Application results

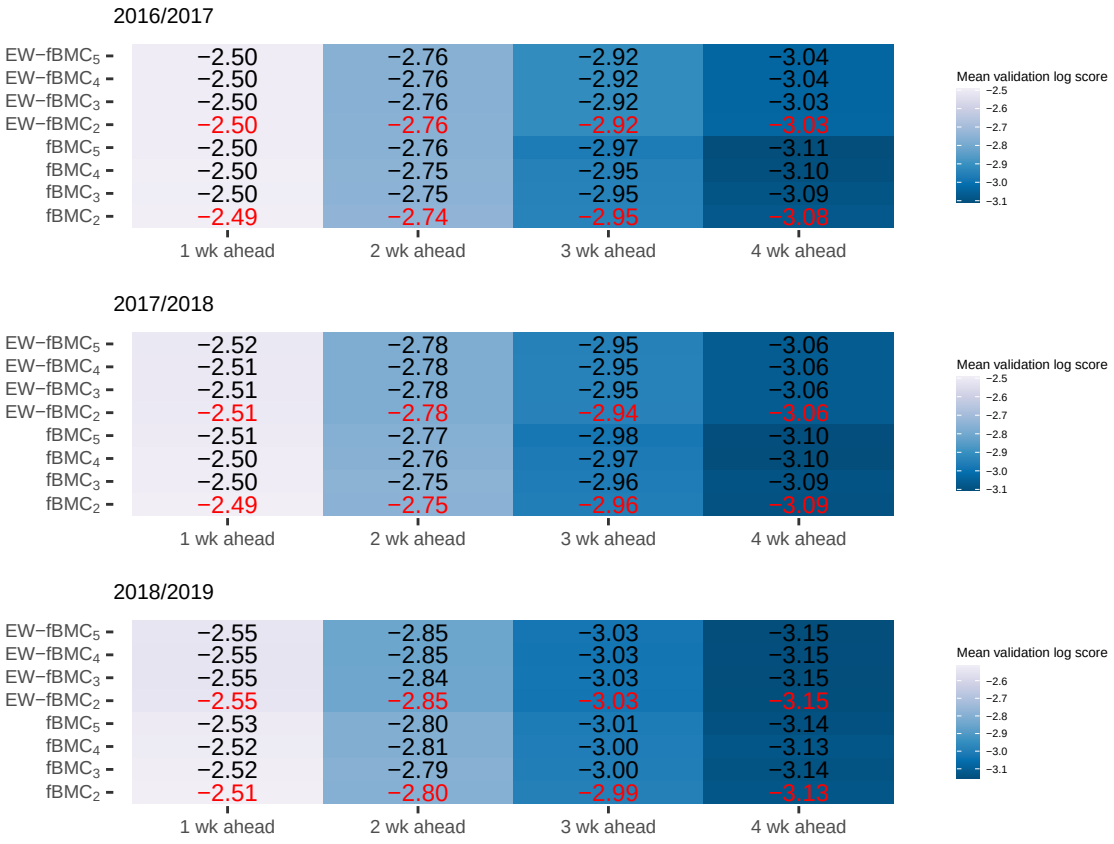
Nutcha Wattanachit

06/29/2021

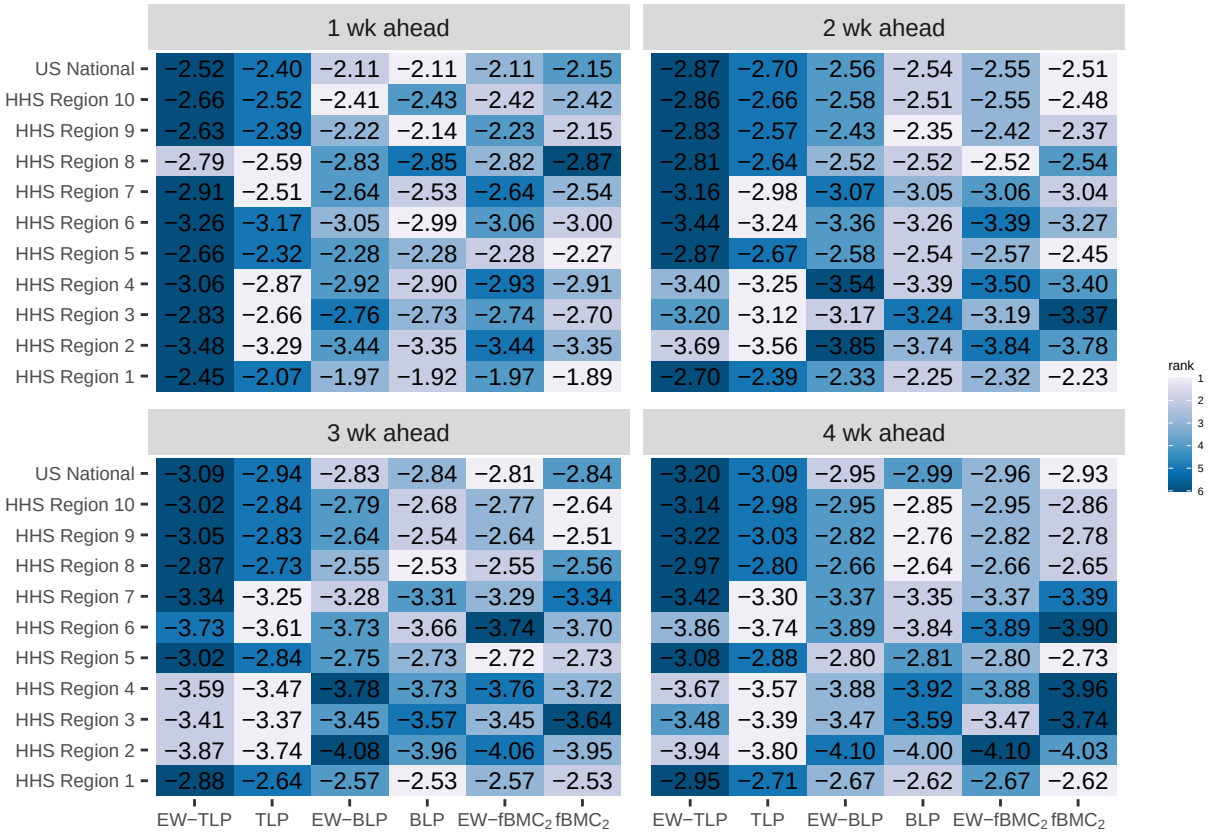
Log scores

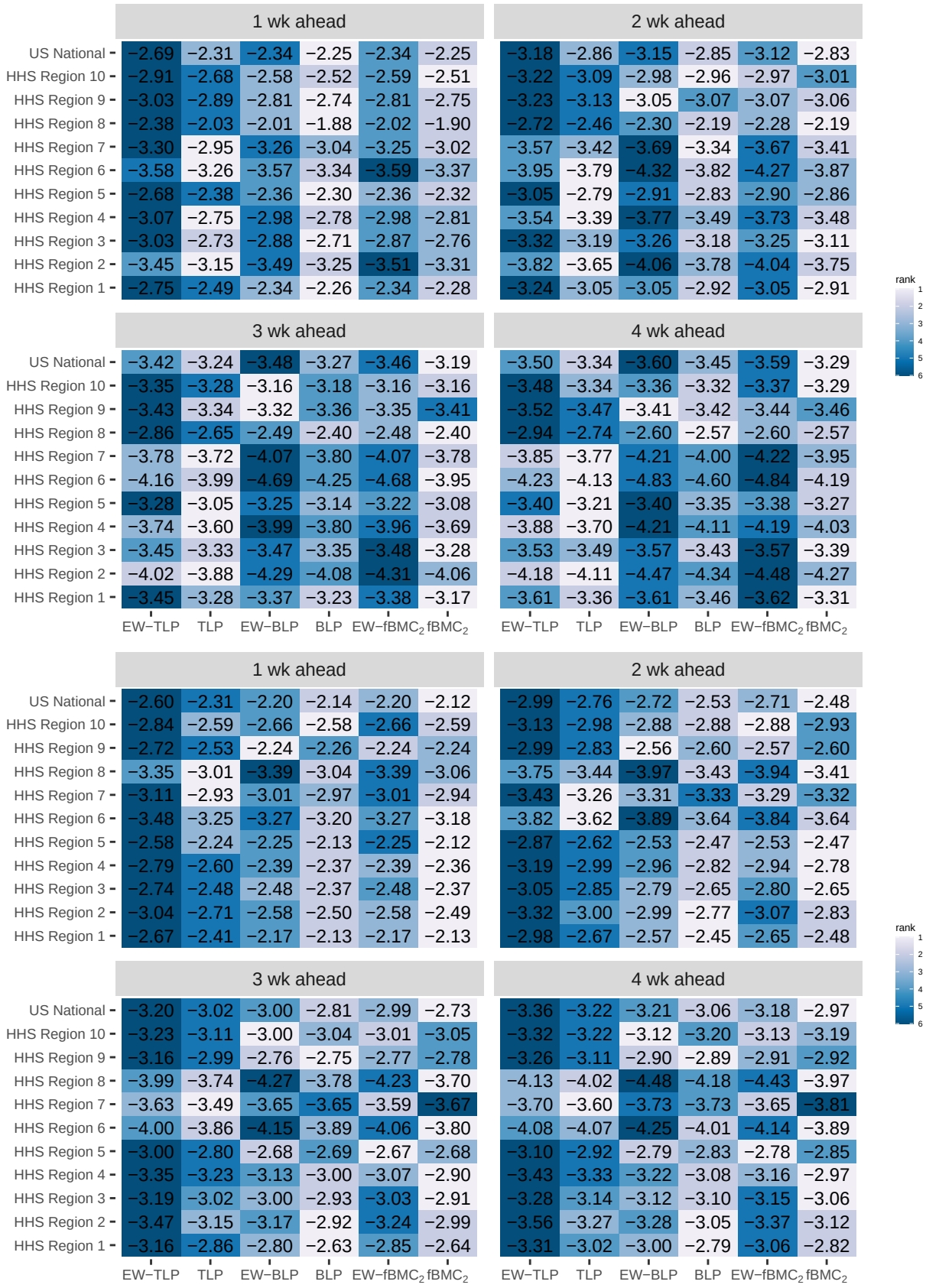
Cross-validated mean log scores for BMC_k and EW-BMC_k

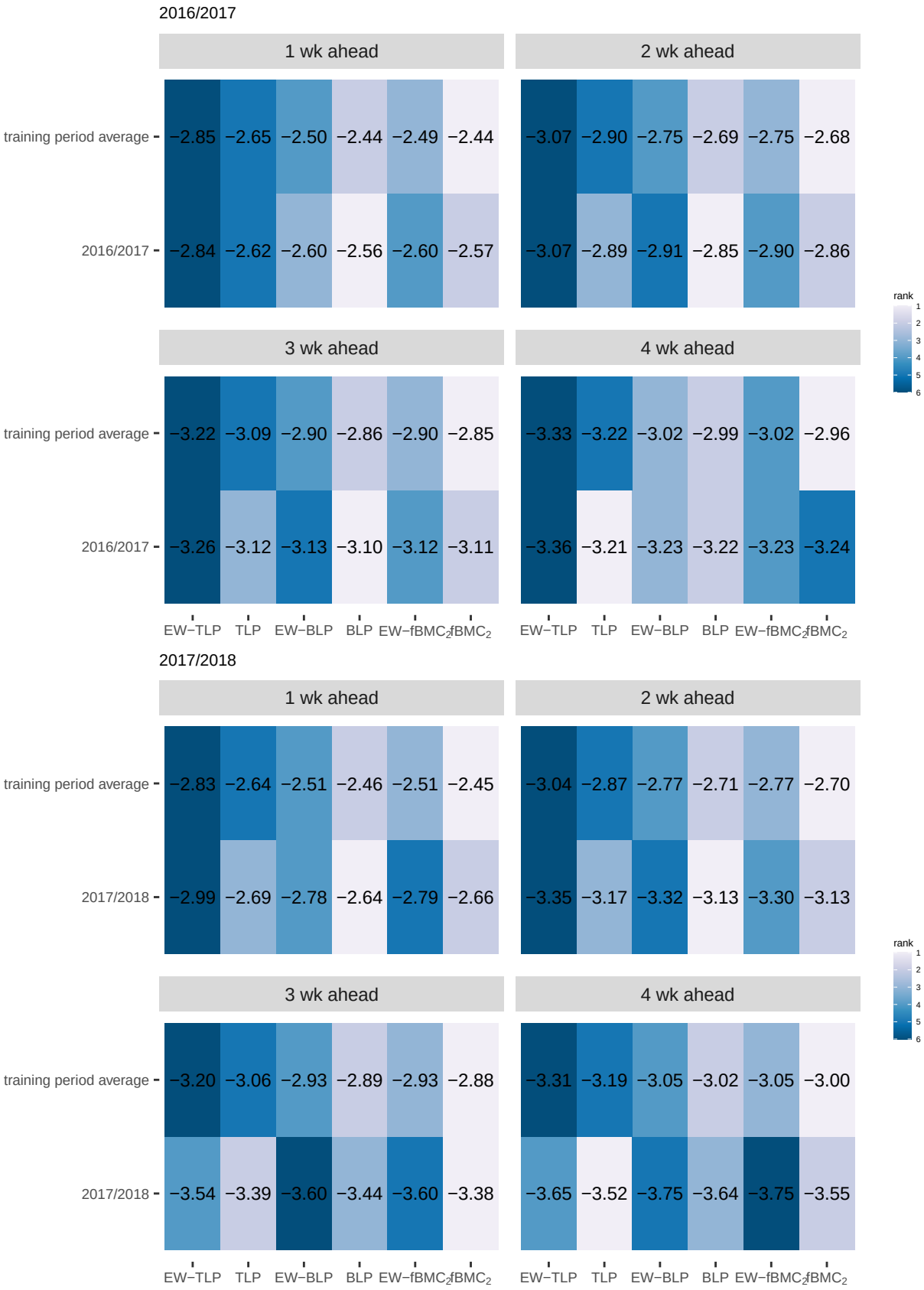
Select method with log score within 1 sd of maximum (negatively oriented). So, the cutoff is max log score plus 1 sd and any methods with mean validation log scores (rounded to the second decimal point) higher or equal to the cutoff (rounded to the second decimal point) will satisfy the criteria. Then, among those that make the cutoff, select one method with the lowest beta components (least complex method).

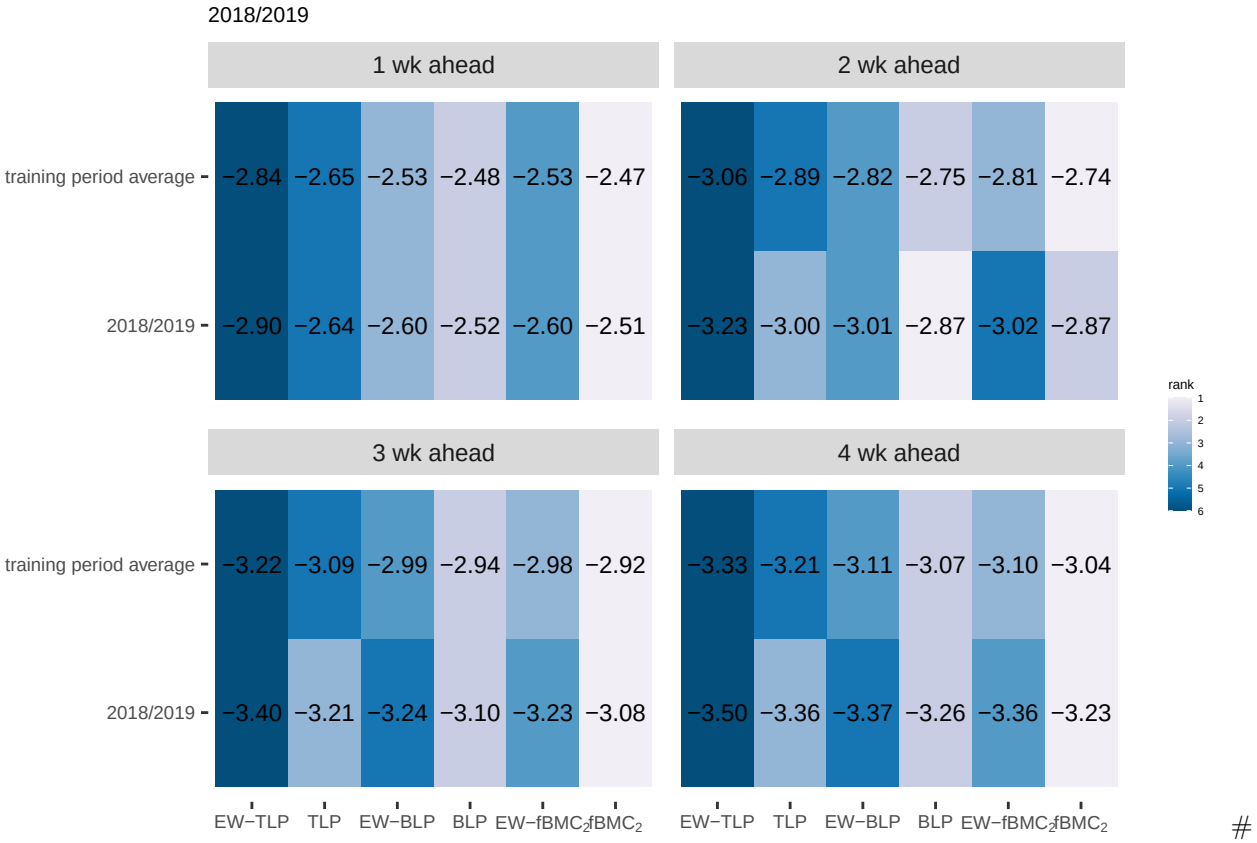


Mean train and test log scores




1 wk ahead
2 wk ahead
3 wk ahead
4 wk ahead
US National
HHS Region 10
HHS Region 9
HHS Region 8
HHS Region 7
HHS Region 6
HHS Region 5
HHS Region 4
HHS Region 3
HHS Region 2
HHS Region 1
EW−TLP
TLP
EW−BLP
BLP
EW−fBMC2
fBMC2
rank
1
2
3
4
5
6

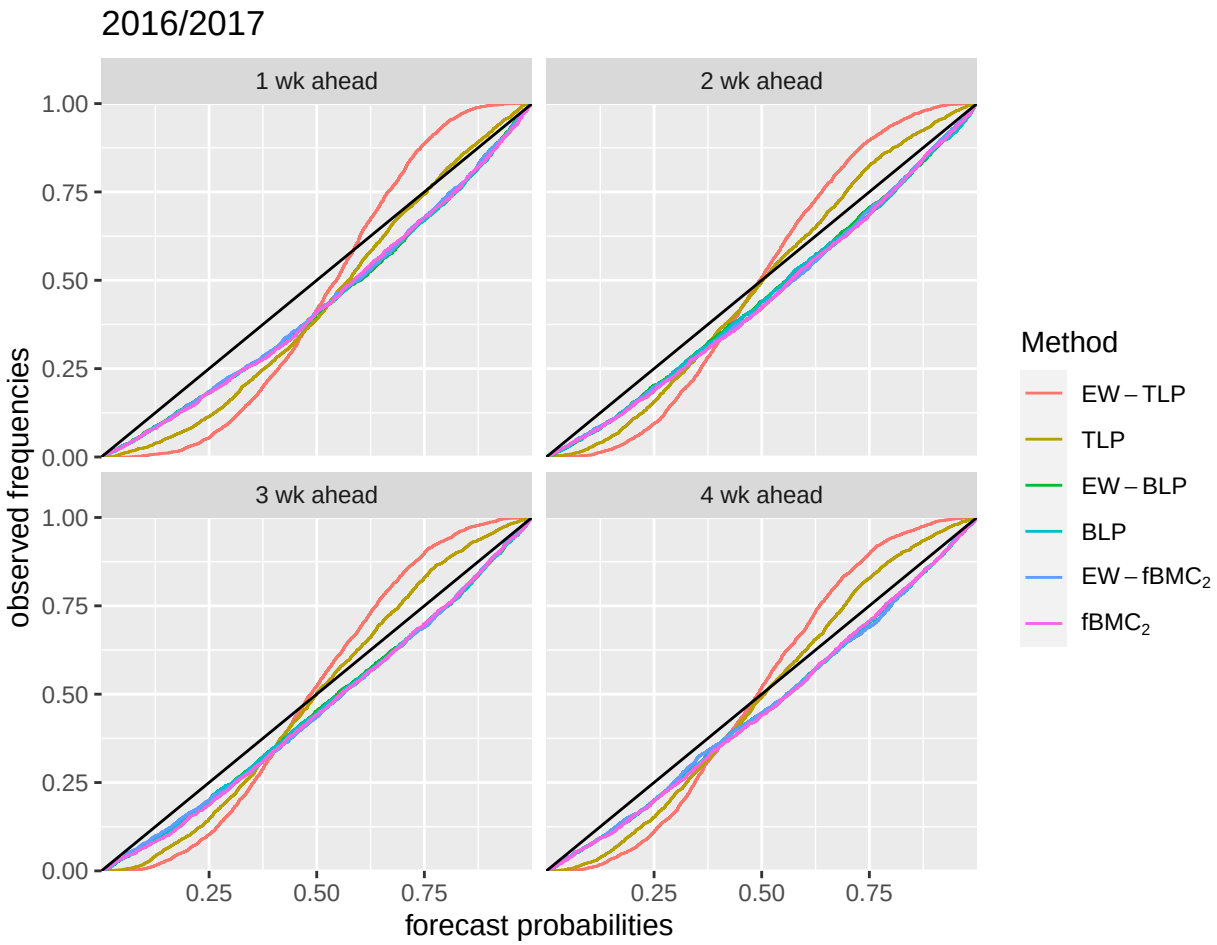

2016/2017
1 wk ahead
2 wk ahead
3 wk ahead
4 wk ahead
training period average
2016/2017
−2.85 −2.65 −2.50 −2.44 −2.49 −2.44
−2.84 −2.62 −2.60 −2.56 −2.60 −2.57
−3.07 −2.90 −2.75 −2.69 −2.75 −2.68
−3.07 −2.89 −2.91 −2.85 −2.90 −2.86
−3.22 −3.09 −2.90 −2.86 −2.90 −2.85
−3.26 −3.12 −3.13 −3.10 −3.12 −3.11
−3.33 −3.22 −3.02 −2.99 −3.02 −2.96
−3.36 −3.21 −3.23 −3.22 −3.23 −3.24
EW−TLP TLP EW−BLP BLP EW−fBMC2 fBMC2
rank
1 2 3 4 5 6
2017/2018
1 wk ahead
2 wk ahead
3 wk ahead
4 wk ahead
training period average
2017/2018
−2.83 −2.64 −2.51 −2.46 −2.51 −2.45
−2.99 −2.69 −2.78 −2.64 −2.79 −2.66
−3.04 −2.87 −2.77 −2.71 −2.77 −2.70
−3.35 −3.17 −3.32 −3.13 −3.30 −3.13
−3.20 −3.06 −2.93 −2.89 −2.93 −2.88
−3.54 −3.39 −3.60 −3.44 −3.60 −3.38
−3.31 −3.19 −3.05 −3.02 −3.05 −3.00
−3.65 −3.52 −3.75 −3.64 −3.75 −3.55
EW−TLP TLP EW−BLP BLP EW−fBMC2 fBMC2
rank
1 2 3 4 5 6

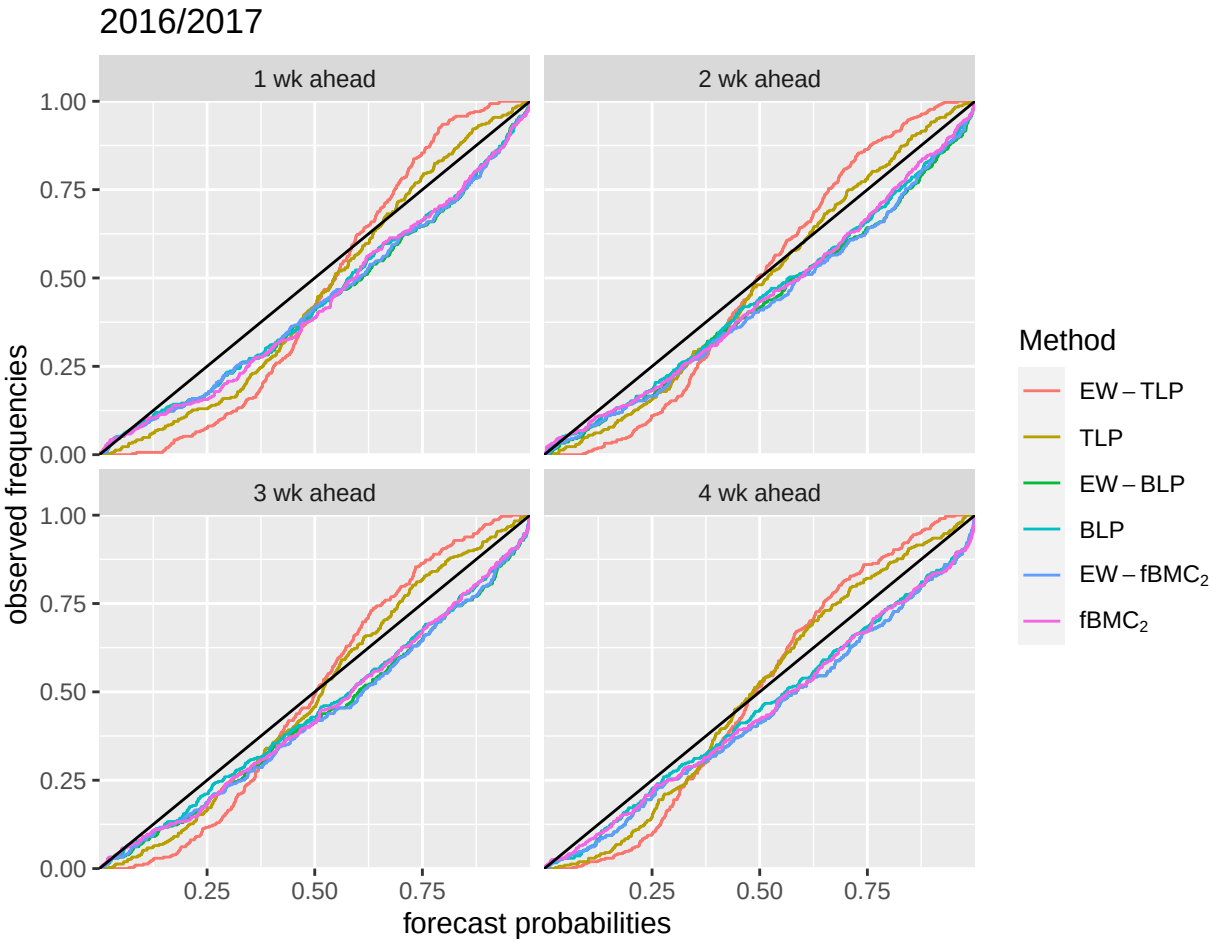


#

Reliability plots

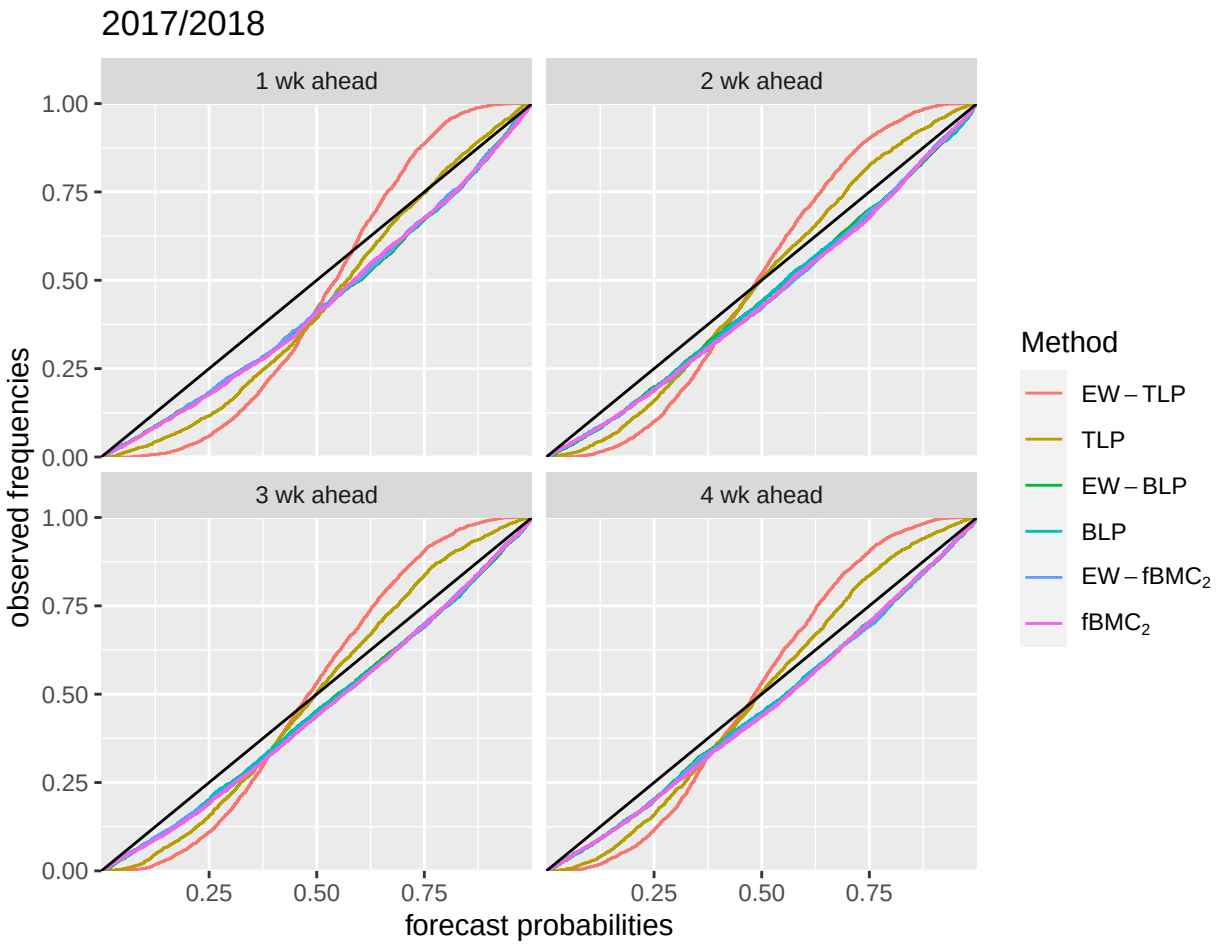
Test season 2016/2017

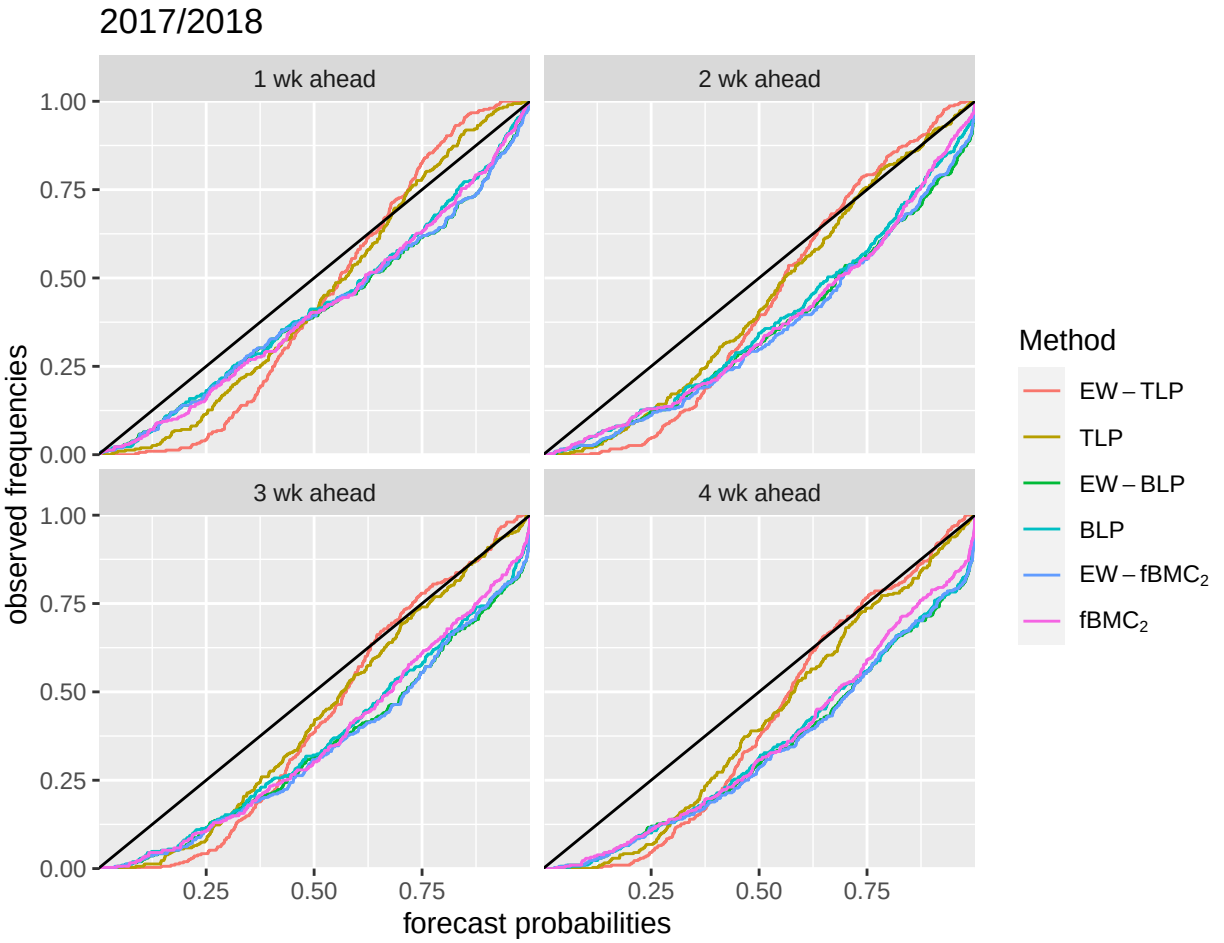




- 1 wk ahead: There is evidence of bias in the PIT histograms in both the training and test seasons. The BLP, which outperforms other methods, has a more uniform PIT histogram in the test season.
- 1 wk ahead: The training PIT histograms for BMC2, EW-BLP, and EW-BMC2 look more uniform compared to that of BLP, maybe overfitting?
- 2 Week Ahead: There is evidence of some bias in the PIT histograms in both the training and test seasons but less than the previous target. The BLP, which outperforms other methods, has a more uniform PIT histogram in the test season.
- 3 Week Ahead: There is evidence of some bias in the PIT histograms in both the training and test seasons. The BLP, which outperforms other methods, has a more uniform PIT histogram in the test season.
- 3 Week Ahead: There might be some overfitting going on, the train PIT histograms look a lot better than the test PIT histograms.
- 4 Week Ahead: There is evidence of a little bias in the PIT histograms in both the training and test seasons.
- 4 Week Ahead: The BLP, EW-BLP, BMC2, and EW-BMC2 are relatively well-calibrated in the training seasons which outperforms other methods, has a more uniform PIT histogram in the test season.
- 4 Week Ahead: TLP outperforms other methods in terms of mean test log score, but the beta methods seem to have more uniform PIT histograms.

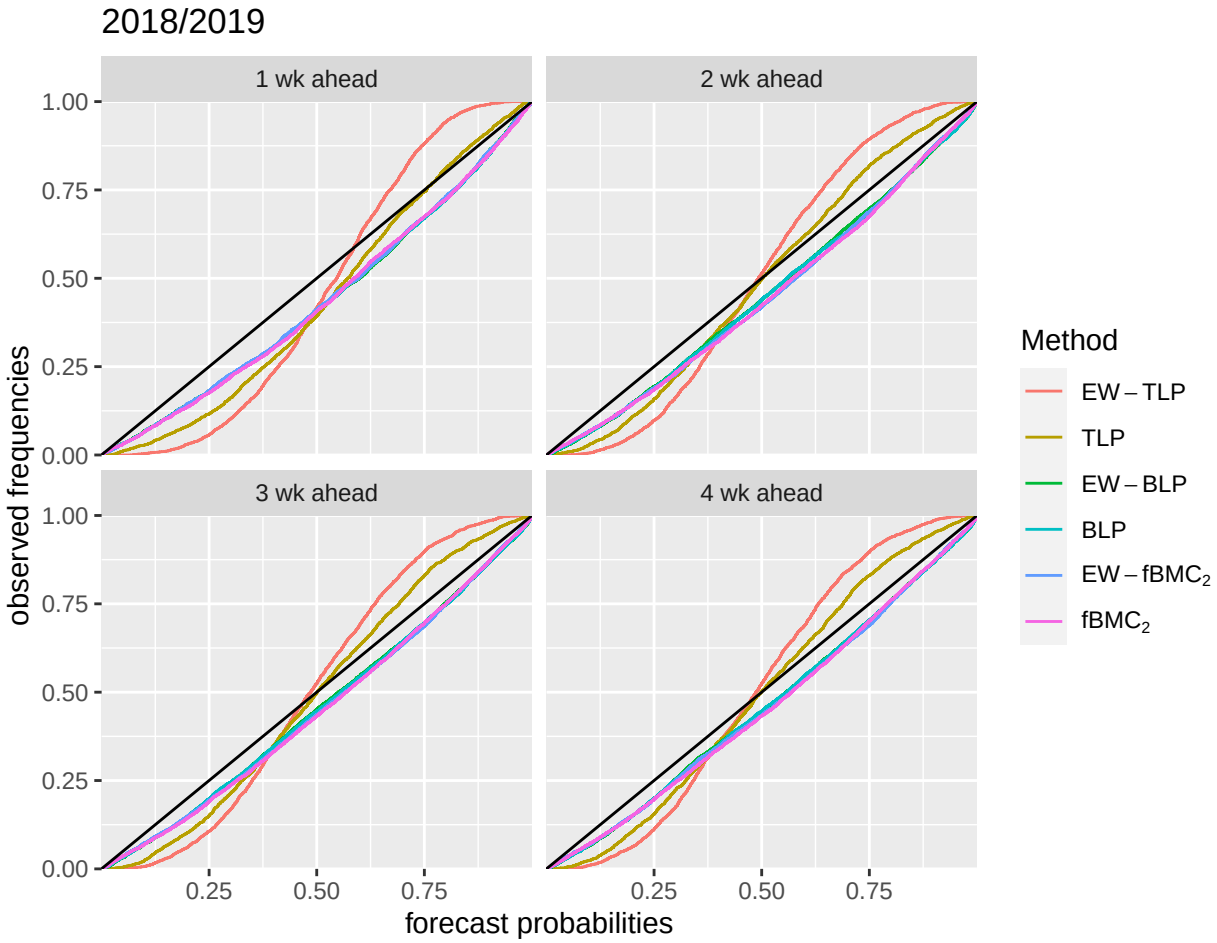
Test season 2017/2018

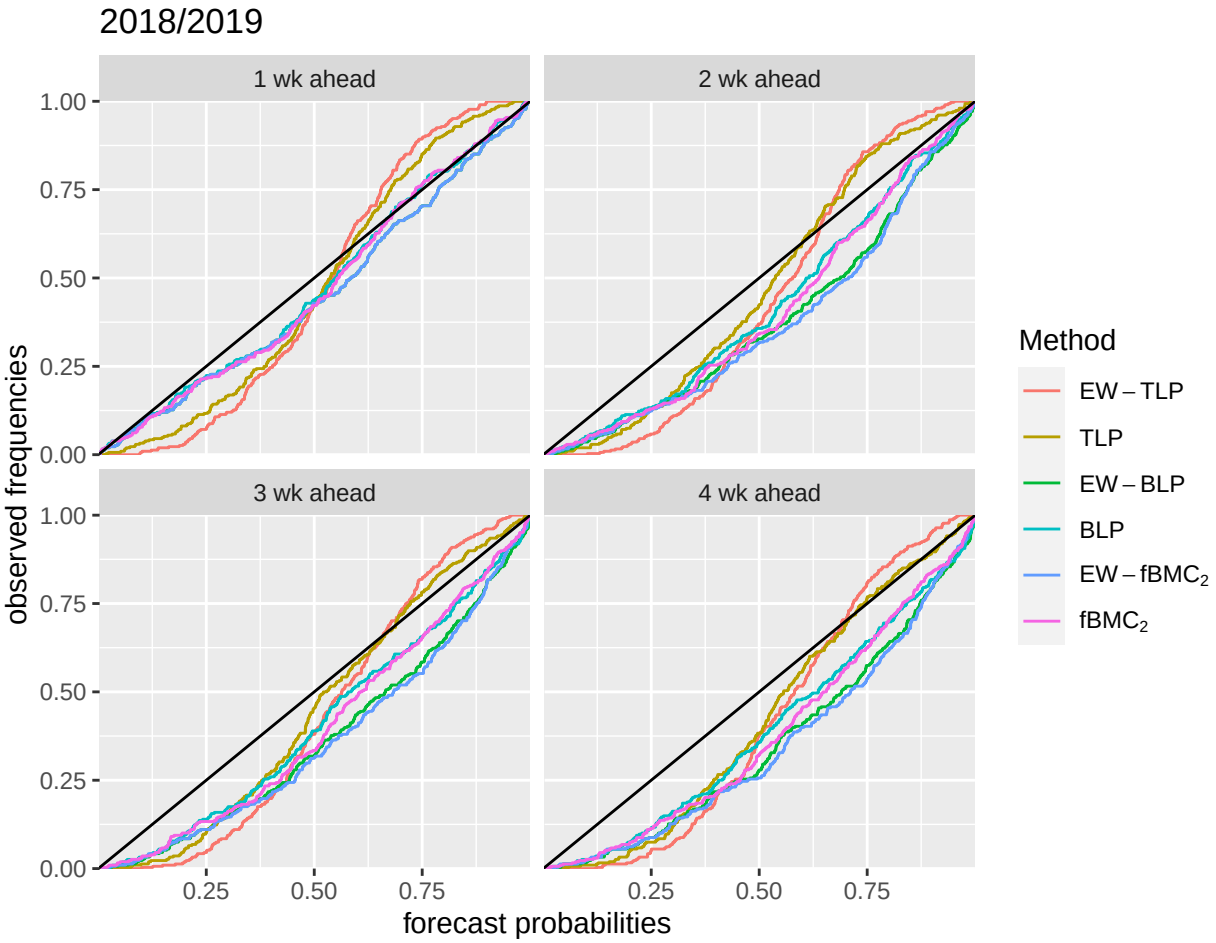




- 1 Week Ahead There is evidence of some bias in the PIT histograms in both the training and test seasons. The BLP, which outperforms other methods, has a more uniform PIT histogram in the test season, but it is not very calibrated.
- 2 Week Ahead: There is evidence of some bias in the PIT histograms in both the training and test seasons. Overall the PIT histograms for the beta methods do not look uniform for the test season, despite being relatively well calibrated for the training seasons
- 2 Week Ahead: The BLP, which outperforms other methods, does not seem to be more calibrated than the TLP in the test season.
- 3 Week Ahead: We have a similar situation here as in the previous target of the same year, but the tail calibration is much worse.
- 3 Week Ahead: BMC2 is the best performing method in terms of mean log score, but again it does not seem more calibrated than TLP (or worse even).
- 4 Week Ahead: PIT histograms for the training season look well calibrated, but very uncalibrated for the test seasons.
- 4 Week Ahead: TLP outperforms other methods here in terms of log score, and the PIT histograms agree.
- 4 Week Ahead: For this season, it is possible the poor calibration is a result from training seasons being very different from the test season (bad flu season in 2017/2018), so we have a lot of overfitting. This phenomenon is more apparent for 3-4 week ahead targets.

Test season 2018/2019





- 1 Week Ahead: There is evidence of some bias in the PIT histograms in the training seasons, but look more calibrated for the test season.
- 1 Week Ahead: The BMC2, which outperforms other methods, does not seem to have a more uniform PIT histogram in the test season compared to other beta methods.
- 2 Week Ahead: The PIT histograms in the training and test seasons look similar for the beta methods. The BLP, which outperforms other methods, has a more uniform PIT histogram in the test season.
- 2 Week Ahead: There is evidence of bias.
- 3 Week Ahead: There is evidence of some bias in the PIT histograms in the test season, but look more calibrated for the training seasons (no surprise here).
- 3 Week Ahead: The BMC2, which outperforms other methods, does not seem to have a more uniform PIT histogram in the test season compared to other beta methods.
- 4 Week Ahead: We see a lot bias in the PIT histograms in test seasons, especially for the equally-weighted beta methods, despite the PIT histograms looking well-calibrated for the training seasons.
- 4 Week Ahead: The BMC2, which outperforms other methods, has a more uniform PIT histogram in the test season compared to other methods. However, these don't look well-calibrated overall.