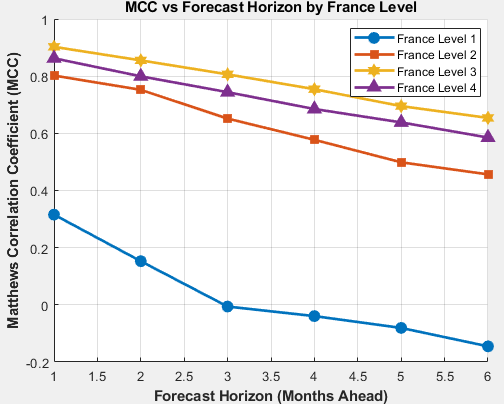


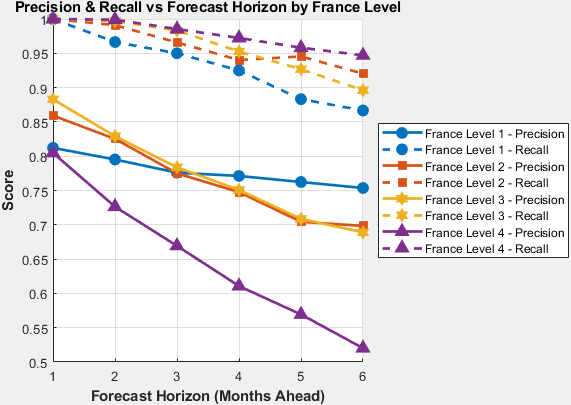
This graph shows **Balanced Accuracy** over time for each France Level.

* **France Level 3 and 4** perform the best, staying above ~0.84 even at month 6.
* **France Level 1 drops fastest**, falling to nearly 0.45 at the 6th horizon — possibly because of class imbalance or poor signal.
* The consistent spacing between lines shows stability in ranking across time
  + The model is more reliable at shorter month horizons and for higher-numbered France Levels.



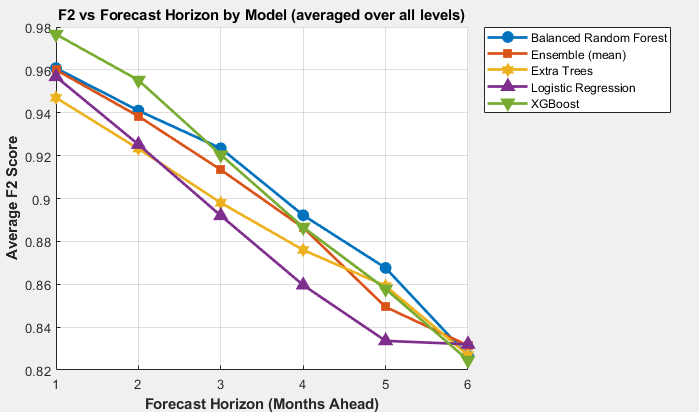
This chart shows the **Matthews Correlation Coefficient (MCC)**, a balanced metric

* MCC drops over time
* **France Level 1 sharply declines**, even going negative (indicating inverse correlation).
* **France Levels 3 and 4** shows the results well — suggesting great predictive performance even under class imbalance.
  + France Levels 3 and 4 show **strong and stable correlation** between predictions and ground truth. A dropping MCC at Level 1 showss predictions are becoming worse than random guessing



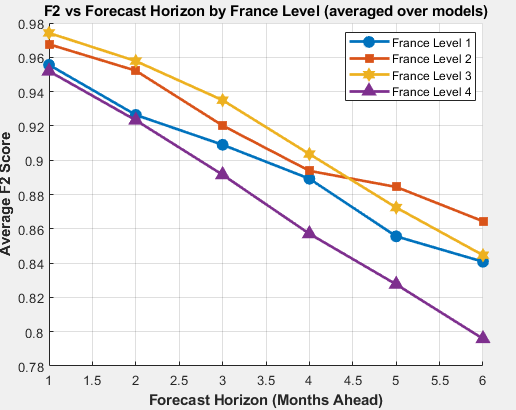
Both **Precision and Recall** plotted together for each France Level over time.

* All curves trend downward over time, especially **Recall**
* Precision shows slightly better than Recall.
* **France Level 4 shows the steepest drop in Recall**, suggesting many missed positives over time
* F2 score is Recall-focused, so this helps explain drops seen in earlier F2 plots
  + Recall is the key factor that makes the F2 score down. The model is **missing more true positives** as the forecast horizon increases



This graph shows each **model architecture** performs over time, averaged across all France Levels

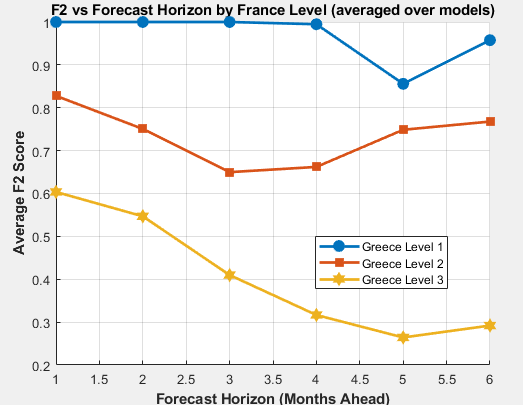
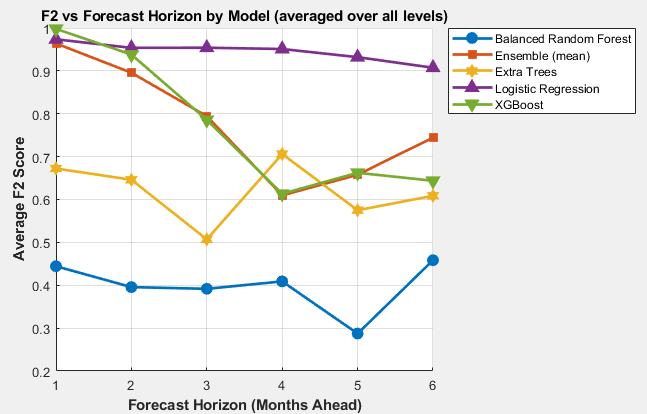
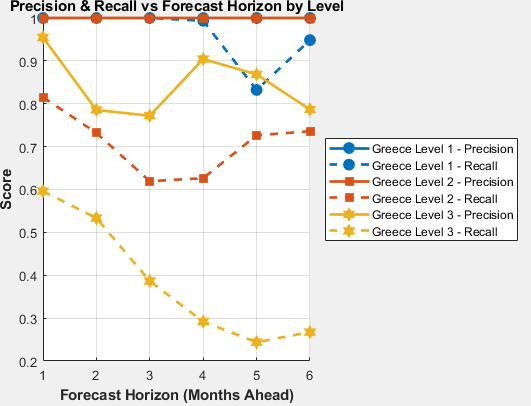
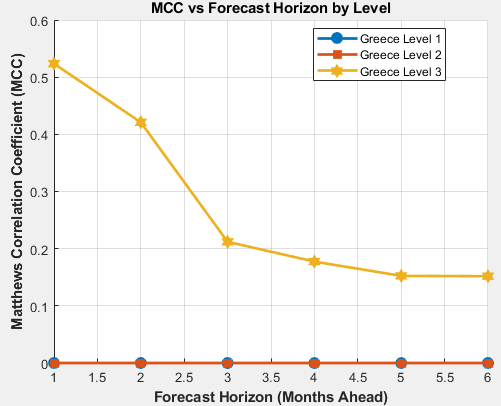
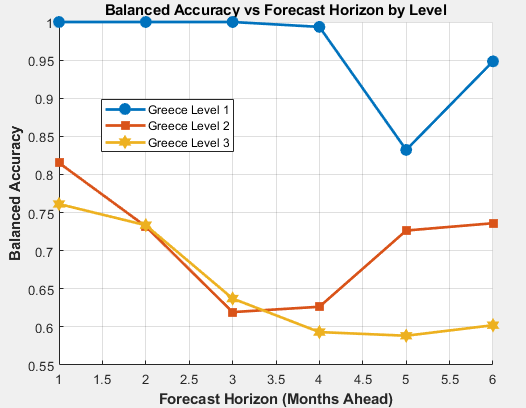
* All models’ F2 scores **decrease slightly** as horizon increases
* **XGBoost starts strongest** but falls fastest
* **Balanced Random Forest and Ensemble (mean)** degrade slower and maintain good performance at month 6
* **Logistic Regression performs worst overall** at longer horizons
  + Some models are more robust to longer timeframes (Random Forest), while others degrade quickly.



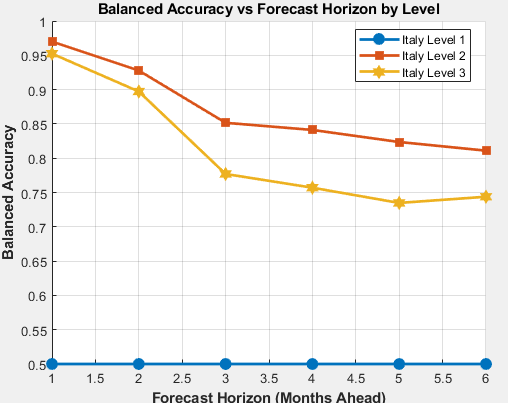
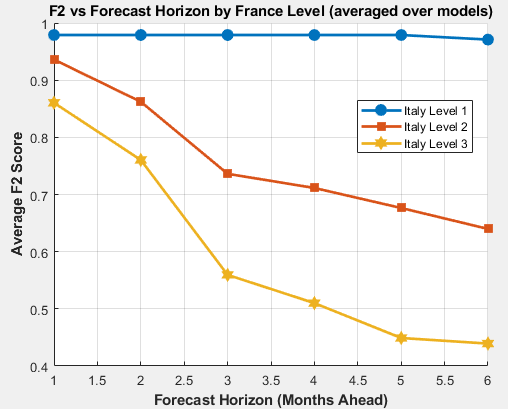
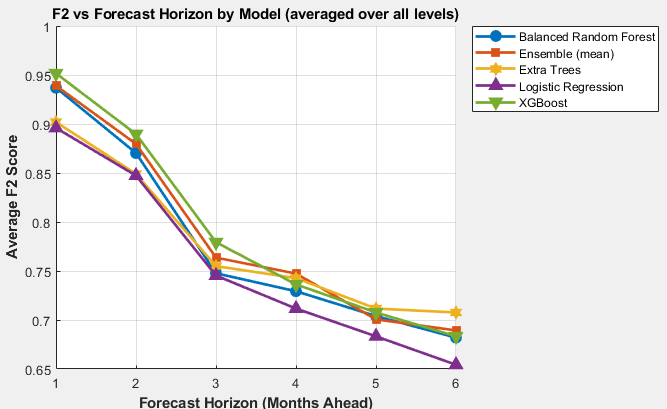
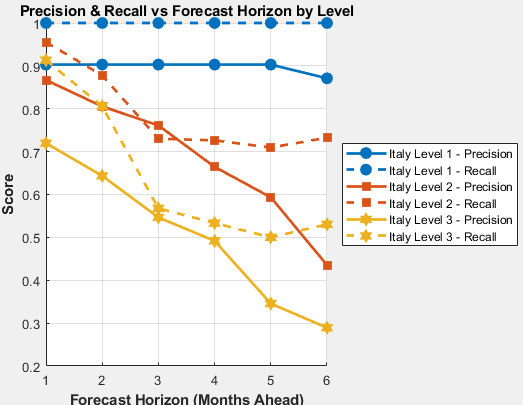
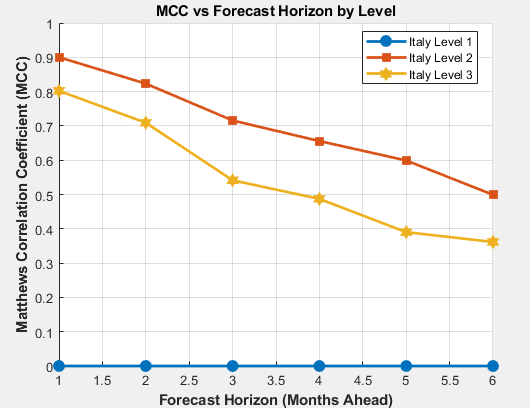
F2 score evolution per **France Level**, averaged over all models

* **France Level 3 performs best**, maintaining high F2 even at month 6
* **France Level 4 declines fastest**, likely because of Recall going down (showed above)
* All levels decline, but **some drop more steeply**, showing level-specific prediction difficulty
  + Forecasting gets harder the further the model trying to predict— but **some France levels are more predictable** than others

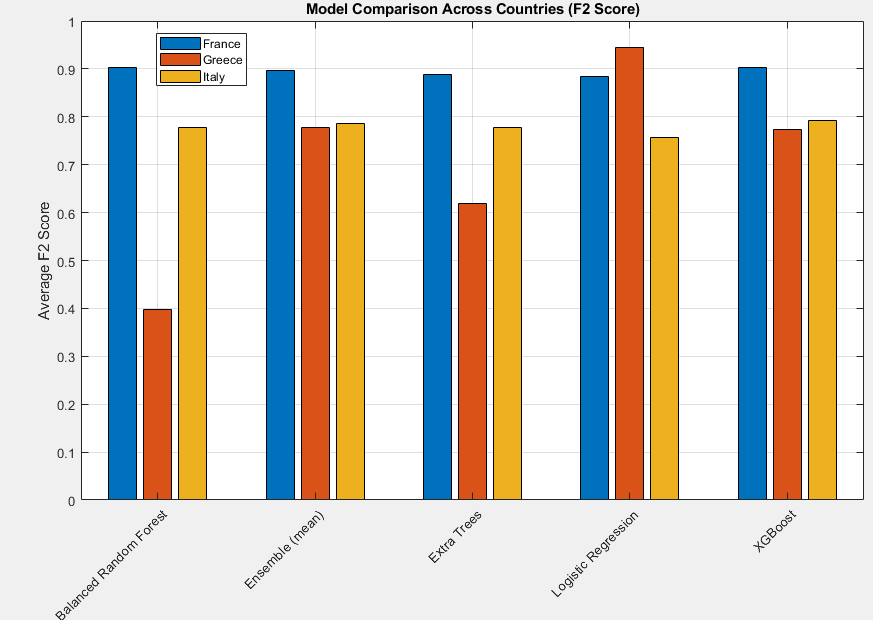
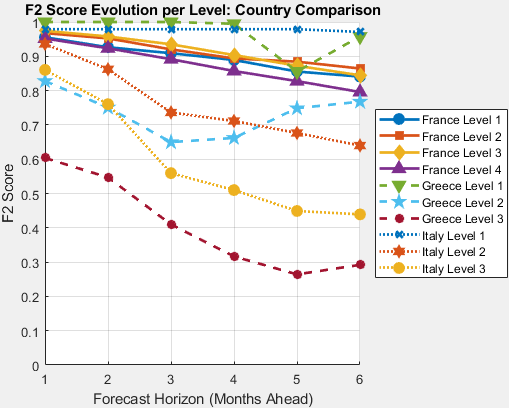
GREECE:



Italy:

Compare:



=== Best Models per Country ===

Country: France

Top models by F2:

1. XGBoost (0.904)

2. Balanced Random Forest (0.902)

3. Ensemble (mean) (0.896)

Top models by Balanced Accuracy:

1. XGBoost (0.830)

2. Ensemble (mean) (0.803)

3. Balanced Random Forest (0.801)

Country: Greece

Top models by F2:

1. Logistic Regression (0.945)

2. Ensemble (mean) (0.778)

3. XGBoost (0.774)

Top models by Balanced Accuracy:

1. Logistic Regression (0.922)

2. Ensemble (mean) (0.847)

3. XGBoost (0.822)

Country: Italy

Top models by F2:

1. XGBoost (0.791)

2. Ensemble (mean) (0.787)

3. Balanced Random Forest (0.778)

Top models by Balanced Accuracy:

1. XGBoost (0.732)

2. Ensemble (mean) (0.730)

3. Logistic Regression (0.727)