Improvement of weather type forecast at medium range

Jouan Gabriel [1] Cuzol Anne [2] Monbet Valérie [3] Monnier Goulven [4]

[1] Institut de recherche mathématiques de Rennes / Scalian

[2] Université Bretagne Sud

[3] Institut de recherche mathématiques de Rennes

[4] Scalian

17 décembre 2018

Overview

Context

Objective

Improve weather forecasting for horizon greater than 3 days.

Why ? You want to organise a barbecue in 10 days. You need :

- Check the weather forecast (Temperature, Precipitation, Wind)
- Choose the date
- Call your friends
- Clean your barbecue
- Buy food

Context

Check the weather forecast for 10 days

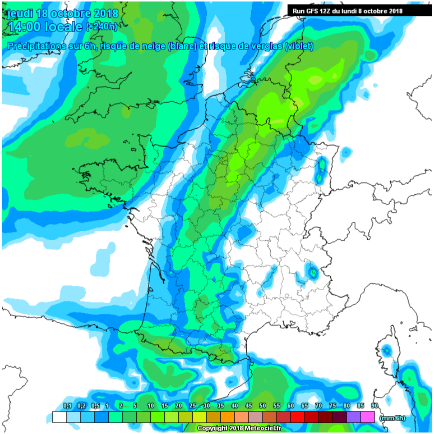


FIGURE – Thursday 18/10/18-14h (forecasted 10 days before)

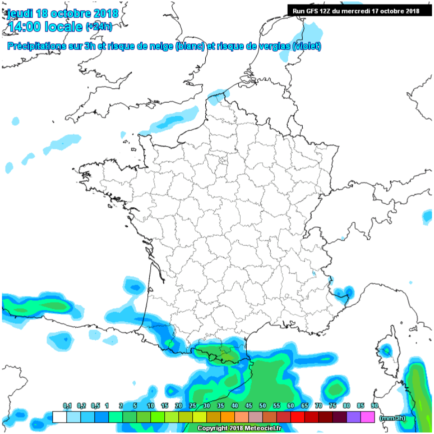


FIGURE – Thursday 18/10/18-14h (observed)

Context

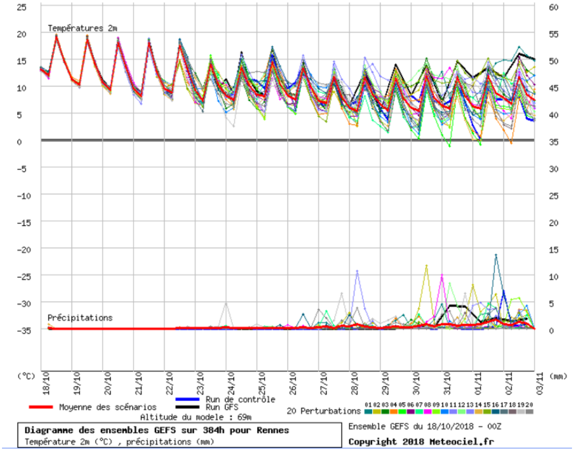


FIGURE – Temperature and precipitation of the ensemble forecasting model at Rennes Thursday 18/10/18-14h

Context

To correct ensemble's model biases use Post processing method.

European Centre for Medium-range Weather Forecast TIGGE archive :

- 11 ensemble models (American, Asian, European ensemble models)
- 11 years (2008 to 2018) of data
- Low spatial Resolution

ECMWF ensemble model :

- 50 forecasting realisations or the perturbed numerical model
- 2 run by day
- 70 variables
- 15 days of forecasting range

Context

Ensemble calibration for weather variable

Y_{obs} is an observation of a weather variable and $(X_1, ..., X_M)$ is an ensemble of M realisations of a random variable X associate to the same weather variable. Calibration provides estimation of $Y_{obs}|X$:

$$P(Y_{obs} \leq y|(X_1, ..., X_m))$$

Ensemble model output statistics [1] (EMOS)

- Multiple linear regression, one physical variable (ground temperature, precipitation, etc...)

Machine learning algorithms :

- Random forest / Gradient boosting [2] [3], one physical variable (ground temperature, precipitation, etc...)
- Non parametric estimation

1. Gneiting et al. 2005
2. Meinshausen et al. 2006a
3. Taillardat et al. 2017

Goal

To choose a date for a barbecue we want : no precipitation, few wind, good temperature, few cloud, etc(...)

Back to our problem, we want to perform a multivariate calibration :

Multivariate conditional distribution

$$P((Y_{obs}^{Pre}, Y_{obs}^{WS}, ...) \leq (y^{Pre}, y^{WS}, ...)|(X_1^{Pre}, ..., X_M^{Pre}), (X_1^{WS}, ..., X_M^{WS}), ...)$$

$\rightarrow$ **Hard Problem**

For the rest of the presentation, we will study the weather variable : Precipitation (Pre) and wind speed (WS).

Class variables

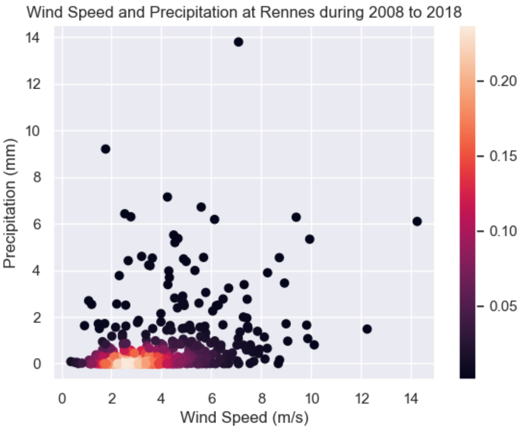


FIGURE – Wind speed and precipitation observations with their density

Solution → Discretise weather variables.

Class variables

Creation of 4 balanced classes with observation of wind speed and precipitation, $Z \in \{0, 1, 2, 3\}$

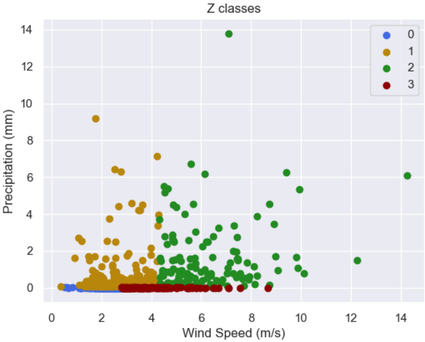


FIGURE – Wind speed and precipitation observations with their classes

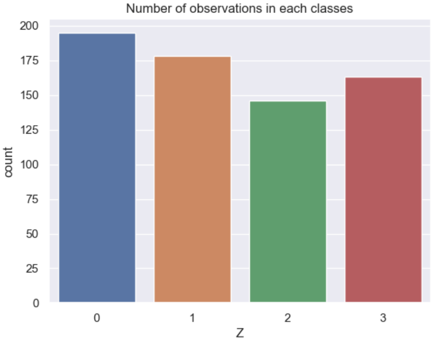


FIGURE – Number of observation in each class Z

Multiclass performance criterion

- Accuracy score in a multiclass problem is a the ratio between the sum of all elements well classified and the total number of elements in the test set.
- In a binary problem $\{0,1\}$, area under the curve (AUC) score is giving by the area under the curve of the relation between the sensitivity $(\frac{\sum \text{True positive}}{n_1})$ and the false positive rate $(\frac{\sum \text{False positive}}{n_0})$.AUC in a multiclass problem is the mean of all AUC of all pair you can form.
- F-measure is the harmonic mean between the precision and the recall of a class.

Raw ensemble mean performance

Evaluation of the capacity of the ensemble model (raw model) to automatically classify Z without postprocessing

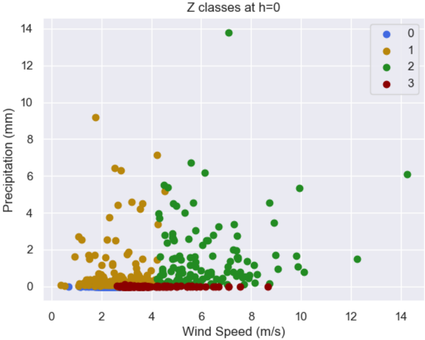


FIGURE – Z at 0 hour of forecasting horizon

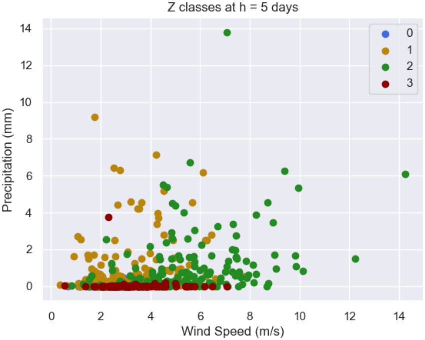


FIGURE – Z at 5 days (120 hours) of forecasting horizon

Raw ensemble mean performance

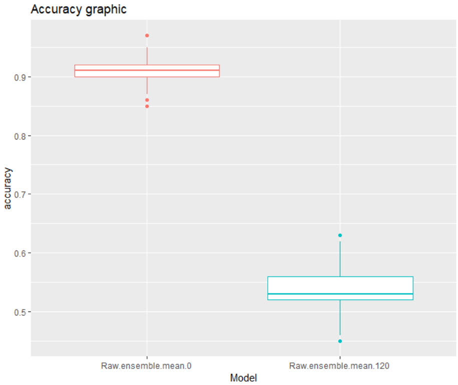


FIGURE – Accuracy score of raw ensemble mean at 0h and 120h of horizon

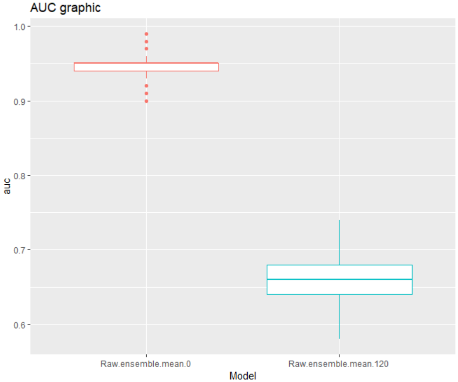


FIGURE – AUC score of raw ensemble mean at 0h and 120h of horizon

Classification

With the Z variable, our problem become a classification problem with ensemble data

Classification problem

$$P(Z = z | (\tilde{X}^{(Pre,h)}, \tilde{X}^{(WS,h)}))$$

Problem : ensembles realisations $\tilde{X} = (X_1, ..., X_M)$ are **excheangeable** $\rightarrow$ $\forall m \in \{1, ..., M\}$ each X_m member carries the same statistical characteristic of X

Classification

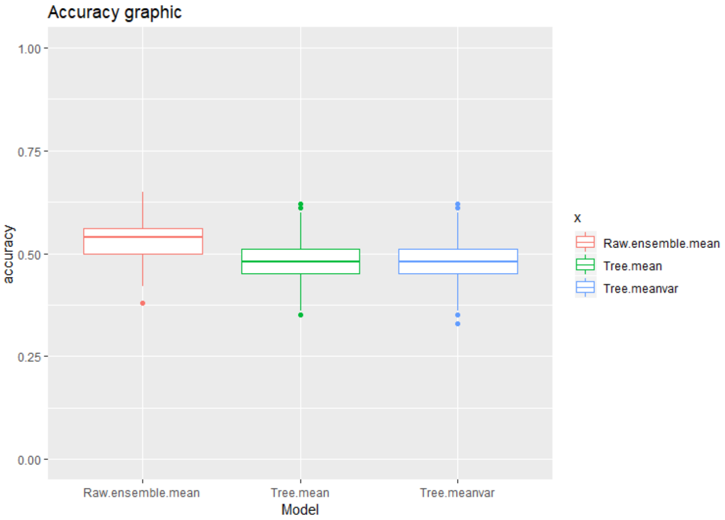
Solution $\rightarrow$ Use ensemble statistics as features

Ensemble statistics

$$\hat{V}_{\hat{\mu}} = \{\hat{\mu}_{X^{(WS,h)}}, \hat{\mu}_{X^{(Prec,h)}}\}$$
$$\hat{V}_{\hat{\mu},\hat{\sigma}^2} = \{\hat{\mu}_{X^{(WS,h)}}, \hat{\sigma}^2_{X^{(WS,h)}}, \hat{\mu}_{X^{(Prec,h)}}, \hat{\sigma}^2_{X^{(Prec,h)}}\}$$

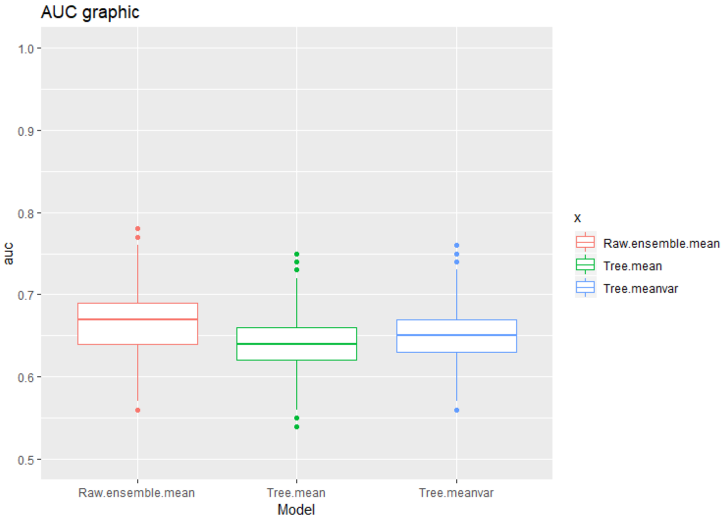
- Training set 407 elements, Test set 275 elements
- AUC, Accuracy, F-measure evaluated by resampling method on test set (1000 random resampling of size 83).
- Algorithm of classification chosen : Tree model

Results



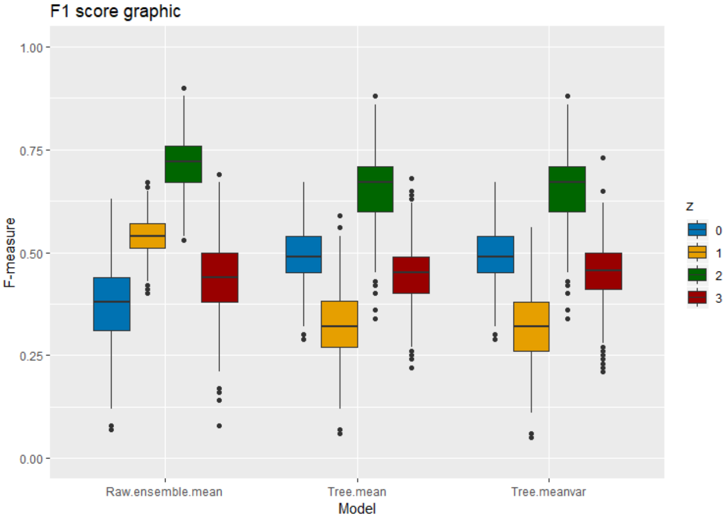
Raw ensemble mean has a better precision of classification than tree classifier with few features.

Results



AUC score reveals that raw ensemble mean classification is not so far from tree algorithm.

Results



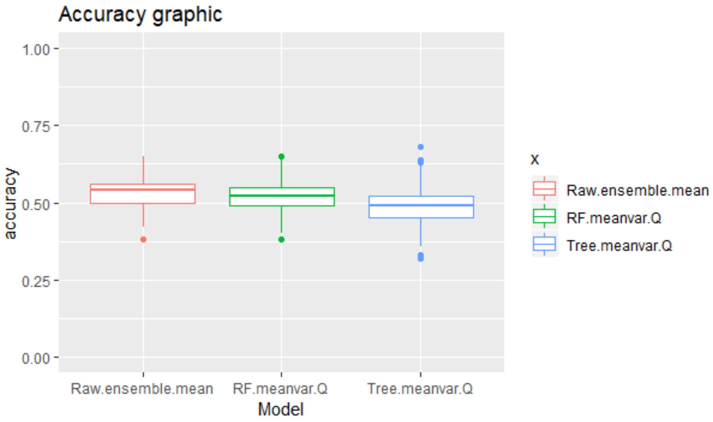
One class overfits the raw ensemble mean classification.

Summary

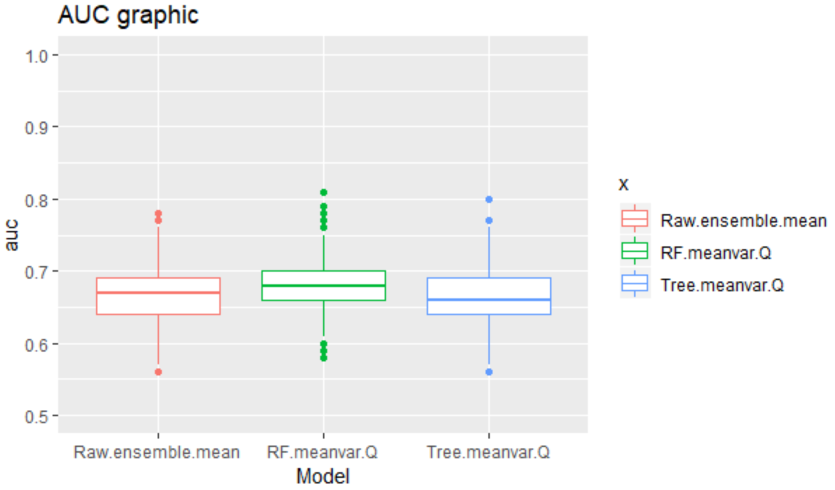
- Raw ensemble mean is a better classifier than tree with mean or mean and variance but the performance of the tree algorithm with 2 and 4 features are not so far than raw ensemble mean.
- No raining weather types (class 0 and 3) are hard to predict by raw ensemble mean.
- No significant difference between tree with mean and the model with mean and variance.

More statistics

Lets study the adding of ensemble statistics ($Q_{0.1}$, $Q_{0.9}$ and median) of wind speed and precipitation with the tree (Tree.meanvar.Q) and random forest (RF.meanvar.Q) classifiers :

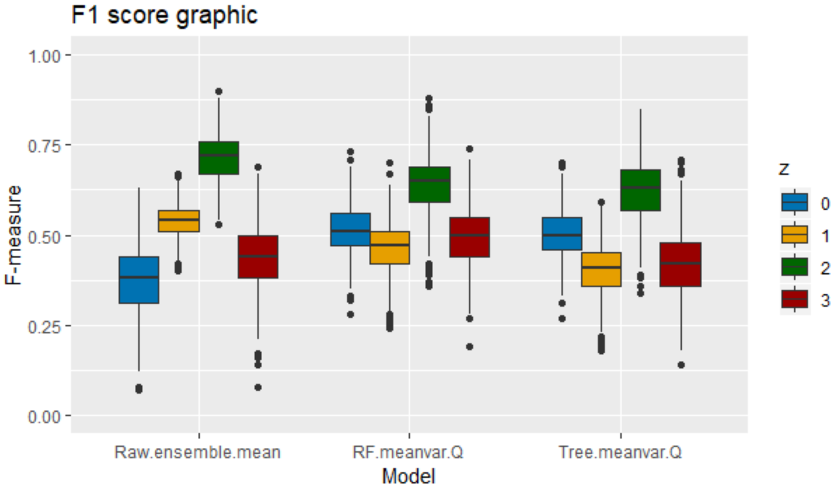


More statistics



Improvement of AUC and Accuracy with the random forest classifier.

More statistics



Improvement of the detection of the class 0 and 3.

More statistics

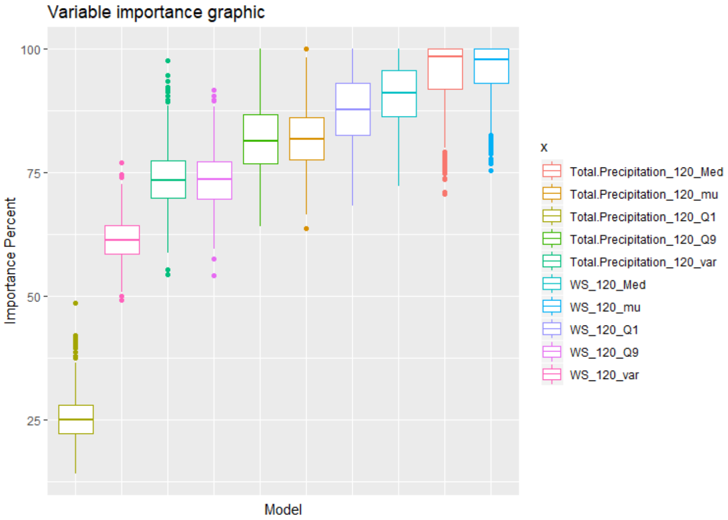


FIGURE – Variable importance of random forest with $\hat{\mu}$, $\hat{\sigma}^2$, $Q_{0.1}$,median and $Q_{0.9}$

Conclusion

- Raw ensemble mean shows good performance of classification for medium range forecasting horizon of weather events.
- First and second statistical moments give not enough information to the classification model to characterise weather type better than the ensemble mean .
- Use of other ensemble statistics from weather variables with random forest classifier show an improvement of the weather type detection.
- Ensemble model classifier like random forest could be a good choice with more features than the tree classifier.

Discussion

- Statistical information from weather random variable at 5 days of forecasting horizon are are underdispersed and biased, mismatch the weather event $\rightarrow$ Integration of statistical information from greater horizon.
- We created a discrete variable Z formed with two variables and try to predict it using ensemble statistics $\rightarrow$ not enough information on intervariable dependencies, study the use of covariance as covariates.
- Statistical information from other weather random variable like temperature, humidity will be studied to help to .
- Other classification algorithm like neural network classifier or gradient boosting tree could be tested with more covariates to help to create a good relation between Z and ensemble statistics.