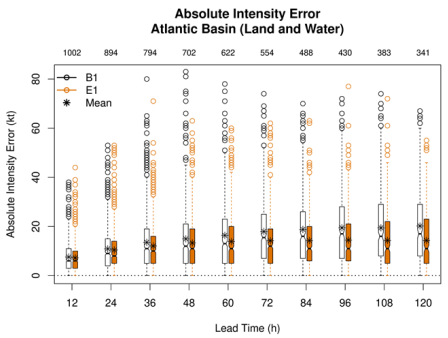
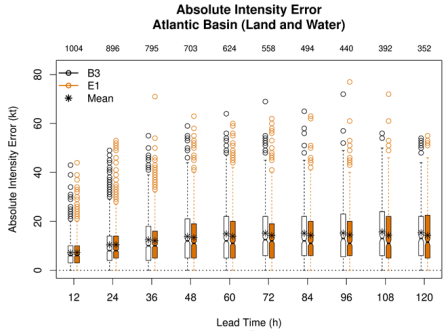
Interpretation of error distributions





- How would you summarize – in words – the differences in performance for each pair of models?
  - B1 vs. E1
    - E1 appears to have better overall performance (smaller errors) than B1, especially for longer lead times, as represented by the mean and median error values and the 0.75th quantiles.
    - The B1 errors are more variable than the E1 errors, as represented by the ***sizes of the boxes*** and the ***lengths of the "whiskers"*** above the boxes.
  - B3 vs. E1
    - Differences in error distributions between models are not as notable as for B1 vs. E1.
      - Boxes have similar sizes (central variability) for most lead times, though B3 boxes are somewhat larger than E1 boxes for lead times between 36 and 108 h.
- What can you say about average performance of each of the models as a function of lead time?
  - For B1 vs. E1, the mean error value (represented by the *s) is larger for B1 than for E1.
  - The differences appear to be smaller for the B3 vs. E1 comparison than for the B1 vs. E1 comparison.
  - Also, the median values (the horizontal lines inside the "box" areas) for B1 are generally larger than the median errors for E1, while the median values seem quite similar for E1 and B3.
  - ***However***: When uncertainty is taken into account [via the "bends" in the boxes, which represent 95% confidence intervals (CIs) on the median values], only a few of the differences between the median errors for B1 and E1 appear to be significantly different..
    - For example: The B1 and E1 medians for lead times 60-120 h can be considered significantly different from each other because the CIs for the two models do not overlap.
- What differences can you describe regarding the characteristics of the outliers for each pair of models?
  - For lead times 48-72 h and 120 h, B1 appears to have more outliers than E1.
  - The frequencies of outliers for B3 and E1 appear to be similar. At lead times of 36, 96, and 108 h, E1 had at least one outlier that was larger than the outliers for B3.
  - Thus, *it appears that B1 tends to have more extreme error values than E1, while the frequencies of extreme values are similar for B3 and E1*.

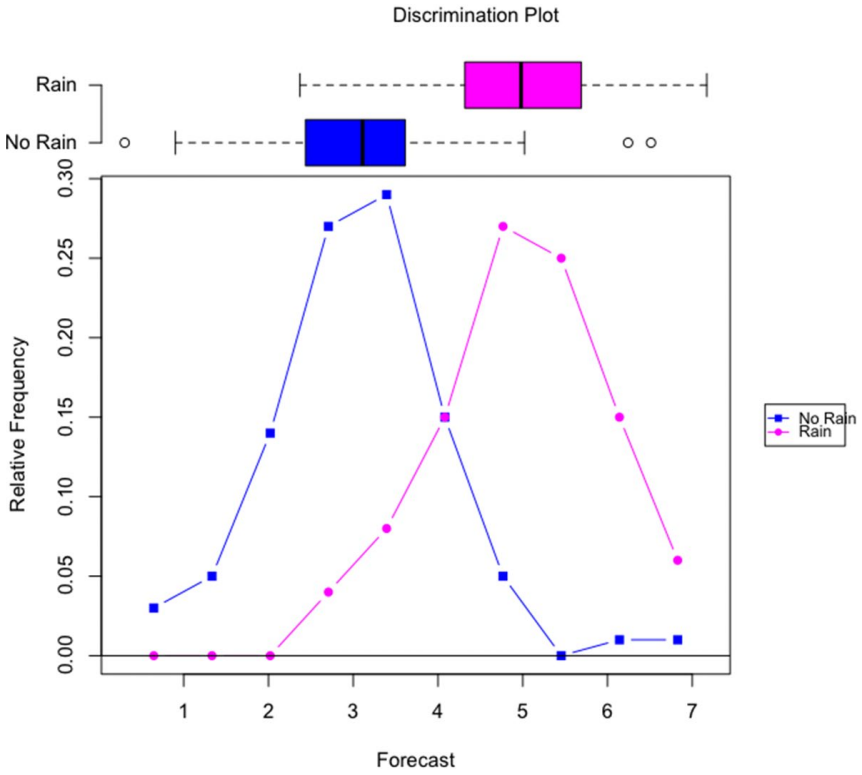
Examination of a 2x2 contingency table (aka "confusion matrix")

	Observed		
Forecast	**Tornado**	**No tornado**	**Total**
Tornado	28	72	100
No Tornado	23	2680	2703
Total	51	2752	2803

Summary table for a set of 2803 forecasts of the occurrence of tornados. The contingency table shows the frequencies associated with each possible forecast and outcome: (a) tornado was forecast and tornado occurred; (b) tornado was forecast and no tornado occurred; (c) no tornado was forecast and a tornado occurred; and (d) no tornado was forecast and no tornado occurred.

- What can you deduce about the quality/performance of the tornado predictions (just looking at the numbers in the table)?
  - An important thing to notice is that the occurrence of tornados is quite rare (51/2803 = 0.02)
  - Tornados are predicted about twice as frequently as they occur (100/2803 = 0.04)
- How would you summarize the performance of these forecasts?
  - Categorical scores would be applied (POD, CSI/ETS, GSS)
  - Percent correct is not a good score to use!

Analyzing “discrimination”



This “discrimination” plot shows the empirical distributions of forecast probabilities of rain for times when measurable rain *did not* occur (blue) vs. the distribution of rain forecast probabilities when measurable rain *did* occur (pink). The box plots at the top also summarize these distributions.

How would you interpret the performance of this set of forecasts to predict rainfall based on this distribution diagram?

- The plotted conditional distributions suggest that there is some separation between the 2 conditional distributions of forecasts when the rain occurs and when it does not occur.
- That is, higher probabilities were applied when rain occurred than when rain did not occur.

What kinds of information can you gain from this type of distribution plot?

- The forecasting system is at least somewhat successful at discriminating between times when rain occurred from times when rain did not occur
- One indicator of this capability is the relatively large separation between the peak values shown on the line plot.
- The box plots at the top also indicate that the central parts of the conditional distributions (between 0.25th to 0.75th quantiles) are totally non-overlapping.

- Would you say that this is a “good” set of forecasts? Why or why not?
  - Without comparison to a standard (e.g., another forecast, climatology, persistence) it is difficult to say how good it is in a relative sense.
  - However, the notable separation of the distributions suggests that the forecasts have some skill.