Project FY22-MG-006: Improved Model Ensembles for Prediction of Fusarium Head Blight

1. **What are the major goals and objectives of the research project?**
   - **a)** Continue expanding the FHB data matrix with observations from the Integrated Management Cooperative Project (IM-CP), but also estimate how many observations are needed for model stability.
   - **b)** Search for a stable FHB predictive model or ensemble of models that generalizes well to any environment (in the future or in other states).
   - **c)** Formally assess the stability of variable selection in Logistic Regression and Random Forest Machine Learning models proposed for FHB prediction.
   - **d)** Assess the predictive stability of Logistic Regression and Random Forest Machine Learning models developed to date.

2. **What was accomplished under these goals or objectives?** *(For each major goal/objective, address these three items below.)*

What were the major activities?

- Objective a: The modeling team (KSU/OSU) successfully completed integrating new observations into the data sets used for the development and evaluation of the FHB models. The total dataset used for modeling FHB was expanded by 20% during the 2023 project year and now contains over 1,200 observations.
- The addition of this information brings important new observations about the development of FHB from 22 states. The expanded dataset includes information about many contemporary cultivars and a range of genetic resistance levels (Table 1).

Table 1. Categories of genetic resistance represented in current FHB modeling dataset.

Resistance level	Number of Observations	Percent of total dataset
VS	115	9
S	502	39
MS	288	23
MR	366	29

- The expanded dataset describes disease development in a wide range of environments and disease intensities with FHB indexes ranging from 0 to >80% (Figure 1).
- The modeling team also completed a detailed evaluation of the base weather information used to develop the predictor variables for the FHB models. After some careful evaluation, we are exploring the possibility of the ERA-5 climate datasets for the FHB modeling projects. This ERA-5 datasets would allow us to leverage ongoing investments in the atmospheric sciences to develop and curate the best available environmental data. We are in the process of comparing these gridded weather products with station-based observations already contained in our datasets. The initial assessments of ERA-5 appear positive (temperature within 1.5 C and RH within 5%),

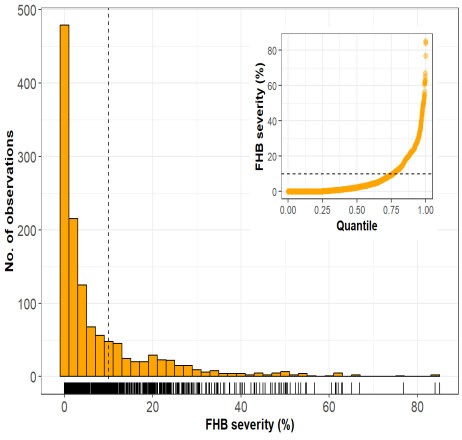
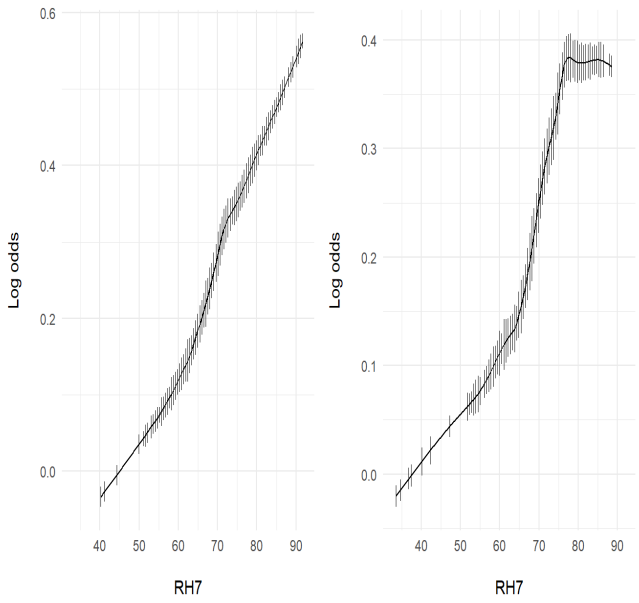


Figure 2. Intensity of FHB index and FHB severity across the ~1,200 observations in the FHB modeling dataset.

- We have calculated weather-based predictors identified in the previous generations of modeling FHB as well as promising variables identified by other FHB modeling groups globally. Graphical evaluations of these variables enable us to examine the distribution of these variables compared to the variables calculated based on stations-based observations.
- We have also fit Logistic Regression models to these new variables set to ensure that the biological interpretation of the variables remains consistent with our current understanding of FHB epidemiology. Plans are in place to fully transition to the new source of environmental data in the coming year (Figure 2).



- Objectives b-d: We are working through the complex coding required to conduct the evaluation of model stability required for objectives b-d. Preliminary results are helping us evaluate the model complexity (number of predictor variables) that can be supported by the current dataset without overfitting the model and compromising the stability of the model predictions in new environments. These assessments are essential if we are to obtain realistic estimates of model accuracy when the models are deployed for use by wheat and barley growers in the US.

What were the significant results?

- The development of forecasting models for FHB has direct influence on the disease management decisions of wheat and barley growers in the US. The cooperative FHB forecasting project and the IM-CP enable us to greatly expand the datasets used for modeling during this past project year.
- Comparisons of the ERA-5 gridded weather datasets with station-based observations of weather provide evidence that the gridded weather products offer key improvements to over station-based observations of weather without compromise in quality.

List key outcomes or other achievements.

- We now using ERA-5 gridded weather products in the FHB modeling projects. This transition accelerates the disease modeling process and effectively leverages the investments made in atmospheric sciences worldwide.
- Preliminary results of the model stability analysis are already helping guide the complexity of the models we develop and will enable realistic estimates of model accuracy when the models are deployed for use by wheat and barley growers in the US.

3. What opportunities for training and professional development has the project provided?

This project provides professional development opportunities for the research associate conducting the analysis and integrates advancements in statistical science into agricultural research and plant pathology specifically.

4. How have the results been disseminated to communities of interest?

The models developed in this project are applied within the FHB Prediction Center which delivers daily estimates of disease risk for 35 states where FHB is a critical production issue.

5. What do you plan to do during the next reporting period to accomplish the goals and objectives?

We plan to proceed with the proposed a methods applying stabilized regression to FHB forecasting. This approach involves fitting a model to each environment. It is possible that the set of predictors is different among the environments. However, what we are looking for are models that do not vary much between environments. Stabilized regression is then a weighted average of the individual models. We give higher weights to those predictors that generalize across environments and consider these variables as more important in prediction. We will evaluate the results graphically and examine which predictors are stable, those that are unstable and those that don't have much of an influence on prediction. The method will be implemented in the `StabilizedRegression` package within the R programing environment. The recently expanded dataset developed in the last project year now provides enough environments (including multiple observations per environment) to apply the proposed method.