is minimal and relates to the presentation in some parts of the manuscript which may need to be improved. Overall, this article is welcome to the weather forecasting and atmospheric community in the region, and my recommendation is to publish this article in Nat. Hazard journal after considering some minor comments provided below to improve the presentation.

Minor Comments

Line 53: Replace “cost” by “coast”

The typo was corrected.

In the section 3, the comparison of the model output forecasts against the observations were done considering separately the sites or point with observations or using a regional average with observation? Please explain a little bit more about this point.

In section 3, we considered the precipitation averaged in all the observed station precipitation dataset in Portugal to define “yes” or “no” extreme precipitation observations. These “yes/no” extreme precipitation events are compared against forecasts for precipitation and IVT. These correspond to the outputs from the forecast model within the same domain over Portugal for both variables, and a forecast is considered as an extreme one if it exceeds the 95^{th} percentile

Line 132: what period does it correspond to the model climatologies? Please specify.

For all series considered as “extreme” (both forecast and observations) in the ROC curves analysis, we defined thresholds based on the 95th percentiles, and using the longest period available for the

dataset. This is needed to ensure that we obtain thresholds which are representative for the specific realm of each independent dataset (station data VS model data), which obviously have different natures and magnitudes/ranges. Thus, they need to be compared using percentiles, and not absolute values. In the case of the Operational/Ensemble forecasts, we considered the period of available data, i.e. winters between 2011-2012 and 2015-2016. For each forecast day we defined the specific percentiles in the target domain using this period. We acknowledge that, in this context, the use of the word "climatology" might be abusive, using this rather short period. In this sense, this information was included in the new version of the manuscript (see the answer to the following query).

Line 133: what does it mean a sufficient number of ensemble members? Please specify.
The minimum fraction of ensemble members presenting a "yes forecast" varies between 0.1 and 1. So 0.1 means that 10% of the ensemble members have a "yes forecast" while 1 correspond to the totality of the ensemble members. This is how the ROC curves are computed: Hit Rates and False Alarm Rates are calculated repeatedly for each one of these varying thresholds (of minimum ensemble members presenting a "yes forecast"), thus enabling the computation of the data presented in Figure 1.

Considering the last three comments of the reviewer, the first paragraph has been revised as follows:

Firstly, a Receiver Operating Characteristic (ROC, Wilks 2006) curve analysis was performed for IVT and precipitation forecasts for mainland Portugal. To begin with, using the observed precipitation dataset presented in Section 2.2, the mean precipitation (averaged over all mainland Portuguese stations) was computed. Afterwards, a list of extreme precipitation events associated with ARs was obtained by considering observations where the 12h-cumulated precipitation averaged over Portugal (using the surface stations) exceeded the 95th percentile , considering: i) only events with spatially averaged precipitation >0.1mm; ii) that an AR was detected simultaneously in the region (IVT >450kg/m/s), following the threshold found by Ramos et al. (2015) for ERA-Interim reanalysis.
In addition, for the forecasts of extreme IVT and precipitation, we computed the 95th percentile of the corresponding period of analysis (2012-2016). To the computation of the percentile we had into account the data for the winters spanning between 2011-2012 and 2015-2016 and have defined the specific percentiles for each forecast day -1 to -14.
These forecasts are then compared against extreme precipitation observations, considering a "yes" forecast if a sufficient number of Ensemble members surpass that given threshold. The minimum fraction of ensemble members presenting a "yes forecast" varies between 0.1 and 1. So that 0.1 means 10% of the ensemble members have a "yes forecast", while 1 corresponds to the totality of the ensemble members. A ROC curve is then obtained by computing Hit Rates versus False Alarm Rates (Wilks, 2006), and considering these different minimum fraction of Ensemble members above the 95th percentile.

In Fig 3 may be is not necessary adding all subpannels with all the days. Perhaps the authors can incorporate subpanels only every 3 days would be sufficient to show the idea and not overcharge the figure. These are very small and hard to visualize.

We agree with the reviewer that different forecast subpanels where very hard to read. This particular case was chosen to show the relatively larger differences that occur in the IVT field and intensity in the different lead times. Therefore, we choose to maximize the visible area of the sub-panels which allows us to keep the entire set of the forecast, and at the same time increasing the figure readability.

Fig 8 has too much information and of the different type. Considering splitting into two figures: the upper (percentages) and lower (contingency tables) panels. In caption indicate that the percentages correspond to the case study shown in Fig 3. The shading color codes of bars in contingency tables could be discrete (using the 5 subdivision) rather than continuous to better visualize the percentages. The first sentence in caption for these contingency tables could be rewritten as follow: Contingency tables for the accuracy of AR-related IVT forecasts by the ECMWF ensemble system, for lead times ranging between 1 and 15 days during winters spanning 2012-2016.

We agree with both reviewers comments regarding Figure 8. Therefore, it was divided into two separate figures, the new Figures 7 (percentages) and 8 (contingency tables). In addition, all the suggestions of improvement were included in the new version of the figure.

Figure 7 (former upper panel of Fig. 8): Percentage of Ensemble members forecasting IVT above 450 Kg/m/s in each of the regional boxes and for each lead time for the case study presented in Figure 3 (January 4 2016). Green bars represent a spatially accurate forecast (in the box where the maximum IVT was observed). Yellow bars represent a forecast in an adjacent box to where it was actually observed. Red bars represent a forecast in one of the remainder boxes. The bars in the last line represent a completely missed forecast, by either: i) no AR forecast; ii) AR forecast outside of the 6 considered boxes in Western Iberia. (upper panel).

Figure 8 (former lower panel of Fig. 8): Contingency tables for the accuracy of AR-related IVT forecasts by the ECMWF ensemble system, for lead times ranging between 1 and 14 days, during the winters spanning 2012-2016. The red shading represents the percentage of observations versus forecasts. Note that a perfect forecast system would only present shadings in the diagonal, as the y-axis represents observed events in each box (as presented in Figure 2) and the x-axis represents forecasts in each box. The number of events in each box is shown in the y-axis by the blue arrow. The last row/column represent either: i) observations/forecasts outside of the 6 considered boxes; ii) no AR observed/predicted (lower panels)

Reviewer #2

This manuscript quantifies the predictability and forecast skill of winter time atmospheric rivers affecting the Iberian peninsula using ensemble forecasts from ECMWF. Given the impact of precipitation associated with atmospheric rivers in society this is a very worthwhile study. The main results include that integrated water vapour transport is more skilfully predicted than precipitation at longer lead times and that the IFS has a systematic error which results in landfall of the atmospheric rivers being predicted too far north. While most of the manuscript is easy to understand, some parts such as the explanation of the diagnostics and what observations / analysis the forecasts are verified against are hard to understand and lacking critical details. These two major points are further explained in major comments 1-8 and other minor issues and typos which should be addressed are described under minor comments.

Major comments:

1. Section 2.1, lines 111 – 113. Here it is stated that daily values of IVT and precipitation from the IFS are used. Please clarify what is actually done here. I assume for precipitation it is the daily accumulated (so time integrated over 24 hours from 00 UTC - 00 UTC) precipitation but it is not clear what is meant by the daily IVT. Is this also integrated over time (24 hr) at each point?

For the IVT we considered instantaneous values, at 00UTC and 12UTC. These are compared with 12h-cumulated precipitation, centered at that time-steps. For example, 12UTC IVT is compared with precipitation falling between previous 06UTC and the following 18UTC. We acknowledge these details were not sufficiently clear in the original version of the manuscript, so we added this information.

The new version of the text reads as follows:

The data considered here consists in instantaneous IVT values (for both direction and magnitude), at 00UTC and 12UTC, and 12-h accumulated precipitation centered in these time steps. The IVT was computed using the specific humidity and zonal and meridional winds between 300hPa and 1000hPa levels (e.g. Ramos et al., 2015).

2. If IVT is integrated over time, does this act to smooth out (or zonally blur) the ARs and how does this impact the skill scores and the predictability. Previous studies for other forecast variables such as clouds and radiation (e.g. Hogan et al, Tuononen et al, 2019) have shown that while 24-hour integrated values are forecast with a large degree of skill, 6 hourly and 1 hourly values have much less skill.

As mentioned in the previous comment the IVT is not integrated over time, corresponding to instantaneous values, at 00UTC and 12UTC, so we don´t have the smoothing problem as the reviewer mentioned. Please see comment above.

3. Section 2.2, lines 124. Here it is stated that precipitation observations are accumulated into 12 hourly periods whereas the forecast precipitation is 24 hour accumulated values. Is this correct? Please clarify the time accumulations.

We agree with the reviewer that this information was not clear in the text. Both observation and forecast precipitation are accumulated into 12 hourly periods centered at 00UTC and 12UTC. Therefore, the 12UTC embraces the precipitation that occurs between 06UTC and 18 UTC for the same day, while the 00UTC embraces the precipitation registered between 18UTC from the previous day and the 6UTC of the following day.

The 10 minutes precipitation were accumulated into consecutive 12h periods centered at 00UTC and 12UTC of each day. Therefore, the 12UTC embraces the precipitation that occurs between 06UTC and 18 UTC for the same day, while the 00UTC embraces the precipitation registered between 18UTC from the previous day and the 6UTC of the following day.

4. Section 2.2. Were these precipitation observations, that are used to verify the forecasts, assimilated into these forecasts or are these independent observations?

The precipitation observations are only used in section 3. In addition, it must be noted that precipitation station data observations are not assimilated by the ECMWF Integrated Forecasting System, and therefore the observed dataset is totally independent from forecasts. The feasibility of assimilating SYNOP rain gauge data in the ECMWF model has been evaluated (see Lopez, P. Experimental 4D-Var Assimilation of SYNOP Rain Gauge Data at ECMWF. Mon. Weather Rev. 2013, 141, 1527–1544.) showing a very small impact in the operational system due to the large amount of other data already assimilated. The only rainfall product assimilated operationally by ECMWF is the Radar data over USA (Lopez, P. Direct 4D-Var Assimilation of NCEP Stage IV Radar and Gauge Precipitation Data at ECMWF. Mon. Weather Rev. 2011, 139, 2098–2116.).

5. It is hard to follow how the forecasts for IVT are verified. It is said later on in the manuscript that the analysis fields are taken from ECMWF to verify IVT but this should be mentioned much earlier, for example after section 2.2. This is because it is confusing to read how precipitation forecasts will be evaluated but not the IVT forecasts. I am also not sure if the precipitation analysis is used or not, and if not, why not.

Regarding the IVT, the forecast validation is entirely model based, using the analysis field from the ECMWF. We agree that this information was not clearly provided in the submitted version of the manuscript, especially the difference between IVT validation and precipitation validation. This information is now included in the new version of the manuscript.

Based on reviewer #1 comment and also on reviewer #2 comments the fist part of section 3, reads as follows:

Firstly, a Receiver Operating Characteristic (ROC, Wilks, 2006) curves analysis was performed for IVT and precipitation forecasts for mainland Portugal. To begin with, using the observed precipitation dataset presented in Section 2.2, the mean precipitation (averaged over all mainland Portuguese stations) was computed. Afterwards, a list of extreme precipitation events associated with ARs was obtained by considering observations where the 12h-cumulated precipitation averaged over Portugal (using the surface stations) exceeded the 95th percentile , considering: i)

only events with spatially averaged precipitation >0.1mm; ii) that an AR was detected simultaneously in the region (IVT >450kg/m/s), following the threshold found by Ramos et al. (2015) for ERA-Interim reanalysis.
In addition, for the forecasts of extreme IVT and precipitation, we computed the 95th percentile of the corresponding period of analysis (2012-2016). To the computation of the percentile we had into account the data for the winters spanning between 2011-2012 and 2015-2016 and have defined the specific percentiles for each forecast day -1 to -14.
These forecasts are then compared against extreme precipitation observations, considering a "yes" forecast if a sufficient number of Ensemble members surpass that given threshold. The minimum fraction of ensemble members presenting a "yes forecast" varies between 0.1 and 1. So that 0.1 means 10% of the ensemble members have a "yes forecast", while 1 corresponds to the totality of the ensemble members. A ROC curve is then obtained by computing Hit Rates versus False Alarm Rates (Wilks, 2006), and considering these different minimum fraction of Ensemble members above the 95th percentile.

6. Section 4. Line 155. How do you verify the precipitation over the boxes which are located over sea where there are no observation stations?
The precipitation forecast and validation is only done in section 3, where the authors compare the predictive skill of precipitation and IVT. Based on the results obtained in section 3 (higher IVT predictability), all the remaining results sections (i.e. sections 4 and 5) are focused specifically on the IVT (ARs) predictability.

Taking this into account we now stress, at the end of section 3, that only the IVT will be analyzed from this point onwards:

Based on the results presented in Figure 1, we show that the IVT can provide an added value for mid-range operational forecast of extreme precipitation events. Therefore, from this point onward we will focus our analysis on the performance of the ECMWF probabilistic forecasts for IVT and the AR-related IVT forecasts over Portugal, exploring potential systematic biases, and trying to access model behavior and accuracy metrics at different forecast lead times.

7. Section 4.1, lines 165 – 171. It is very hard to understand these diagnostics and as such this is the biggest weakness of this manuscript. This must be improved. Specific points are:
(a) Landfall distance. As this is described (line 165) this is the scalar distance simply measured between two points which in theory should always have a positive value and no direction. However when this is discussed in the text and shown in Figure 4 this parameter can have negative values and a direction. Is this then the difference in the meridional direction with positive (negative) errors indicating a northward (southward) forecast relative to the analysis? Please clarify.
We agree with the reviewer that the wording used in the submitted version of the manuscript lacked sufficient clarity. Regarding the landfall distance, the values presented correspond simply to the meridional distance (in km) between the landfall (location of the maximum IVT) in the analysis and in the forecast. As stated by the reviewer, this value can be positive (negative) errors indicating a northward (southward) forecast landfall error.

(b) How is the landfall location identified? Is this the first point in time when IVT exceeds the threshold value over a land point in any of the boxes? Again please clarify this in the revised manuscript.

The landfall location corresponds to the latitude of the maximum IVT within the coastal area used for detection (the box domains presented in figure 2).

(c) The landfall IVT error is sensitive to both intensity and displacement errors. This should be noted more clearly. It would also be interesting to include a diagnostic which solely measures the intensity error e.g. the difference in the maximum value in the forecast and the analysis regardless of where they occur.

We agree with the reviewer that the landfall IVT error is sensitive to both intensity and displacement errors. That is why we developed three different metrics to test it: (1) Landfall distance and (2) Landfall IVT error and (3) AR-axis IVT error.

Regarding the suggestion raised by the reviewer to have a new diagnostic that measures the intensity error, this is already included in metric (3), which is exactly the difference in the maximum value in the forecast and in the analysis, regardless of where these two maxima occur. We believe that with the new addition to the text, this information will become clear.

(d) The AR-axis angle error. Two points (or a vector) are always need to calculate an angle e.g. you need to identify the axis of the AR yet this is not done here. I do not fully understand how this angle is calculated in the forecast / analysis and therefore I do not understand how the difference can be calculated. I assume it is the angle of the IVT vector but where and when? Please clarify this. A schematic diagram may be helpful as would adding the IVT vectors to the large panel in Figure 3 to make it clearer to readers that IVT is vector and the shading is the magnitude of that vector.

The AR-axis angle is relative to the landfall region, not to the "entire" AR, in this regards it is more appropriate to state that we compute the angle of incidence of the AR in the target area. In this sense, it is not very easy to depict it as suggested in Fig.3, due to the relatively small spatial scale. Regarding its computation, we simply detect the latitude of the maximum IVT for each longitude within the target area. Then, using those latitudes, the "mean" angle is calculated, using a west-east direction as the 0º reference. As for other metrics, this is computed for analysis and forecast, providing the error in the angle. Positive (negative) errors denote a counterclockwise (clockwise) error. We added this information in the revised manuscript, to make it clearer.

Considering the review comments the diagnosis description now reads as follows:

Afterwards, forecasts up to 14 days in advance from the control and ensemble members where compared against the analyses, through the computation of the following metrics that consider the landfall IVT error sensitivity to both intensity and displacement errors:

1) Landfall distance: the meridional distance (in km) between the landfall (location of the maximum IVT) in the forecast and in the analysis. This value can be positive (negative), indicating a northward (southward) forecast landfall error.

2) Landfall IVT error: the difference (forecast minus analysis) between the IVT (in kg/m/s) at the correct location of the landfall, i.e., where the maximum IVT was actually observed in the analysis;

3) AR-axis IVT error: the difference (forecast minus analysis) between the IVT (in kg/m/s) at the specific individual locations of the landfall in the analysis and forecast. It considers the difference in the maximum IVT value in the forecast and the analysis, regardless of where they occur;

4) AR-axis angle error: the difference (forecast minus analysis) between the incidence angle (in º, respective to W→E) at the specific locations of the landfall (Figure 2) in the analysis and forecast. The latitude of the maximum IVT is detected for each longitude within the target area. Then, using those latitudes, the "mean" angle is computed, using a west-east direction as the 0º reference. As for other metrics, this is computed for analysis and forecasts, providing the error in the angle. Positive (negative) errors denote a counterclockwise (clockwise) error.

8. It is not clear how the diagnostics described in section 4.1 are calculated in the cases that no AR is forecast. Are these included as missing data? How does this impact the overall results and conclusions? Please add some information about this.

To create a catalogue of Observed events we only considered analysis where the 450kg/m/s threshold is surpassed within the target area (boxes). However, the maximum IVT (intensity and location) is detected in a much wider latitudinal window (further north/south). For example, in the Case Study presented to explain the methodology (new Figure 4), if the maximum IVT is detected further north/south than the target domain, or if the maximum is below 450kg/m/s, it is considered as no-AR in the domain (as depicted by the open circles). However, regardless of being detected as AR in the domain or not, the maximum IVT at the longitudes where the target area is located (as well as the latitude where that maximum is located) is always kept. These values are used for the computation of "mean statistics" presented in figure 5 and figure 6 (new version of the manuscript), in the same way that values Forecasted as AR are.

Minor comments and typos:

1. Title. I'm not 100% sure this title is grammatically correct. Would "Predictive skill of atmospheric rivers in the western Iberian Peninsula" be more correct?

The title was changed.

2. Line 75. Should read "These kind of studies..."

The sentence was corrected.

3. Lines 77-80. The information presented here about the AR reconnaissance program is somewhat out of place. Either this program needs to be further explain and links made to the research presented in this paper or this should be removed.

We agree with the reviewer that the AR reconnaissance program sentence was not needed in the context of this paper. Therefore, it was deleted from the text.

4. Line 91. What is meant by this statement “The EFI for IVT became control at ECMWF….”? Please clarify the text here. I think it should read “became operational at...”
Thank you for spotting this error. The information was corrected in the text.

5. Line 98 / objective 1. This objective does not make sense. I think what it meant here is to compare the impact of forecast lead time of the forecast values of both IVT and precipitation. Please revise.
The objective 1 was revised in order to become clearer.: “*The main objective here is twofold: a) the comparison between the predictive skill of precipitation and IVT at different lead times during extreme ARs striking western Iberia, using ECMWF ensemble forecasts up to 15 days for winters between 2012/2013 and 2015/16;….. *“

6. It would be helpful to add letters to the panels in the figures and refer to the panels using the letters rather than “upper panel” etc.
We agree with the reviewer’s suggestion. Therefore, letters were added to all the figures with multiple panels. In addition, the text was also changed accordingly.

7. Line 290. There is a typo in the reference here and “to” is missing in the sentence “This is due the….”
Thank you for spotting that. The typo was corrected.

8. Line 309. “control context”. I think what is meant here is “in an operational context...”.
The text was corrected.

Figure comments:
1. Figure 1, top panel. The colour bar is hard to read since it is an continuum. Can this be changed to have discrete colours and only the number of colours that there are lines on this figure (I think 4 colours). The yellow lines are also hard to see so using darker colours would be better.
We agree with the reviewer, and changed the figure accordingly. Also, confidence intervals have been added to panel b), regarding the “area under the ROC curve”, following a bootstrap procedure.

2. Figure 1. bottom panel. What is the grey bar above this panel for?
This was a problem with the figure exporting the title for the panel. We thank the reviewer for noticing it, and it is corrected in the revised version of the manuscript.

3. Figure 3. The boxes are hard to see in the top panel as they are similar colours to the shading. Furthermore, the panels at the bottom are very small and hard the see. These smaller panels would be clearer if the area boxes were removed and if the titles were shortened as this would allow the images to be made larger.
We agree with the reviewer that different forecast subpanels where very hard to read. This particular case was chosen to show the relatively larger differences that occur in the IVT field and intensity in the different lead times. Therefore, we choose to maximize the visible area of the sub-panels, which allowed us to keep the entire set of forecasts, and to increase the figure readability.

4. Figure 4 caption. “Solid blue line represents the error in the location of the maximum IVT between observations and each forecast”. What observations of IVT is available or should this read “...between the verifying analysis and each forecast”. Also see major point 5 above.
We agree with the reviewer that this is not clear in the text. As mentioned in reply to major point 5, the error is between the verifying analysis and each forecast. As mentioned before, figures 4 and 5 were combined in this new Figure 4 and the caption was changed accordingly.

Figure 4. Example of the evolution with lead time for the accuracy of IVT probabilistic forecasts, for the event presented in Figure 3. In a) the black line represents the error in the location of the maximum IVT (i.e. landfall distance) in the Operational run (in km), while the blue thick solid line represents the landfall distance for the Ensemble Forecasts. The blue shaded envelope accommodates the Ensemble spread, considering the 25th and 75th percentiles. In addition, the black arrows represent the errors in the angle (in degrees) of the AR axis for each forecast. Panel b) shows the error in the IVT intensity (Kg/m/s) for each forecast at the observed landfall location. Black solid line, red solid line and red shaded envelope are as in panel a). Panel c) shows the error in the maximum IVT at the specific locations where it has been observed and forecasted for each lead time, regardless of the landfall distance. Black solid line, dashed red line and red shaded envelop as in a) and b). The open circles represented in some lead times represent forecasts where the maximum IVT did not surpass a minimum threshold of 450 Kg/m/s within the target domain (i.e. regional boxes over Western Iberia).

5. Figure 5. The shading is not very clear in the top panel and appears to change shade. Can this be improved? Also please add information to the caption about how the “spread” is calculated. For example, is this the maximum and minimum differences or the 25th - 75th percentile that is shaded?
Figure shading in this figure (old Figure 5, part of new Figure 4) was improved in order to become clearer. The information about the percentiles has been added to the caption.
See previous comment.

6. Figure 8 is very small and hard to see. The caption is also very long and hard to follow. Could this figure be split into two figures e.g. top panel and then the middle and bottom panels as a separate figure?
As suggested by both reviewers, we split Figure 8 into two figures (New figure 7 and 8), and the captions where changed accordingly.

7. Figure 9 is also hard to see and could be made large. The colours could be explained briefly in the caption here rather than expecting a reader to return to Figure 2. e.g. The darkest blue bar represent the most northerly box and the yellow bars the most southerly box.
We agree with the reviewer suggestion. The figure was improved in order to become clearer and the caption was also improved. It reads as follows:

Figure 9. Forecast verification metrics for IVT exceedances (>450 Kg/m/s) using the ECMWF Ensemble forecast system during the 2012-2016 extended winters in Western Iberia, and for lead times between 1 and 14 days. Colored bars represent metrics for individual regional boxes, as

where the darkest blue bar represents the most northerly box and the yellow bars the most southerly box (as depicted in Figure 2).

References:

Hogan, R. J., O'Connor, E. J., and Illingworth, A. J.: Verification of cloud–fraction forecasts, Q. J. Roy. Meteorol. Soc., 135, 1494–1511, https://doi.org/10.1002/qj.481, 2009.

Tuononen, M., O'Connor, E. J. and Sinclair, V. A.: "Evaluating solar radiation forecast uncertainty."
Atmospheric Chemistry and Physics 19.3 (2019): 1985-2000.

Figure Captions

Figure 1. Receiver Operating Characteristic curves (ROC curves) for the IVT and precipitation ensemble forecasts during Atmospheric River days (ARs) from the ECMWF model, using Portuguese surface meteorological stations during the 2012-2016 extended winters (October-March) as a benchmark, and considering events above the 95th percentile (a). The solid lines are for the IVT and dashed lines for precipitation. Different curve colors represent different lead times for the forecasts (1, 5, 9 and 13 days). Area under the ROC curves for lead times up to 14 days (b), where the confidence intervals are also shown. The mean percentage of ensemble members forecasting IVT (pink) and precipitation (purple) above the 95th percentile for lead times up to 14 days during extreme rainfall events associated to ARs (observed precipitation above the 95th percentile associated to an AR over Western Iberia) is shown in (b).

Figure 2. The six regional boxes considered for the verification of IVT probabilistic forecasts in Western Iberia at lead times up to 14 days: i) sea North; ii) Galicia; iii) North Portugal; iv) Central Portugal; v) South Portugal; vi) sea South.

Figure 3. Example of the evolution of the Operational Forecast of the IVT in an event affecting Western Iberia. a) Analysis of the IVT fields on January 4 2016 at 12UTC. In addition, operational forecasts for that date at different lead times, from 1 to 14 days.

Figure 4. Example of the evolution with lead time for the accuracy of IVT probabilistic forecasts, for the event presented in Figure 3. In a) the black line represents the error in the location of the maximum IVT (i.e. landfall distance) in the Operational run (in km), while the blue thick solid line represents the landfall distance for the Ensemble Forecasts. The blue shaded envelope accommodates the Ensemble spread, considering the 25th and 75th percentiles. In addition, the black arrows represent the errors in the angle (in degrees) of the AR axis for each forecast. Panel b) shows the error in the IVT intensity (Kg/m/s) for each forecast at the observed landfall location. Black solid line, red solid line and red shaded envelope are as in panel (a). Panel c) shows the error in the maximum IVT at the specific locations where it has been observed and forecasted for each lead time, regardless of the landfall distance. Black solid line, dashed red line and red shaded envelop as in a) and b). The open circles represented in some lead times represent forecasts where the maximum IVT did not surpass a minimum threshold of 450 Kg/m/s within the target domain (i.e. regional boxes over Western Iberia).

Figure 5. Statistics for the verification of the accuracy of the Operational Forecast of IVT for all events affecting Western Iberia during the extended winters between 2012 and 2016 relative to mean errors (a) and absolute errors (b). Solid blue line represents the error in the location of the maximum IVT between observation and each forecast (in km). The solid red line shows the error in the IVT (Kg/m/s) for each forecast at the real landfall location (where the maximum IVT was observed), while the red dashed curve represents the error in the maximum IVT between the observed and each lead time forecast, independently of the location in each forecast. Black arrows represent the errors in the angle (in degrees) of the AR axis.

Figure 6. Statistics for the verification of the accuracy of the Ensemble Forecast of IVT for all events affecting Western Iberia during the extended winters between 2012 and 2016. a) mean Landfall distance errors (in km) for the Operational forecast (thin black line), the mean of the Ensemble Forecast (thick solid colored line) and the spread of the Ensemble (shading). b) As in a), but for the mean IVT error (in Kg/m/s) at the location of observed landfall. c) As in b), but at the location of the maximum IVT in each forecast.

Figure 7. Percentage of Ensemble members forecasting IVT above 450 Kg/m/s in each of the regional boxes and for each lead time for the case study presented in Figure 3 (January 4 2016). Green bars represent a spatially accurate forecast (in the box where the maximum IVT was observed). Yellow bars represent a forecast in an adjacent box to where it was actually observed. Red bars represent a forecast in one of the remainder boxes. The bars in the last line represent a completely missed forecast, by either: i) no AR forecast; ii) AR forecast outside of the 6 considered boxes in Western Iberia. (upper panel).

Figure 8. Contingency tables for the accuracy of AR-related IVT forecasts by the ECMWF ensemble system, for lead times ranging between 1 and 14 days, during the winters spanning 2012-2016. The red shading represents the percentage of observations versus forecasts. Note that a perfect forecast system would only present shadings in the diagonal, as the y-axis represents observed events in each box (as presented in Figure 2) and the x-axis represents forecasts in each box. The number of events in each box is shown in the y-axis by the blue arrow. The last row/column represent either: i) observations/forecasts outside of the 6 considered boxes; ii) no AR observed/predicted (lower panels).

Figure 9. Figure 9. Forecast verification metrics for IVT exceedances (>450 Kg/m/s) using the ECMWF Ensemble forecast system during the 2012-2016 extended winters in Western Iberia, and for lead times between 1 and 14 days. Colored bars represent metrics for individual regional boxes, as where the darkest blue bar represents the most northerly box and the yellow bars the most southerly box (as depicted in Figure 2).

Figures

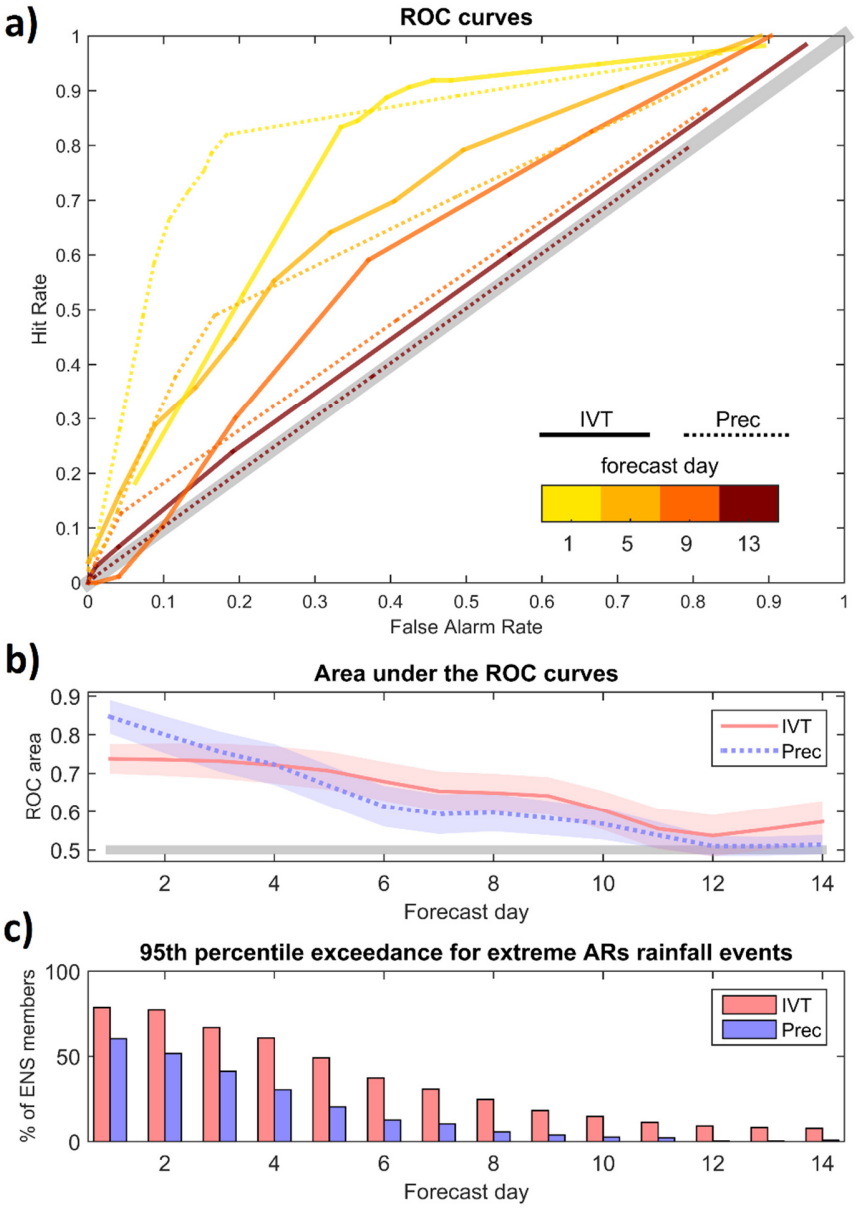

Figure 1.

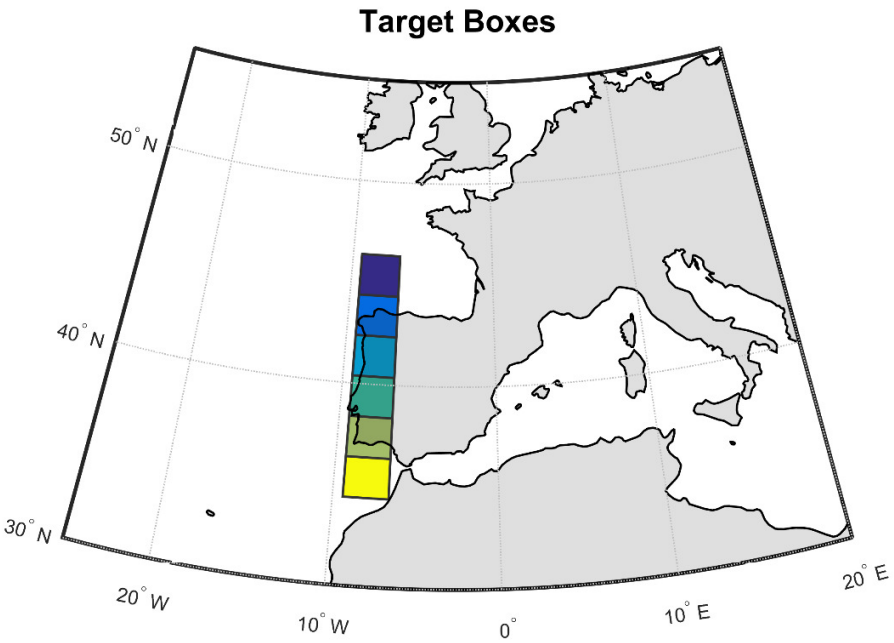

Figure 2.

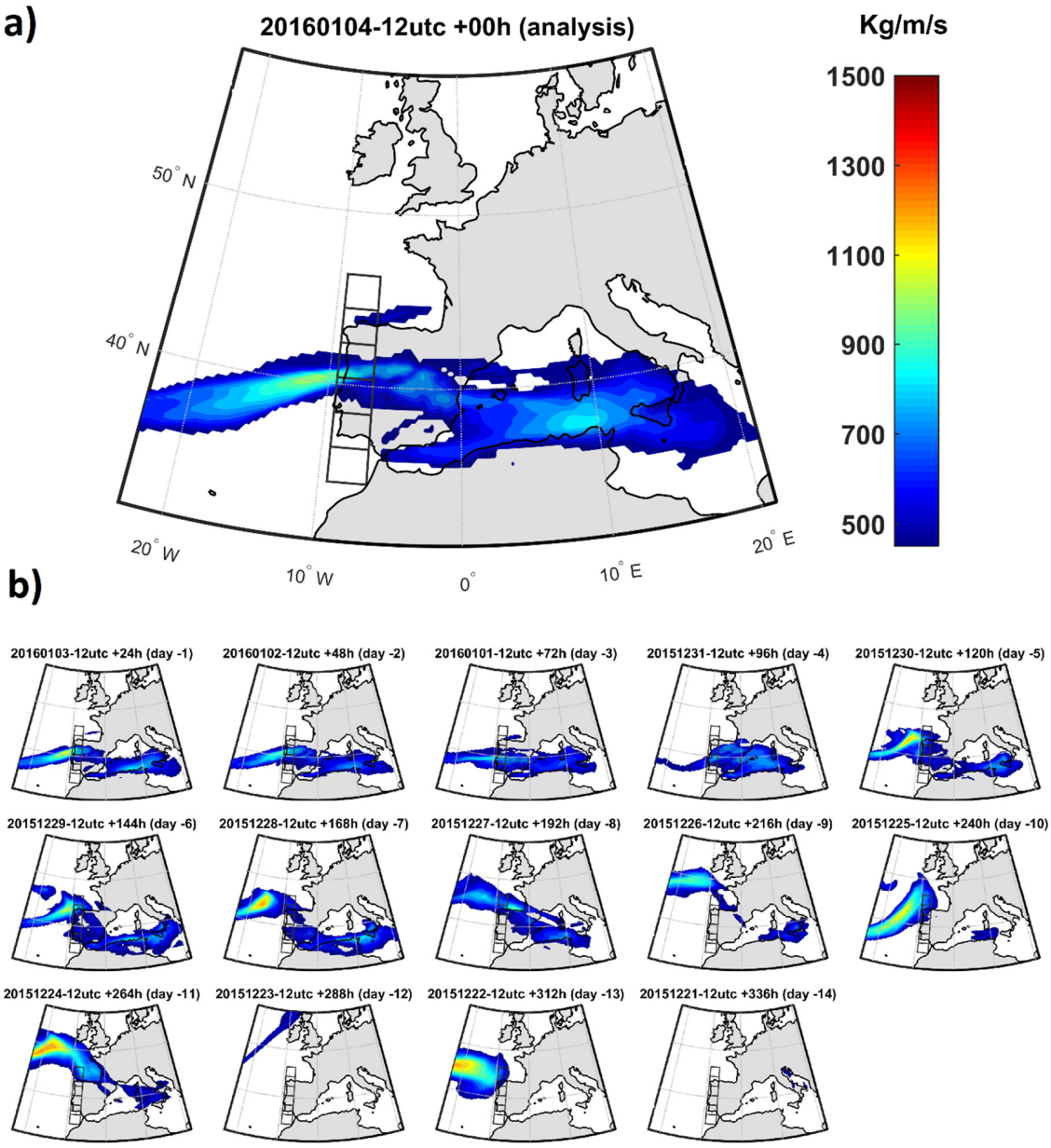
a)
20160104-12utc +00h (analysis)
Kg/m/s
1500
1300
1100
900
700
500
50° N
40° N
20° W
10° W
0°
10° E
20° E
b)
20160103-12utc +24h (day -1)
20160102-12utc +48h (day -2)
20160101-12utc +72h (day -3)
20151231-12utc +96h (day -4)
20151230-12utc +120h (day -5)
20151229-12utc +144h (day -6)
20151228-12utc +168h (day -7)
20151227-12utc +192h (day -8)
20151226-12utc +216h (day -9)
20151225-12utc +240h (day -10)
20151224-12utc +264h (day -11)
20151223-12utc +288h (day -12)
20151222-12utc +312h (day -13)
20151221-12utc +336h (day -14)

Figure 3.

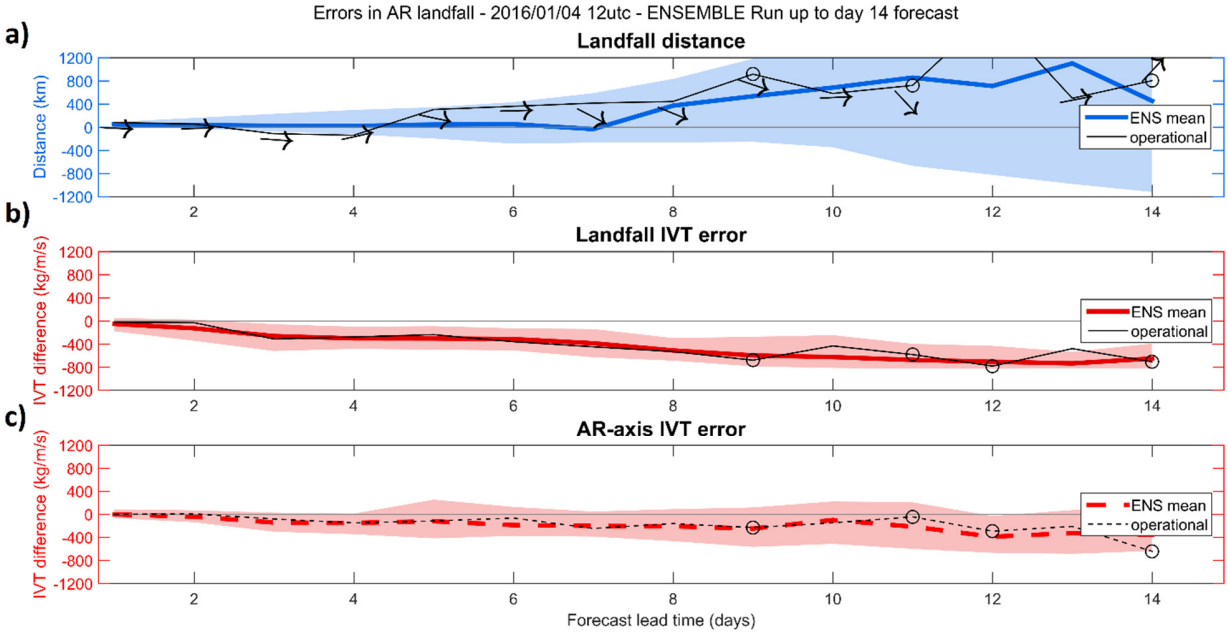
Errors in AR landfall - 2016/01/04 12utc - ENSEMBLE Run up to day 14 forecast
a)
Landfall distance
Distance (km)
1200
800
400
0
-400
-800
-1200
2
4
6
8
10
12
14
ENS mean
operational
b)
Landfall IVT error
IVT difference (kg/m/s)
ENS mean
operational
c)
AR-axis IVT error
IVT difference (kg/m/s)
ENS mean
operational
Forecast lead time (days)

Figure 4.

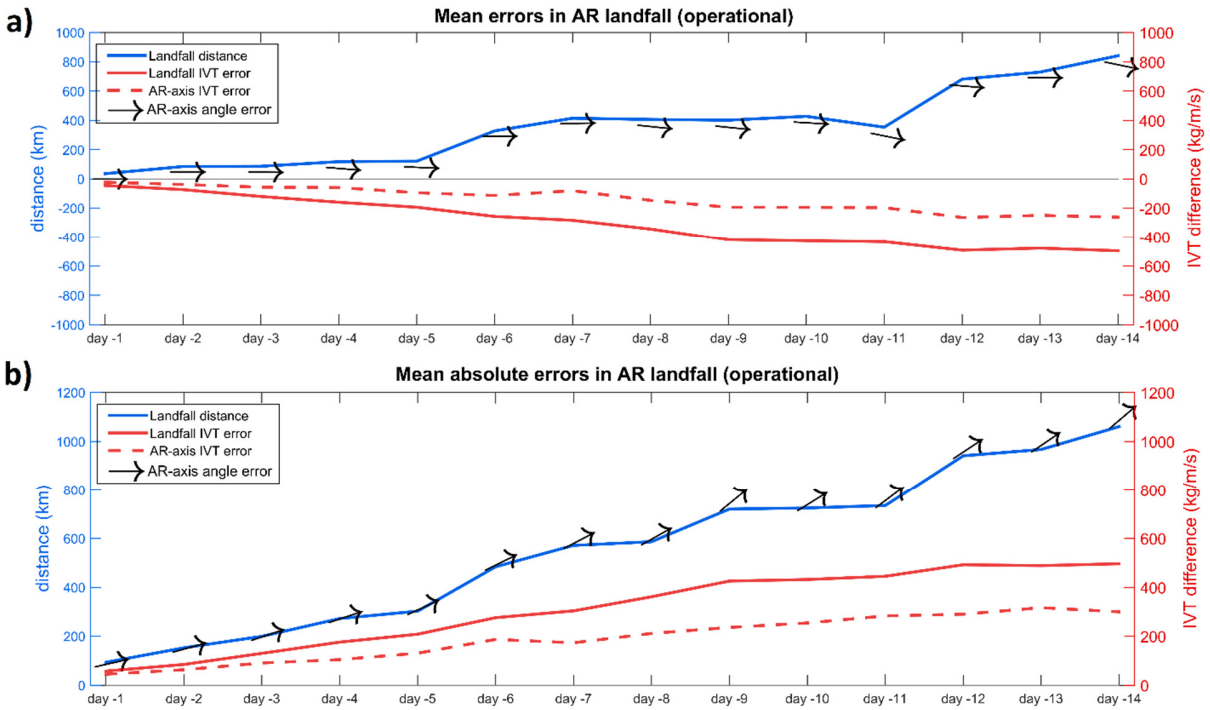
a)
Mean errors in AR landfall (operational)
Landfall distance
Landfall IVT error
AR-axis IVT error
AR-axis angle error
distance (km)
IVT difference (kg/m/s)
1000
800
600
400
200
0
-200
-400
-600
-800
-1000
day -1
day -2
day -3
day -4
day -5
day -6
day -7
day -8
day -9
day -10
day -11
day -12
day -13
day -14
b)
Mean absolute errors in AR landfall (operational)
Landfall distance
Landfall IVT error
AR-axis IVT error
AR-axis angle error
distance (km)
IVT difference (kg/m/s)
1200
1000
800
600
400
200
0
day -1
day -2
day -3
day -4
day -5
day -6
day -7
day -8
day -9
day -10
day -11
day -12
day -13
day -14

Figure 5.

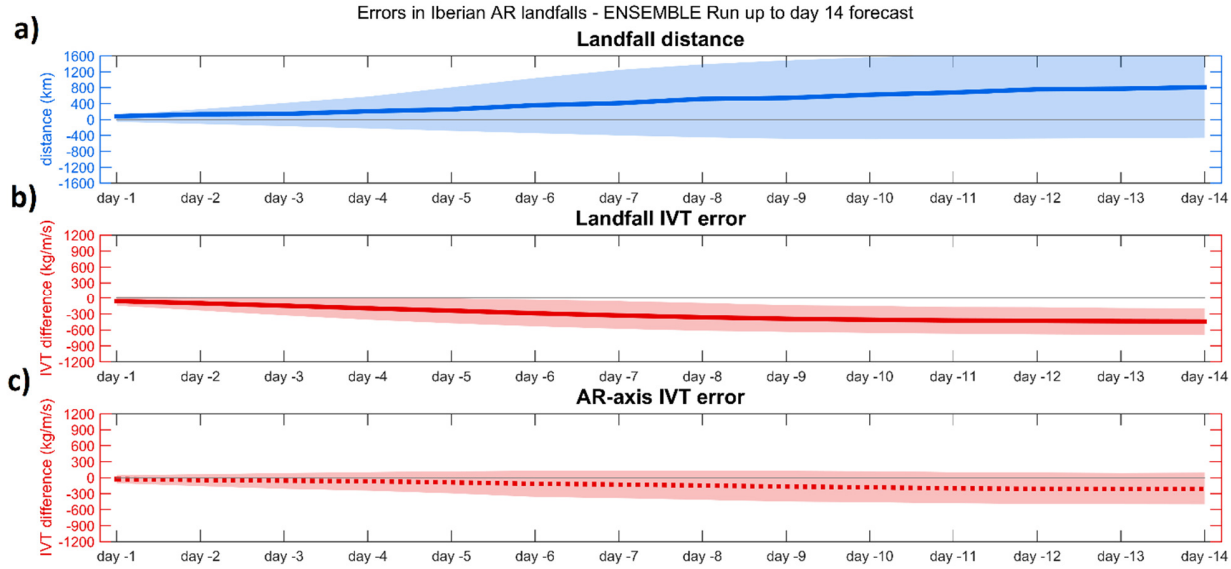
Errors in Iberian AR landfalls - ENSEMBLE Run up to day 14 forecast
a)
Landfall distance
distance (km)
1600
1200
800
400
0
-400
-800
-1200
-1600
b)
Landfall IVT error
IVT difference (kg/m/s)
1200
900
600
300
0
-300
-600
-900
-1200
c)
AR-axis IVT error
IVT difference (kg/m/s)
day -1
day -2
day -3
day -4
day -5
day -6
day -7
day -8
day -9
day -10
day -11
day -12
day -13
day -14

Figure 6.

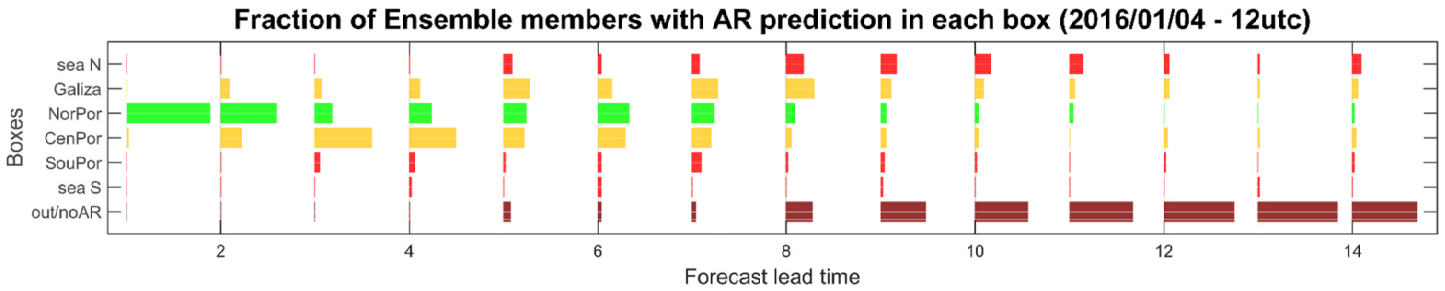
Fraction of Ensemble members with AR prediction in each box (2016/01/04 - 12utc)
Boxes
sea N
Galiza
NorPor
CenPor
SouPor
sea S
out/noAR
2
4
6
8
10
12
14
Forecast lead time

Figure 7.

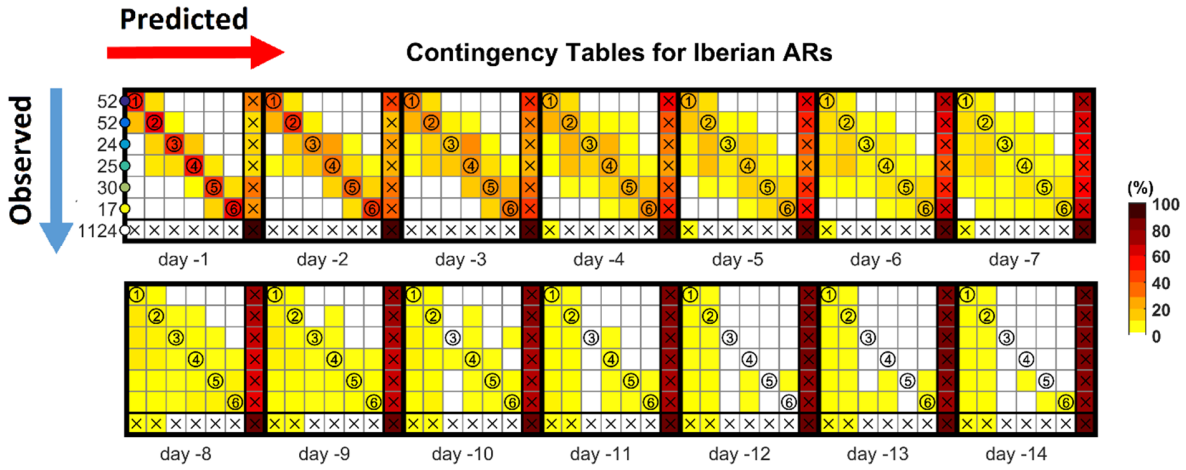
Predicted
Contingency Tables for Iberian ARs
Observed
52
52
24
25
30
17
1124
day -1
day -2
day -3
day -4
day -5
day -6
day -7
day -8
day -9
day -10
day -11
day -12
day -13
day -14
(%)
100
80
60
40
20
0

Figure 8.

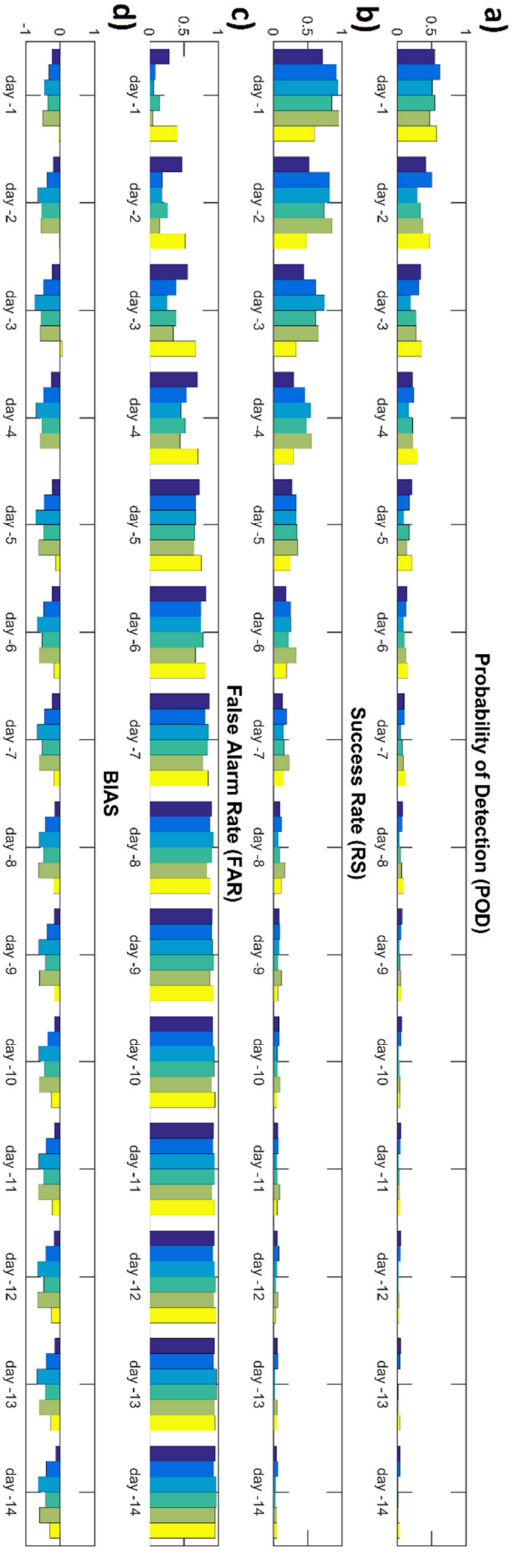
a)
Probability of Detection (POD)
b)
Success Rate (RS)
c)
False Alarm Rate (FAR)
d)
BIAS
day -1
day -2
day -3
day -4
day -5
day -6
day -7
day -8
day -9
day -10
day -11
day -12
day -13
day -14

Figure 9.

Predictive skill of Atmospheric Rivers in the western Iberian Peninsula

Alexandre M. Ramos[1], Pedro M. Sousa[1], Emanuel Dutra[1], Ricardo M. Trigo[1,2]

[1] Instituto Dom Luiz (IDL), Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2]Departamento de Meteorologia, Instituto de Geociências, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941-916, Brazil

Correspondence to: Alexandre M. Ramos (amramos@fc.ul.pt)

Abstract. A large fraction of extreme precipitation and flood events across Western Europe are triggered by Atmospheric Rivers (ARs). The association between ARs and extreme precipitation days over the Iberian Peninsula has been well documented for western river basins.

Since ARs are often associated with high impact weather, it is important to study their medium-range predictability. Here we perform such an assessment using the ECMWF ensemble forecasts up to 15 days for events where ARs made landfall in western Iberian Peninsula during the winters spanning between 2012/2013 and 2015/16. IVT and precipitation from the 51 ensemble members of the ECMWF Integrated Forecasting System (IFS) ensemble (ENS) were processed over a domain including western Europe and contiguous North Atlantic Ocean.

Metrics concerning the ARs location, intensity and orientation were computed, in order to compare the predictive skill (for different prediction lead times) of IVT and precipitation. We considered several regional boxes over Western Iberia, where the presence of ARs is detected in analysis/forecasts, enabling the construction of contingency tables and probabilistic evaluation for further objective verification of forecast accuracy. Our results indicate that the ensemble forecasts have skill to detect upcoming ARs events, which can be particularly useful to better predict potential hydrometeorological extremes. We also characterized how the ENS dispersion and confidence curves change with increasing forecast lead times for each sub-domain. The probabilistic evaluation, using ROC analysis, shows that for short lead times precipitation forecasts are more accurate than IVT forecasts, while for longer lead times this reverses (~10 days). Furthermore, we show that this reversal occurs for shorter lead times in areas where the ARs contribution is more relevant for winter precipitation totals (e.g. northwestern Iberia).

1. Introduction

Extreme precipitation events in the Iberian Peninsula are often due to anomalous vertically integrated water vapor transport (IVT) which are generally associated with an Atmospheric River (AR, Ramos et al., 2015, Eiras et al., 2016, Ramos et al., 2018). According to the definition of the American Meteorological Society glossary of Meteorology, ARs correspond to "a long, narrow, and transient corridor of strong horizontal water vapor transport that is typically associated with a low-level jet stream ahead of the cold front of an extratropical cyclone. The water vapor in ARs is supplied by tropical and/or extratropical moisture sources and these systems frequently lead to heavy precipitation where they are forced upward—for example, by mountains or by ascent in the warm conveyor belt. Horizontal water vapor transport in the midlatitudes occurs primarily in atmospheric rivers and is focused in the lower troposphere" (Ralph et al., 2018).

Extreme precipitation and floods in other regions of the world have been shown to be also associated with ARs, especially on the western continental coastlines of the mid-latitudes (Guan and Waliser, 2015, Gimeno et al., 2016). The preferred regions for ARs to strike are the western coast of the continents like: California (e.g. Gershunov et al., 2017, Ralph et al., 2017, Rutz et al., 2015), South Africa (e.g. Blamey et al., 2018, Ramos et al., 2019), Chile (e.g. Viale et al., 2018, Valenzuela and Garreaud, 2019) the Iberian Peninsula (Ramos et al., 2015, Eiras et al., 2016) or even Southern Australia and New Zealand (Guan and Waliser, 2015, Kingston et al., 2016). However, their climate and socio-economics impacts are significant also in other regions of the world, as western and northwestern Europe (Lavers and Villarini, 2015, 2013, Sodemman and Stohl, 2013) or even the east coast of the US (Mahoney et al., 2016, Miller et al., 2019).

AR impacts are not always hazardous, as they can be also responsible for providing beneficial water supply (e.g. Dettinger, 2013). Lavers and Villarini, 2015 show that for western Europe, the west coast of the United States, and for the central and northeastern United States, the AR contribution to the total precipitation occurring during the winter months is in the range between 30% to 50%. In addition, Ralph et al., 2019, introduced a new scale for the intensity and impacts of ARs along the west coast of the United States, where the authors showed that weak ARs are mostly beneficial, since they can enhance water supply and snowpack, while stronger ARs tend to be frequently hazardous.

Due to its importance in the hydrological cycle, in the last years, there has been an increase in the number of studies dealing with the predictive skill of the different forecast systems in terms of ARs at short and medium range forecasts, as well as seasonal to sub-seasonal scales. Regarding the short (1–3 days) and medium-range (3–14 days) forecasts, most studies are focused on the US. Among them, Nayak et al., 2014 analyzed the skill of global numerical weather prediction models to forecast atmospheric rivers over the central United States showing that, for five different numerical models, AR occurrences are predicted quite well for short lead times, with an increase in the location errors as the lead time increases to about 7 days. The authors show that, as expected, the skill of the forecast decreases with increasing lead time in both occurrence and location.

On the other hand, Martin et al., 2018, examined in detail the skill of a mesoscale numerical weather prediction system (WRF) and compared it with a global numerical weather prediction model (Global Ensemble Reforecast Dataset, Hamill et al. 2013) during AR events for the western United States. It was shown that both models present similar and important errors in low-level water vapor flux, and consequently on the magnitude of precipitation. However, it was found that that WRF (at 9 km horizontal

resolution) can add value to the forecast when compared with a global numerical weather prediction model by means of dynamical downscaling of the medium-range forecast.

Using an adjoint model framework (Errico, 1997), it was shown for short-range forecast that both low-level winds and precipitation, for ARs striking California, are most sensitive to mid- to lower-tropospheric perturbations in the initial state in and near the ARs (Doyle et al., 2014; Reynolds et al., 2019). These kind of studies can help identify locations of greatest sensitivity in the forecast, thus helping to plan observational campaigns that probe ARs using research aircraft and dropsondes; the dropsonde observations will then be assimilated into operational forecast models (Lavers et al., 2018; Guan et al., 2017).

Weather forecasting uses a process called ensemble forecasting to generate multiple realizations of possible future atmospheric conditions or states. This is undertaken to take into account uncertainties in the initial atmospheric state and inadequacies in the numerical model formulations. In recent years a new approach based on the IVT forecasts (Lavers et al, 2014) has revealed that the IVT may provide earlier awareness of ARs and extreme precipitation than precipitation forecasts in different regions of the world (Lavers et al., 2016, Lavers et al., 2017, Lavers et al., 2018). The rationale behind it is to use higher IVT predictive skill (e.g. Lavers et al., 2014; 2016) and then use the ECMWF Extreme Forecast Index (EFI, Zsoter et al., 2014). The EFI assesses how extreme the ensemble forecasts are with respect to the model climate and provide values that range between −1 and 1. Regarding the Iberian Peninsula, Lavers et al., 2018, showed, using a high-density daily surface precipitation observation, for the winters of 2015/2016 and 2016/2017, the IVT EFI has slightly more skill (than the precipitation EFI) in discriminating extreme precipitation anomalies across the western Iberian Peninsula (Portugal and northwestern Spain) from forecast day 11 onwards. The EFI for IVT became operational at ECMWF in the summer of 2019.

Since the ARs are relatively narrow corridors of strong horizontal water vapor transport, its landfall position will influence the location of a possible extreme precipitation event. In the case of the Iberian Peninsula, it was shown by Ramos et al., 2015, that the occurrence (or not) of extreme precipitation days over western river basins is highly sensitive to the latitudinal location of the AR landfall. Therefore, it is important to obtain an objective assessment of the forecast accuracy, at different lead times regarding ARs landfall position by using the IVT. This will be explored here, through a validation procedure that is based on observational precipitation records.

The main objective here is twofold: a) the comparison between the predictive skill of precipitation and IVT at different lead times during extreme ARs striking western Iberia, using ECMWF ensemble forecasts up to 15 days for winters between 2012/2013 and 2015/16; and b) to assess the skill (or accuracy) of IVT probabilistic forecasts in terms of location landfall and intensity, through a probabilistic verification procedure, thus allowing the identification of possible model biases during extreme ARs events, and to define simple metrics which may be suitable for operational purposes.

2. Dataset

2.1 Forecast dataset

The ECMWF integrated forecasting system (IFS) ensemble (ENS) operational forecasts were processed for the extended winter seasons (October to March) in four winters: 2012/2013, 2013/2014, 2014/15 and 2015/16. ENS has a control run and 50 ensemble members. The two daily forecasts initialized at 00UTC and 12 UTC with a lead time of 15 days were processed. The control forecast is produced with the best estimate of the initial atmospheric state. The remaining 50 members are generated by perturbating the initial conditions. The data considered here consists in instantaneous IVT values (for both direction and magnitude), at 00UTC and 12UTC, and 12-h accumulated precipitation centered in these time steps. The IVT was computed using the specific humidity and zonal and meridional winds between 300hPa and 1000hPa levels (e.g. Ramos et al., 2015). The ECMWF operational forecasts in the four winters had several upgrades, including model and resolution updates (accessed 18 September 2019: https://www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model). A detailed evaluation of the impact of these model changes on forecast skill would have required a detailed analysis of the past forecasts (hindcasts) of each model version, which is beyond the scope of this study.

2.2. Observed precipitation dataset

In order to evaluate the operational forecasts, we have used the Portuguese national network of automatic weather stations surface provided by the Portuguese Institute of Meteorology (Instituto Português do Mar e da Atmosfera, IPMA). The data include 10 minutes accumulated precipitation from around 100 automatic weather stations over mainland Portugal, which were chosen based on a combination of tests for completeness and quality. The 10 minutes precipitation were accumulated into consecutive 12h periods centered at 00UTC and 12UTC of each day. Therefore, the 12UTC embraces the precipitation that occurs between 06UTC and 18 UTC for the same day, while the 00UTC embraces the precipitation registered between 18UTC from the previous day and the 6UTC of the following day.

3. Comparing the predictive skill of precipitation and IVT

Firstly, a Receiver Operating Characteristic (ROC, Wilks 2006) curve analysis was performed for IVT and precipitation forecasts for mainland Portugal. To begin with, using the observed precipitation dataset presented in Section 2.2, the mean precipitation (averaged over all mainland Portuguese stations) was computed. Afterwards, a list of extreme precipitation events associated with ARs was obtained by considering observations where the 12h-cumulated precipitation averaged over Portugal (using the surface stations) exceeded the 95th percentile , considering: i) only events with spatially averaged precipitation >0.1mm; ii) that an AR was detected simultaneously in the region (IVT >450kg/m/s), following the threshold found by Ramos et al. (2015) for ERA-Interim reanalysis.

In addition, for the forecasts of extreme IVT and precipitation, we computed the 95th percentile of the corresponding period of analysis (2012-2016). In the computation of the percentiles we took into account the data for the winters spanning between 2011-2012 and 2015-2016 and defined the specific percentiles for each forecast day -1 to -14.

These forecasts are then compared against extreme precipitation observations, considering a "yes" forecast if a sufficient number of Ensemble members surpass that given threshold. The minimum fraction of ensemble members presenting a "yes forecast" varies between 0.1 and 1. So that 0.1 means 10% of the ensemble members have a "yes forecast", while 1 corresponds to the totality of the ensemble members. A ROC curve is then obtained by computing Hit Rates versus False Alarm Rates (Wilks, 2006), and considering these different minimum fraction of Ensemble members above the 95th percentile.The area under the ROC curve above 0.5 denotes skilful forecasts in respect to climatology. In Fig. 1, the ROC curves and areas for forecasts at different lead times are presented.

The ROC curves analysis (Fig. 1 a) clearly shows that lead time is crucial for the performance of both IVT and precipitation forecasts. For both variables, the forecast skill becomes negligible beyond 10 days lead time. While this is not unexpected, our results show that IVT can add potential value to extreme precipitation forecasts during ARs in the mid-range time frame (5 to 10 days). As Fig. 1 b), around day 4-5, ROC areas are higher for IVT than for precipitation which is also confirmed by the confident intarvals also shown. This intervales were computed using a bootstrap process (with 1000 repetitions). This result is in line with a previous work by Lavers et al. (2018), which already highlighted the potential of probabilistic forecasts for IVT (over precipitation forecasts) in western Iberia for 2 winters over longer lead times. Furthermore, as shown in Fig. 1–c), when considering those days with extreme precipitation associated to ARs in Portugal, the number of ensemble members warning for a forecast above the 95th percentile is higher for IVT than for precipitation, for all lead times. While during the days where the precipitation ROC areas are above the IVT (<5 days) which likely reflects the more False Alarms using the IVT, it is clear that between days 5-10 the use of probabilistic IVT forecasts can give a significant added warning value for extreme precipitation forecasts.

Based on the results presented in Fig. 1, we show that the IVT can provide an added value for mid-range operational forecast of extreme precipitation events. Therefore, from this point onward we will focus our analysis on the performance of the ECMWF probabilistic forecasts for IVT and the AR-related IVT forecasts over Portugal, exploring potential systematic biases, and trying to access model behavior and accuracy metrics at different forecast lead times.

4. Model bias during extreme ARs landfall events

We have defined 6 regional boxes (Fig. 2) over Western Iberia, 3 of them covering the area where IPMA surface stations are located (North Portugal, Central Portugal and South Portugal), and also extending further north/south (Sea-North, Galicia and Sea-South), in order to define metrics for the accuracy of the location of ARs landfall, including "hits" and "misses" in forecasts.

4.1. An exemplificative case-study (January 4 2016)

To check the model performance for IVT forecast over the domain during AR events, we considered only cases where the analysis (+00 hours forecast) exceeded the 450kg/m/s threshold for IVT (as defined in Ramos, 2015 - where the ERA-Interim dataset was used). These cases were considered as the

"benchmark" for model forecast verification, being performed for 00UTC and 12UTC analysis during the 4 winters spanning 2012-2016. Afterwards, forecasts up to 14 days in advance from the control and ensemble members where compared against the analyses, through the computation of the following metrics that consider the landfall IVT error sensitivity to both intensity and displacement errors:

1) Landfall distance: the meridional distance (in km) between the landfall (location of the maximum IVT) in the forecast and in the analysis. This value can be positive (negative), indicating a northward (southward) forecast landfall error.

2) Landfall IVT error: the difference (forecast minus analysis) between the IVT (in kg/m/s) at the correct location of the landfall, i.e., where the maximum IVT was actually observed in the analysis;

3) AR-axis IVT error: the difference (forecast minus analysis) between the IVT (in kg/m/s) at the specific individual locations of the landfall in the analysis and forecast. It considers the difference in the maximum IVT value in the forecast and the analysis, regardless of where they occur;

4) AR-axis angle error: the difference (forecast minus analysis) between the incidence angle (in º, respective to W→E) at the specific locations of the landfall (Fig. 2) in the analysis and forecast. The latitude of the maximum IVT is detected for each longitude within the target area. Then, using those latitudes, the "mean" angle is computed, using a west-east direction as the 0º reference. As for other metrics, this is computed for analysis and forecasts, providing the error in the angle. Positive (negative) errors denote a counterclockwise (clockwise) error.

The relevance of these metrics can be easily understood looking at a case study. In Fig.3 ((a), analysis), an example of an AR affecting the North Portugal box is presented. The 14 small panels (Fig. 3b) show how the control forecast changed with increasing forecast daily lead time. While at short lead times the location, intensity and angle are quite similar to the analysis, at longer lead times the control forecast becomes less accurate, and some of them predicting an IVT magnitude, axis orientation and landfall totally disconnected from reality. This example higliths the importance of considering ensemble forecasts for long lead times, as discussed below.

The evolution of the forecasts is depicted in Fig. 4, where the metrics for each lead time are summarized using both control and ensemble. As lead time increases, it is notable how the AR was being predicted further north by the control forecast (Fig 4a) , black line). In fact, for lead times of -9 days, -11 days and -14 days, there was no predictive skill of an AR affecting the defined boxes (either forecasted much further north, or not forecasted at all, as depicted by the open circles), and in accordance with the corresponding subplot observed in Fig. 3. Consequently, as the landfall distance error increases, the IVT error at that location also increases (black line line in Fig. 4b). When considering the AR-axis IVT error (Fig. 4c, dashed black line), the decrease with longer lead times is smaller, showing that despite the error in the actual position of the AR, its maximum intensity was well forecasted for most of the period. Regarding the angle of incidence of the AR, it was mostly zonal, both in the analysis and most forecasts, thus with small error. Nevertheless, and as seen in Fig. 3, during mid-range lead times (around 7-10 days) forecasts tended to tilt the AR NW→SE over Western Iberia, as depicted by the arrows in Fig. 4a).

Moreover, we computed the same metrics described above for each of the 50 ensemble members of ENS, being these results are also presented in Fig. 4. The errors of the ensemble mean are presented for landfall distances (blue thick line) and Landfall IVT error (red thick line) and AR-axis IVT error (dashed black line)

as well as the spread in the ensemble forecast (shaded areas), shown here as the 25th and 75th percentiles of the ensemble computed metrics distribution. The control forecast errors and the ensemble mean are in good agreement with the dispersion of the ensemble forecast increasing with lead time. Regarding the landfall distance, it was found that the error increases with lead time, and in this case the control forecast error at lead time -12 days is higher than the ensemble mean and even the ensemble dispersion. The 50 members of the ensemble for lead time -5 day (Fig. S1) and for -12 days (Fig. S2) are shown in supplementary material Fig. S1 where it can be compared with the control IVT forecast shown in Fig. 3. One can see, that the ensemble members for the shorter lead time are in better agreement and closer to reality. When looking to -12 days lead time (Fig. S2), the dispersion of AR location is much higher when compared with the reality and even with the control forecast shown in Fig. 3. In addition, both Landfall IVT error and AR-axis error (Fig. 4), for both control and ensemble members, show a decrease in the forecasted IVT as we consider longer lead times.

4.2. Mean performance of the ECMWF forecasts during 2012-2016

In the previous section we have analyzed one case study, evaluating both control and ensnemble forecasts against analysis. We extended the evaluation of the IVT forecast metrics for all events occurring during the extended winters of the study period (from 2012-2013 to 2015-2016). Similarly to the case study presented in Fig. 4., the same metrics have been computed for all AR landfall cases (207), and the average error is obtained as presented for the control forecast in Fig.5 and for the ensemble forecast in Fig. 6.

When analyzing the control forecast of ARs over Western Iberia (Fig .5), it is possible to identifysome systematic errors/biases. Regarding the mean errors (Fig. 5a) a northward landfall bias is systematically found for longer lead times, especially for those longer than 5 days, which can reach up to 800km at +14 days. In addition, regarding the AR-axis angle error, the results show a slightly negative bias in respect to those actually observed. Since we consider the 0º angle as west-east orientated and that AR events over Portugal tend to present a southwest-northeast orientation (Ramos et al., 2015), this systematic bias reflects a lower tilt in the AR orientation at longer time forecasts, or in other words, a tendency for more zonal forecasts (Fig. 5a). When analyzing the IVT magnitude, both landfall IVT error and AR-axis IVT error have a negative bias, as a result of: i) the error in the landfall location; ii) an underestimation of the ARs intensity. Comparing both metrics in more detail, it can be noted that the ARs-axis IVT bias is lower than the landfall IVT error, showing that while the IFS is forecasting the intensity of the ARs quite well (with just a small underestimation in intensity associated to lead times), the landfall location bias leads to significant IVT forecast errors at the location where the AR actual landfall is observed. The mean absolute errors were also computed for the same metrics (Fig. 5b), unsurprisingly presenting higher errors for landfall distance, location biases occur both northwards and southwards, thus partially canceling themselves, as shown in Fig. 5a. Nevertheless, this difference is not that large, thus reinforcing the systematic tendency for a bias towards the north in longer lead time forecasts. A similar rationale is applicable to the incidence angle, where errors in the tilt of the ARs in different directions tend to cancel out. As so, mean absolute errors tend to be around 45º at longer lead times (Fig. 5b). Overall it is possible to affirm that the case study evaluated in Fig. 4 presents similar biases to those obtain here with the average of the entire set of ARs considered.

The mean weighted errors/biases of the ensemble forecasts (i.e. all the 50 ensemble members) for all events are presented in Fig. 6. The same methodology as in Fig. 4 is followed here, and the ensemble

mean is presented, along with the spread in the ensemble forecast (shown here as the 25th and 75th percentiles forecast distribution). The results are very similar to the ones found for the control forecast (Fig. 5) with a positive bias (northly) in the position of the AR landfall as we move to higher lead times, along with a negative bias (less AR intensity) in the landfall IVT error and AR -axis IVT error. However, the dispersion in the ensemble forecast is higher in the landfall distance than the in the other IVT error metrics, reinforcing that the model forecasts the ARs but lacks skill in forecasting the right location of their landfalls.

It is vital to stress the use of the ensemble forecast in this kind contitions. In Fig. 4 and Fig. 6., we compare the control forecast with the ensemble mean and the ensemble errors metrics. Both control forecast and ensemble forecast Landfall and IVT error shows a northly bias and a negative IVT bias on the landfall location, respectively. Ramos et al., 2016, showed that the number of ARs affecting Iberia is relatively lower when compared with the contiguous western France or even the UK using the ERA-Interim reanalysis. This means that on average most of the AR go further north, and the ones hitting Iberia are not that frequent. Taking this into account, one can hypothesize~~,~~ that the northerly bias and respective negative bias in IVT intensity in the ARs forecasts at longer lead times can reflect that the model tends to its climatology, which is to have ARs further north, as shown in Ramos et al., 2016. A similar behaviour occurs for the blocking frequency (Euro-Atlantic sector and the Pacific sector) using the the NCEP Climate Forecast System version 2 (CFSv2), in which for longer lead time the model tends to reach its climatology (Jia et al., 2014).

5. Objective verification of the IVT forecasts

In the previous section we proposed some metrics to analyze control and probabilistic IVT/ARs forecasts errors in the IFS for the Iberian Peninsula. In this section, we provide an objective verification of the IVT forecast, but considering in greater detail the landfall location and how to use it for a possible application in terms of control forecast for ARs landfall. Firstly, and bearing in mind the regional boxes presented in Fig. 2, and the case study presented in Fig. 3, the percentage of ensemble members providing correct/incorrect forecasts regarding the regional boxes is summarized in Fig. 7. As it can be seen, the forecast issued the day before the event was almost perfect, with most members predicting the location correctly in the North Portugal box (green bar). As lead time increases, the percentage of correct forecast decreases until day 5, but still a large fraction of members predicts the AR to make landfall in one of the adjacent boxes (yellow bars), until around day 7. At longer lead times, the percentage of members predicting landfall in boxes further away (red bars) and predicting no landfall or no AR (brown bar) increases significantly.

We now present the contingency table for the ensemble forecasts of IVT for all the events (>450 kg/m/s) during the 4 winters considered in this study (Fig. 8). Each different box corresponds to a lead time (from 1 to 14 days) and the different boxes corresponds to the observed vs predicted landfall location corresponding the bluish colors correspond to the location of each landfall box shown on Fig. 2. The box for lead time 1 day presents additional info to help reading the contingency tables: i) the x-axis representing “yes” forecasts for each box; ii) the y-axis representing “yes” observations for each box; iii) the color code presented for day +1 corresponds to the boxes presented in Fig. 2); iv) the last column and row, with the white circle and crosses, represent events that have been observed but not predicted and

vice-versa, respectively. Note that the numbers in the left axis represent the number of observed events in each box. A perfect forecasting model would only present values in the diagonal (NxN).

Results confirm what was partly shown in Fig. 6, as the error in the landfall locations increases with the lead time, and an increasing fraction of the ensemble members forecasting the landfall outside the Iberian Peninsula (further north of Galicia or further south of Algarve) or not even forecasting an AR. However, for shorter lead times (day -1 to day-3) the forecast error is quite low, with the ARs landfall being predicted very well, considering the small size of the regional boxes (less than 1º latitude each). In addition, the contingency table also confirms the northward landfall bias of most forecasts, with the left side of the table being more populated, meaning that forecasted location is more frequent in the northern boxes, when compared to observations. It is also shown that a few ensemble members pick up the AR, therefore it can be argued that system IFS is skillful, however with low probability of occurrence in the right location.

Finally, different verification forecast metrics widely used are also computed for the different ARs landfall cases, and for each box (as in Fig. 2). The metrics used are the probability of detection (POD), Success Rate (RS), False Alarm Rate (FAR) and the BIAS (Wilks, 2006) and their formulation is shown in the supplementary material (Fig. S3). As mentioned before, and due to the increase in landfall error with lead time, these systematic errors are also expected to be reflected in the verification forecast metrics. Both POD and RS decrease as we move further ahead in lead times, getting closer to zero from lead time higher than 5 days (Fig. 9). On the contrary, the FAR is expected to increase with the lead time, staying above 0.5 in all boxes from lead times of 5 days or more. The relatively fast decline with lead time in these metrics is not surprising, as they are computed for very small target areas, and just reflects that a very accurate forecast of landfall location becomes difficult at lead times longer than 5 days, i.e. when considering the mesoscale. Still, as shown before at a synoptic scale, the model is able to forecast high probabilities of an AR affecting Western Iberia at longer lead times. This suggests that an effective warning system can be developed with reasonable lead times, although very detailed local forecasts of AR activity can only be achieved at short time scales.

6. Conclusions

The occurrence (or not) of extreme precipitation days in different river basins is highly sensitive to the latitudinal location of the ARs landfall as shown for the Iberian Peninsula (Ramos etl., 2015). This is due to the ARs being relatively narrow corridors of strong horizontal water vapor transport, therefore their landfall position has influence on the occurrence of a possible extreme precipitation event and its specific location. With this in mind, we assessed the forecast accuracy at different lead times regarding ARs landfall position, intensity and incidence angle by using the IVT. To achieve this goal, we used the ECMWF operational ensemble forecasts up to 15 days, for extended winter seasons between the winters of 2012/2013 and 2015/2016, and assessed the skill (or accuracy) of IVT probabilistic forecasts through a probabilistic verification procedure.

The main conclusions are as follows:

- The IVT forecasts shows higher predictive skill than precipitation forecasts for lead times higher than 5 days, when considering extreme precipitation events associated to ARs over Portugal. In addition,

we show that there is a higher agreement amongst the IVT ensemble members for early awareness at such lead times, than that found for the precipitation ensemble.

- We identified the systematic errors in AR forecasts using a designed objective verification scheme for IVT/ARs applied to the ECMWF Ensemble. There is a good predictive skill of the model in terms of ARs landfall over the domain at short term forecast. However, at longer lead times, the location of the landfall is less reliable~~y~~, and ARs landfall tends to be predicted too much to the north in western Iberian Peninsula, and their intensity tends to be underestimated.

- In addition, when using the ensemble members to check the forecast skill for the specific ARs landfall location (using 6 regional boxes), it becomes clear that the predictive skill at larger spatial scales (the entire domain) tends to be reasonable, while the predictive skill at regional scales tends to be considerably smaller.

- These results show the potential added value to forecast mid-range AR related precipitation events using the IVT, as well as the possibility to develop warning systems based on IVT ensemble forecasts.

Accordingly, we presented a methodology that can be used in an operational context, consisting on the probabilities of ARs striking different regional boxes. This probability is based on the fraction of Ensemble members providing IVT forecasts above a threshold in each box, thus providing an estimate on the probabilities of occurrence and on the expected landfall location. As our analysis for increasing lead times shows, confidence on control forecasts should quickly rise at lead times shorter than a week, but early awareness can be expected at longer lead times. This methodology can be easily replicated using different forecast systems (e.g. the Global Forecast System, GFS) and applied to different regions of the globe after a similar verification as we proposed is performed.

Author contribution

A.M.R., P.M.S., E.D. conceived and designed the experiments, P.M.S and E.D. performed the computations and A.M.R., P.M.S., E.D., R.M.T, analysed the data and wrote the paper.

Acknowledgments

This work was supported by the project Landslide Early Warning soft technology prototype to improve community resilience and adaptation to environmental change (BeSafeSlide) funded by Fundação para a Ciência e a Tecnologia, Portugal (FCT, PTDC/GES-AMB/30052/2017). P.M.S. and R.M.T. were supported by the project Improving Drought and Flood Early Warning, Forecasting and Mitigation using realtime hydroclimatic indicators (IMDROFLOOD) funded by FCT (WaterJPI/0004/2014). A.M.R. was also supported by the Scientific Employment Stimulus 2017 from FCT (CEECIND/00027/2017). The authors would like to thank David Lavers, from the ECMWF, for the helpful comments on this manuscript.

References

Blamey, R. C., Ramos, A. M., Trigo, R. M., Tomé, R., and Reason, C. J.: The Influence of Atmospheric Rivers over the South Atlantic on Winter Rainfall in South Africa, J. Hydrometeor., 19, 127–142, doi:10.1175/JHM-D-17-0111.1, 2018

Dettinger, M.: Atmospheric Rivers as Drought Busters on the U.S. West Coast, J. Hydrometeor., 14, 1721–1732, doi:10.1175/JHM-D-13-02.1, 2013

Doyle, J. D., Amerault, C., Reynolds, C. A. and Reinecke, P. A.: Initial Condition Sensitivity and Predictability of a Severe Extratropical Cyclone Using a Moist Adjoint. Mon. Wea. Rev., 142, 320–342, doi:10.1175/MWR-D-13-00201.1, 2014

Eiras-Barca, J., Brands, S., and Miguez-Macho, G.: Seasonal variations in North Atlantic atmospheric river activity and associations with anomalous precipitation over the Iberian Atlantic Margin, J. Geophys. Res. Atmos., 121, 931– 948, doi:10.1002/2015JD023379, 2016

Errico, R. M.: What Is an Adjoint Model?, Bull. Amer. Meteor. Soc., 78, 2577–2592, doi:10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2, 1997

Gershunov, A., Shulgina, T., Ralph, F. M., Lavers, D. A., and Rutz, J. J.: Assessing the climate-scale variability of atmospheric rivers affecting western North America, Geophys. Res. Lett., 44, 7900–7908, doi:10.1002/2017gl074175, 2017.

Gimeno, L., Dominguez, F., Nieto, R., Trigo, R. M., Drumond, A., Reason, C., and Marengo, J.: Major Mechanisms of Atmospheric Moisture Transport and Their Role in Extreme Precipitation Events, Annu. Rev. Env. Resour., 41, 117–141, doi:10.1146/annurev-environ-110615-085558, 2016.

Guan, B. and Waliser, D. E.: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies, J. Geophys. Res.-Atmos., 120, 12514–12535, doi:10.1002/2015jd024257, 2015.

Guan, B., Waliser D, Ralph, F. 2017: An inter-comparison between reanalysis and dropsonde observations of the total water vapor transport in individual atmospheric rivers. J. Hydrol., 19, 321-337, doi:10.1175/JHM-D-17-0114.1, 2017.

Hamill, T. M., Bates G. T, Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y. and Lapenta, W.: NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset, Bull. Amer. Meteor. Soc., 94, 1553–1565, doi:10.1175/BAMS-D-12-00014.1, 2013

Jia, X., Yang, S., Song, W., He, B.: Prediction of wintertime Northern Hemisphere blocking by the NCEP Climate Forecast System, Acta Meteorol. Sin, 28: 76, doi:10.1007/s13351-014-3085-8, 2014

Kingston, D. G., Lavers, D. A., and Hannah, D. M.: Floods in the Southern Alps of New Zealand: the importance of atmospheric rivers, Hydrol. Process., 30: 5063– 5070. doi: 10.1002/hyp.10982, 2016

Lavers, D. A., and Villarini, G.: The nexus between atmospheric rivers and extreme precipitation across Europe, Geophys. Res. Lett., 40, 3259– 3264, doi:10.1002/grl.50636, 2013

Lavers, D. A. and Villarini, G.: The contribution of atmospheric rivers to precipitation in Europe and the United States, J. Hydrol., 522, 382–390, doi:10.1016/j.jhydrol.2014.12.010, 2015.

Lavers, D. A., Pappenberger, F. and Zsoter, E.: Extending medium range predictability of extreme hydrological events in Europe, Nat. Commun., 5, 5382, doi:10.1038/ncomms6382, 2014

Lavers, D. A., Zsoter, E., Richardson, D. S. and Pappenberger, F.: An Assessment of the ECMWF Extreme Forecast Index for Water Vapor Transport during Boreal Winter, Wea. Forecasting, 32, 1667–1674, doi.:10.1175/WAF-D-17-0073.1, 2017

Lavers, D. A, Richardson, D. S., Ramos, A. M, Zsoter, E., Pappenberger, F., Trigo, R. M.: Earlier awareness of extreme winter precipitation across the western Iberian Peninsula, Meteorol Appl., 25: 622– 628, doi:10.1002/met.1727, 2018

Lavers, D.A., Rodwell, M.J, Richardson, D.S., Ralph, F.M., Doyle, J.D., Reynolds, C.A, Tallapragada, V., Pappenberger, F.: The Gauging and Modeling of Rivers in the Sky, Geophys. Res. Lett., 45, doi:10.1029/2018GL079019, 2018

Mahoney, K., Jackson, D. L., Neiman, P., Hughes, M., Darby, L., Wick, G., White, A., Sukovich, E. and Cifelli, R.: Understanding the Role of Atmospheric Rivers in Heavy Precipitation in the Southeast United States, Mon. Wea. Rev., 144, 1617–1632, doi:10.1175/MWR-D-15-0279.1, 2016

Martin, A., Ralph, F. M., Demirdjian, R., DeHaan, L., Weihs, R., Helly, J., Reynolds, D., Iacobellis, S.: Evaluation of Atmospheric River Predictions by the WRF Model Using Aircraft and Regional Mesonet Observations of Orographic Precipitation and Its Forcing, J. Hydrometeor., 19, 1097–1113, doi:10.1175/JHM-D-17-0098.1, 2018

Miller, D. K., Miniat, C.F., Wooten, R.M., Barros, A. P.: An Expanded Investigation of Atmospheric Rivers in the Southern Appalachian Mountains and Their Connection to Landslides, Atmosphere, 10, 71, 2019

Nayak, M. A., Villarini, G., and Lavers, D. A.: On the skill of numerical weather prediction models to forecast atmospheric rivers over the central United States, Geophys. Res. Lett., 41, 4354– 4362, doi:10.1002/2014GL060299, 2016

Ralph, F. M., and Coauthors: Dropsonde observations of total water vapor transport within North Pacific atmospheric rivers, J. Hydrometeor., 18, 2577–2596, doi:10.1175/JHM-D-17-0036.1, 2017

Ralph, F. M., Dettinger, M. D., Cairns, M. M., Galarneau, T. J. and Eylander, J.: Defining “atmospheric river”: How the Glossary of Meteorology helped resolve a debate, Bull. Amer. Meteor. Soc., 99, 837–839, doi:10.1175/BAMS-D-17-0157.1, 2018

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M., Anderson, M. D. Reynolds, M., Schick, and C. Smallcomb, L. J.: A Scale to Characterize the Strength and Impacts of Atmospheric Rivers, Bull. Amer. Meteor. Soc., 100, 269–289, doi:10.1175/BAMS-D-18-0023.1, 2019

Ramos, A. M., Trigo, R. M., Liberato, M. L. R., and Tome, R.: Daily precipitation extreme events in the Iberian Peninsula and its association with Atmospheric Rivers, J. Hydrometeorol., 16, 579–597, doi:10.1175/JHM-D-14-0103.1, 2015

Ramos, A. M., Nieto, R., Tomé, R., Gimeno, L., Trigo, R. M., Liberato, M. L. R., and Lavers, D. A.: Atmospheric rivers moisture sources from a Lagrangian perspective, Earth Syst. Dynam., 7, 371–384, doi:10.5194/esd-7-371-2016, 2016.

Ramos, A. M., Martins, M. J., Tomé, R. and Trigo, R. M.: Extreme Precipitation Events in Summer in the Iberian Peninsula and Its Relationship with Atmospheric Rivers, Front. Earth Sci., 6:110. doi:10.3389/feart.2018.00110, 2018

Ramos, A. M., Blamey, R. C., Algarra, I., Nieto, R., Gimeno, L., Tomé, R., Reason, C. J. and Trigo, R. M., From Amazonia to southern Africa: atmospheric moisture transport through low-level jets and atmospheric rivers, Ann. N.Y. Acad. Sci., 1436: 217-230. doi:10.1111/nyas.13960, 2019

Reynolds, C. A., Doyle, J. D., Ralph, F. M. and Demirdjian, R.: Adjoint Sensitivity of North Pacific Atmospheric River Forecasts, Mon. Wea. Rev., 147, 1871–1897, doi:10.1175/MWR-D-18-0347.1, 2019

Rutz, J. J., Steenburgh, W. J., and Ralph, F. M.: Climatological characteristics of atmospheric rivers and their inland penetration over the Western United States, Mon. Weather Rev., 142, 905–921, doi:10.1175/mwr-d-13-00168.1, 2019

Sodemann, H. and Stohl, A.: Moisture Origin and Meridional Transport in Atmospheric Rivers and Their Association with Multiple Cyclones. Mon. Wea. Rev., 141, 2850–2868, doi:10.1175/MWR-D-12-00256.1, 2013

Valenzuela, R. A. and Garreaud, R. D., Extreme Daily Rainfall in Central-Southern Chile and Its Relationship with Low-Level Horizontal Water Vapor Fluxes, J. Hydrometeor., 20, 1829–1850, doi:10.1175/JHM-D-19-0036.1, 2019

Viale, M., Valenzuela, R., Garreaud, R. D., and Ralph, F. M.: Impacts of Atmospheric Rivers on Precipitation in Southern South America, J. Hydrometeor., 19, 1671–1687, doi:10.1175/JHM-D-18-0006.1, 2018

Wilks, D. S. (2006) Statistical Methods in the Atmospheric Sciences. 2nd Edition, Academic Press, London.

Zsoter, E., Pappenberger, F. and Richardson, D.: Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index, Meteo. App., 22, 236– 247. doi:10.1002/met.1447, 2014

Figure captions

Figure 1. Receiver Operating Characteristic curves (ROC curves) for the IVT and precipitation ensemble forecasts during Atmospheric River days (ARs) from the ECMWF model, using Portuguese surface meteorological stations during the 2012-2016 extended winters (October-March) as a benchmark, and considering events above the 95th percentile (a). The solid lines are for the IVT and dashed lines for precipitation. Different curve colors represent different lead times for the forecasts (1, 5, 9 and 13 days). Area under the ROC curves for lead times up to 14 days (b), where the confidence intervals are also shown. The mean percentage of ensemble members forecasting IVT (pink) and precipitation (purple) above the 95th percentile for lead times up to 14 days during extreme rainfall events associated to ARs (observed precipitation above the 95th percentile associated to an AR over Western Iberia) is shown in (b).

Figure 2. The six regional boxes considered for the verification of IVT probabilistic forecasts in Western Iberia at lead times up to 14 days: i) sea North; ii) Galicia; iii) North Portugal; iv) Central Portugal; v) South Portugal; vi) sea South.

Figure 3. Example of the evolution of the Operational Forecast of the IVT in an event affecting Western Iberia. a) Analysis of the IVT fields on January 4 2016 at 12UTC. b) control forecasts for that date issued with different lead times, from 1 to 14 days.

Figure 4. Example of the evolution with lead time for the accuracy of IVT probabilistic forecasts, for the event presented in Fig. 3. In a) the black line represents the error in the location of the maximum IVT (i.e. landfall distance) in the control run (in km), while the blue thick solid line represents the landfall distance for the Ensemble Forecasts. The blue shaded envelope accommodates the Ensemble spread, considering the 25th and 75th percentiles. In addition, the black arrows represent the errors in the angle (in degrees) of the AR axis for each forecast. Panel b) shows the error in the IVT intensity (Kg/m/s) for each forecast at the observed landfall location. Black solid line, red solid line and red shaded envelope are as in panel (a). Panel c) shows the error in the maximum IVT at the specific locations where it has been observed and forecasted for each lead time, regardless of the landfall distance. Black solid line, dashed red line and red shaded envelop as in a) and b). The open circles represented in some lead times represent forecasts where the maximum IVT did not surpass a minimum threshold of 450 Kg/m/s within the target domain (i.e. regional boxes over Western Iberia).

Figure 5. Statistics for the verification of the accuracy of the control forecast of IVT for all events affecting Western Iberia during the extended winters between 2012 and 2016 relative to mean errors (a) and absolute errors (b). Solid blue line represents the error in the location of the maximum IVT between observation and each forecast (in km). The solid red line shows the error in the IVT (Kg/m/s) for each forecast at the real landfall location (where the maximum IVT was observed), while the red dashed curve represents the error in the maximum IVT between the observed and each lead time forecast, independently of the location in each forecast. Black arrows represent the errors in the angle (in degrees) of the AR axis.

Figure 6. Statistics for the verification of the accuracy of the Ensemble Forecast of IVT for all events affecting Western Iberia during the extended winters between 2012 and 2016. a) mean Landfall distance errors (in km) for the mean of the Ensemble Forecast (thick solid colored line) and the spread of the Ensemble (shading). b) As in a), but for the mean IVT error (in Kg/m/s) at the location of observed landfall. c) As in b), but at the location of the maximum IVT in each forecast.

Figure 7. Percentage of Ensemble members forecasting IVT above 450 Kg/m/s in each of the regional boxes and for each lead time for the case study presented in Fig. 3 (January 4 2016). Green bars represent a spatially accurate forecast (in the box where the maximum IVT was observed). Yellow bars represent a forecast in an adjacent box to where it was actually observed. Red bars represent a forecast in one of the remainder boxes. The bars in the last line represent a completely missed forecast, by either: i) no AR forecast; ii) AR forecast outside of the 6 considered boxes in Western Iberia.

Figure 8. Contingency tables for the accuracy of AR-related IVT forecasts by the ECMWF ensemble system, for lead times ranging between 1 and 14 days, during the winters spanning 2012-2016. The red shading represents the percentage of observations versus forecasts. Note that a perfect forecast system would only present shadings in the diagonal, as the y-axis represents observed events in each box (as presented in Fig.2) and the x-axis represents forecasts in each box. The number of events in each box is shown in the y-axis by the blue arrow. The last row/column represent either: i) observations/forecasts outside of the 6 considered boxes; ii) no AR observed/predicted.

Figure 9. Forecast verification metrics for IVT exceedances (>450 Kg/m/s) using the ECMWF Ensemble forecast system during the 2012-2016 extended winters in Western Iberia, and for lead times between 1 and 14 days. Colored bars represent metrics for individual regional boxes, as where the darkest blue bar represents the most northerly box and the yellow bars the most southerly box (as depicted in Fig. 2).