Price Forecasting for Risk-Aware Fruit and Vegetable Contracting

A multi-output approach to forecasting prices and yield quantities for cucumbers, strawberries, and bell peppers across the Netherlands, Spain, and Italy

Nicolette Kuijt

June 29, 2026

Supervisors

VU: Prof.dr. Rob van der Mei
Picnic: Julie Witsen Elias

MSc Business Analytics

Reworked version for confidentiality

Vrije Universiteit Amsterdam
Fruit and Vegetables Team, Picnic Technologies

Management Summary

Research context. Picnic's Fruit and Vegetables (F&V) team [...Confidential information removed...]

The aim of this research is to investigate whether forecasting can provide a structured reference point for the buyer. The research tests whether external market and agricultural data can be used to estimate:

- expected future market prices during the contract window;
- the uncertainty range around those expected prices;
- and whether supply-side yield information can add useful context to the buying decision.

Research approach. The research combines external data on market prices, production statistics, weather observations, satellite vegetation indicators, and soil characteristics into one forecasting dataset covering 2004-2023. The focus is on cucumbers, peppers, and strawberries across the Netherlands, Spain, and Italy.

The product-country groups are modelled and evaluated separately, because a Dutch cucumber and a Spanish cucumber represent different buying contexts, with different seasonal calendars, production systems, and price dynamics. The final test period is 2021-2023, which is interpreted as a market-stress period because it overlaps with recent disruptions in European food and agricultural input markets, including COVID-19 and the 2022 energy crisis.

The thesis evaluates three forecasting components:

- **Monthly price forecasts** to estimate future market price levels during the contract window.
- **Annual yield forecasts** to explore whether external agricultural data can provide a supply-side signal.
- **Prediction intervals** to quantify the range of plausible future prices around the selected price forecast.

Finally, the forecast results based on external market data are validated against Picnic's internal buying prices. This step is important because [...Confidential information removed...]

Main findings. The main findings show that the useful outcome of this thesis is narrower than the full multi-output forecasting framework initially tested. Historical price information was the strongest forecasting signal, while the external feature set and annual yield target were less operationally useful in the current data setup.

- **Historical price-based forecasting performed best for monthly prices.** Holt-Winters was the strongest average point price forecast model in the six-month ahead contract-window setting, with an average WAPE of 24.5%. It outperformed the Naive Seasonal benchmark and the feature-based machine learning models.
- **Feature-based machine learning did not improve point price forecasts.** Random Forest and XGBoost did not consistently add value over historical price-based models. This does not mean that weather, satellite, soil, or production information is irrelevant for F&V markets. Rather, the available data depth and feature setup were not sufficient to make these variables improve the point forecasts.
- **Annual yield forecasting was statistically possible, but not operationally useful.** Naive Seasonal performed best for annual yield forecasting, but annual national yield is too coarse for Picnic's buying decision. It does not show whether supply pressure will occur during the actual contract window, nor whether that pressure affects the supplier regions or product specifications Picnic uses.

- **Prediction intervals added useful price-risk context.** Fixed Empirical was selected as the most suitable prediction interval method because it creates a transparent uncertainty range around the Holt-Winters forecast using historical forecast errors. QRF+CQR achieved higher average coverage, but mainly by producing wider and more complex intervals. Fixed Empirical provided the better overall balance between interval score, width, transparency, and usability in the contract pricing process.

Picnic validation. [...Confidential information removed...]

Recommendations. Based on the results, this thesis gives three main recommendations:

1. **Apply the chosen framework selectively.** The Holt-Winters and Fixed Empirical approach should first be used for product-country groups where the external and internal price series match well. [...Confidential information removed...]
2. **Improve the data foundation before adding more complex models.** The results do not suggest that model complexity should be the first next step. Feature-based machine learning, adaptive prediction intervals, and yield-based supply signals may become more useful once more forward-looking and granular data are available. Until that data is available, the transparent historical price-based approach is the most practical starting point.
3. [...Confidential information removed...]

[...Confidential information removed...]

Key takeaway. The main contribution of this thesis is not that a complex model replaces Picnic's current buying logic. This thesis shows where forecasting can realistically support the buying process today. As already implemented in the contract pricing process, historical price-based forecasting remains the strongest foundation, and a transparent uncertainty range can make price risk more explicit. For broader operational use, [...Confidential information removed...]

Contents

1 Introduction

Fresh fruit and vegetable sourcing is a decision problem where uncertainty is present long before the product reaches the customer. A buyer has to make commitments about price, supplier, country of origin, and expected volumes before the actual season has unfolded. At that moment, future market prices are still unknown, the quality and volume of the harvest are uncertain, and the effect of weather or market disruptions often only become visible during the season. This makes fruit and vegetables different from many other retail categories. The product is *biological*, *highly perishable*, and *strongly seasonal*. These characteristics mean that sourcing uncertainty does not only include the expected buying price, but also whether enough product will be available and whether the delivered quality will meet customer expectations. A decision, that looks reasonable before the season starts, can quickly become unfavourable when prices, supply conditions, or product quality move away from the expected pattern.

Recent market instability shows the limits of relying only on average seasonal patterns. This uncertainty has become more visible in recent years. In the European Union, the COVID-19 pandemic and the Russia-Ukraine War created a period of extraordinary market behaviour, during which agricultural input and energy prices increased sharply between 2020 and 2022 (Kornher et al., 2024). EU food price inflation during this period was driven mainly by rising agricultural production costs, while fruits and vegetables were particularly exposed to international price movements (Kornher et al., 2024). In the Dutch vegetable sector, this pressure was reflected in a 5.4% decline in production volumes and a 16.5% increase in fresh vegetable output prices in 2022 (Statistics Netherlands (CBS) and Wageningen Economic Research, 2022). These developments illustrate the effects of many outside conditions on fresh production markets. Average seasonal patterns remain important when predicting season conditions, but they do not always indicate by itself how uncertain the upcoming season is or how strongly it will deviate from previous years.

For Picnic, fruit and vegetables are a [...Confidential information removed...] category. This research was conducted within Picnic's Fruit and Vegetables (F&V) team. **[...Confidential information removed...]** Uncertainty in fresh production markets can therefore affect Picnic through several channels at once: *buying price*, *product availability*, and *delivered quality*.

Quality is especially important in Picnic's operating model because Picnic is an online supermarket, meaning that customers do not physically select their own fruit and vegetables. Instead, Picnic acts as the quality selector for every customer order, trying to fulfil the customer's quality expectations. This creates a trust-based relationship in which the customer expects the delivered product to match what they would have chosen themselves in a physical store (Ling et al., 2023). Fresh production therefore matters not only because of margin and volume, but also as it is one of the most visible indicators of the quality of Picnic products in the customer experience.

This thesis focuses on buying price uncertainty in the contract pricing process. F&V sourcing includes many uncertainties, such as future supply, product quality, availability, and customer demand. This thesis focuses specifically on the margin-related part of sourcing: *uncertainty around future buying prices*. In Picnic's F&V contract pricing process, **[...Confidential information removed...]**

For that reason, this thesis evaluates forecasts separately for crop-country combinations. The thesis focus is on **three products**, cucumbers, bell peppers, and strawberries, across **three countries**, the Netherlands, Spain, and Italy.

These uncertainties in making contracting decisions motivated this thesis's multi-output forecasting framework. To support crop-country specific buying decisions, this thesis investigates whether part of the uncertainty at the decision moment can be quantified using external agricultural and market data. The framework combines three components. *First*, monthly price forecasting is used to estimate the future market price level during the sourcing period. This is directly connected to the contract-versus-spot buying decision, because buying prices determine whether a fixed contract is likely to protect or pressure margin. *Second*, annual yield forecasting is explored as a possible supply-side signal, because harvest outcomes can also influence market availability and price pressure. Yield is treated separately from price, because the available yield data are annual, while market prices are reported at monthly cadence, and the buying decisions are made on a seasonal basis. *Third*, price range prediction intervals are constructed around the price forecast to also quantify the range of plausible future prices.

This last uncertainty layer is important because a point forecast alone does not show how reliable the expected price level is. A contract price slightly above the expected spot price may still be reasonable when future market prices are highly uncertain. The same contract price may be unattractive when the expected price range is narrow and stable. Together, the price forecast, the yield signal, and the expected price range form a multi-output forecasting framework designed to support buyer judgement in making contract-versus-spot buying price decisions.

Existing agricultural forecasting research provides a starting point for this framework. Previous studies show that weather observations, satellite vegetation indices, soil characteristics, and historical production data can contain useful information for agricultural forecasting (Klompenburg et al., 2020; Paudel et al., 2022). Machine learning methods, especially tree-based ensemble models, have been applied successfully in crop forecasting settings where tabular environmental data are available (Huber et al., 2022). However, much of this research focuses on staple crops and yield prediction. Fresh production creates a different forecasting problem, because fruits and vegetables have short production cycles, high perishability, and rapid price responses to external supply disruptions. A frost night or heat wave can quickly change market supply and prices (Hatfield and Prueger, 2015).

The agricultural price forecasting literature is also relevant, because the main commercial decision in this thesis depends on future market prices. This literature includes statistical time-series models, machine learning models, and deep learning approaches (Guindani et al., 2024). However, these studies often evaluate forecast accuracy in general terms, while the connection to a concrete buying decision is usually limited. In Picnic's contract-versus-spot buying setting, average accuracy alone is not enough. The forecast needs to be interpretable because the buyer must be able to connect the output to familiar concepts such as seasonal price patterns, weather stress, and market volatility. The forecast also needs to include uncertainty, because the decision depends on the range of possible future prices rather than only on the expected price. This creates a literature gap for a framework that combines price forecasting, supply-side agricultural signals, and price uncertainty prediction in a way that can be interpreted within an actual buying process. This research addresses that gap in the context of Picnic's F&V buying process, focusing on cucumbers, bell peppers, and strawberries across the Netherlands, Spain, and Italy.

The hypothesis in this thesis is that fruit and vegetable prices are largely driven by strong seasonal patterns, but that external agricultural and market features add value by identifying deviations from seasonal patterns. **[...Confidential information removed...]** The research, therefore, tests whether feature-based models can improve making point predictions over seasonal benchmarks, and whether uncertainty methods can quantify the residual risk around the selected best-performing price forecast model.

Evaluation is based on the 2021-2023 test period. However, since this period overlaps with recent disruptions in European food and agricultural input markets, the final results are interpreted as performance under demanding market conditions rather than under a fully stable price regime. This makes the evaluation stricter, but also relevant to the business problem, because uncertainty support matters most when prices are unstable.

To structure the research, one **main research question** and five sub-questions were defined.

Main research question: How can a multi-output forecasting framework, combining historical market prices and external agricultural features, support risk-aware contract-versus-spot buying decisions for fruit and vegetables at Picnic?

Sub-questions:

1. How can multi-source weather, production, and market data be harmonized into a robust dataset for fresh production price and yield forecasting across crop-country combinations?

2. To what extend do feature-based machine learning models improve price forecasts over seasonal baselines for cucumbers, strawberries, and bell peppers across the Netherlands, Spain, and Italy?

3. Can external agricultural data produce reliable annual yield forecasts that serve as a supply-side input signal for fruit and vegetable buying decisions?

4. How can prediction intervals quantify price uncertainty around the selected price forecast and support the assessment of contract-versus-spot buying risks?

5. How can the developed forecasting framework support Picnic's contract-versus-spot buying process, and what is required to move from forecast evaluation to operational use?

Thesis structure. The remainder of this thesis is structured as follows. Chapter 2 maps Picnic's F&V sourcing process and explains the contract-versus-spot buying decision in more detail. Chapter 3 reviews the relevant literature on agricultural forecasting, price forecasting, feature selection, and prediction interval forecasting. Chapter 4 describes the external data sources and the preprocessing steps used to construct the final modelling dataset. Chapter 5 presents the baseline models, point forecasting methods, prediction interval methods, and yield forecasting setup. Chapter 6 defines the evaluation metrics and validation procedure. Chapter 7 presents the results for point price forecasting, yield forecasting, price range interval forecasting, and Picnic validation. Finally, Chapter 8 answers the research questions and discusses the main limitations, operational implications, and directions for future research.

2 Business Process Mapping

[...Confidential information removed...]

3 Literature Review

Fresh production markets combine three properties that make forecasting particularly difficult: *short production cycles*, *high perishability*, and *strong sensitivity to short-term weather events*. These properties mean that supply-side shocks can translate rapidly into price movements. Price volatility transmission is even stronger for more perishable products (Pan and Zheng, 2023), and historical patterns may be disrupted by a single extreme weather event. Despite this operational relevance, agricultural forecasting literature has concentrated mainly on staple crops rather than fresh fruit and vegetables (Klompenburg et al., 2020). As increased climate variability and extreme weather events reduce the reliability of traditional growing seasons, the need for accurate regional forecasts before harvest has increased (Hernández Hernández et al., 2025; Hatfield and Prueger, 2015). While agricultural forecasting historically relied on process-based models that represent biological growth mechanisms, the past decade has seen a shift toward data-driven machine learning approaches that learn patterns directly from historical observations. This review discusses four areas of literature relevant to this thesis: *agricultural forecasting and supply-side drivers*, *agricultural price forecasting*, *feature selection and interpretability*, and *prediction interval forecasting*.

3.1 Agricultural Forecasting and Supply-Side Drivers

The first stream of literature focuses on agricultural production and yield forecasting, which are important supply-side drivers of fresh production markets.

Process-based models. Traditional agricultural forecasting systems are built on models that simulate plant development through biophysical equations. One example is WOFOST, which is used in the European Union's MARS forecasting platform to produce seasonal production forecasts for EU member states (Velde and Nisini, 2019). A central concept in such models is growing degree days (GDD), which quantify accumulated thermal time and are commonly used to predict phenological stages such as flowering (McMaster and Wilhelm, 1997). Alongside thermal accumulation, temperature stress thresholds are important because extreme heat or frost can damage plant tissue and reduce yields (Hatfield and Prueger, 2015). Velde and Nisini (2019) evaluated the MARS platform over a 20-year period and found that it can provide reliable seasonal forecasts when properly calibrated, but that accuracy declines in years with extreme weather events that are poorly represented in the historical calibration data. This limitation is directly relevant for fruit and vegetable forecasting, where short-term shocks can quickly affect supply.

Remote sensing. More recently, the availability of remote sensing data via satellite images has expanded agricultural forecasting by providing spatially detailed observations across large production regions. For example, López-Lozano et al. (2015) showed that one-kilometre satellite observations can capture within-region variability in grain yield forecasts, although finer resolution also increases sensitivity to cloud contamination. Lecerf et al. (2018) compared forecasts based on biological model outputs with meteorological indicators and found that combining both sources improved accuracy. This suggests that process-based information and data-driven indicators can be complementary rather than substitutes.

Machine learning shift. Despite the high level of interpretability offered by rule-based biological models, these models often struggle with computational scalability when applied across many regions simultaneously and require extensive local calibration (Klompenburg et al., 2020). Machine learning has offered a robust alternative by learning patterns directly from historical data without requiring explicit biological mechanisms. Several systematic reviews have documented this transition. For instance, Klompenburg et al. (2020) analysed 50 studies and found that ensemble methods, particularly random forests, consistently outperformed individual algorithms. However, they also highlighted ongoing challenges regarding the "black box" nature of machine learning models, meaning that the reasoning behind the model's predictions cannot be easily explained (Ryo, 2022). More recent reviews by Jabed and Azmi Murad (2024) and Hernández Hernández et al. (2025) confirmed these findings, while noting growing interest in model explainability techniques and hybrid architectures that attempt to merge biological rules with data-driven flexibility.

Tree-based models. Tree-based ensemble methods have proven particularly effective for agricultural forecasting using tabular data, such as historical weather and soil records (Paudel et al., 2022; Huber et al., 2022). Jeong et al. (2016) demonstrated that random forests could predict yields effectively at both

global and regional levels by capturing non-linear interactions between weather, soil, and management variables. Furthermore, their research showed that models trained on geographically diverse observations generalized better than those using only site-specific calibrations. Building on this research, Paudel et al. (2021) developed a framework for forecasting yields across European NUTS2 regions, the EU's standard geographic classification for regional statistics, by integrating satellite vegetation indices, daily weather data, and soil characteristics into monthly aggregated features. This framework achieved strong predictive performance several months before harvest, and Paudel et al. (2022) later scaled the approach to 26 European countries. Their comparison of random forests, gradient boosting, and neural networks showed that tree-based methods consistently performed well for tabular agricultural data and that spatial feature engineering, such as lagged production values and regional soil properties, significantly improved cross-regional performance. Hybrid approaches have also been explored, for example by using crop simulators to generate features for machine learning models, although calibrating the process-based component remains challenging (Maestrini et al., 2022).

Deep learning. Research has also explored deep learning to capture more complex temporal and spatial dependencies. Khaki and Wang (2019) showed that deep neural networks with dropout regularization could predict corn and soybean yields across U.S. counties using weather sequences and historical yield data. Khaki et al. (2020) extended this work with a CNN-RNN framework to combine spatial feature extraction with temporal dependencies. They found that attention mechanisms outperformed standard long short-term memory (LSTM) networks by learning to emphasize critical growth windows. Other studies similarly found that CNN-LSTM architectures can improve accuracy, but at the cost of interpretability and higher data requirements (Oikonomidis et al., 2022). Notably, Huber et al. (2022) showed that XGBoost matched or exceeded deep learning performance for yield estimation while requiring less training data and computational power. This suggests that deep learning benefits do not always transfer to tabular agricultural datasets. Since most agricultural forecasting research focuses on staple crops such as maize and wheat, fruit and vegetable forecasting remains comparatively underexamined (Klompenburg et al., 2020).

Supply-side takeaway. Overall, this literature shows that weather, satellite, soil, and historical production data can contain useful information for agricultural forecasting. However, the evidence comes mainly from yield forecasting for staple crops at regional or national scale. For this thesis, external agricultural data are therefore relevant as possible supply-side indicators, but not automatically sufficient for the F&V buying problem, where forecasts must be timely, crop-country specific, and connected to price risk.

3.2 Agricultural Price Forecasting

Price as market signal. Where agricultural yield forecasting starts from crop development and supply conditions, price forecasting starts from market behaviour, including historical price dynamics, demand, logistics, and market structure. The two perspectives are complementary, as supply-side shocks caused by weather events can translate into price movements, while prices are also shaped by demand, logistics, and market structure. Understanding both streams of research is therefore necessary for a forecasting approach that aims to support buying decisions in fresh production markets. Recent reviews of agricultural commodity forecasting show that this field includes statistical time-series methods, machine learning methods, neural networks, ensemble models, and hybrid approaches (Guindani et al., 2024).

Time-series models. Classical time-series methods are part of agricultural price forecasting because agricultural prices often contain trend, seasonality, and short-term persistence (Guindani et al., 2024). For example, Sabu and Kumar (2020) forecasted monthly arecanut prices in Kerala using SARIMA, Holt-Winters Seasonal Method, and LSTM, showing how traditional time-series models and deep learning models are both used in agricultural price forecasting. Holt-Winters is especially relevant in this thesis because it is based on exponential smoothing and updates level, trend, and seasonal components over time. Taylor and Letham (2018) describes an additional decomposable time-series model Prophet that separates trend and seasonality and allows for flexible trend changes. Both Holt-Winters and Prophet mainly focus on the structure of the historical price series itself. By contrast, feature-based machine learning models can often also include external explanatory variables such as weather and market features.

Machine learning. Machine learning and deep learning methods have also been applied to agricultural commodity price forecasting. For example, Manogna et al. (2025) compared stochastic, machine learn-

ing, and deep learning approaches for agricultural commodity prices, illustrating that agricultural price forecasting has moved beyond purely statistical time-series models. Other studies focus more specifically on volatility and regime changes, modelling sudden shifts between low- and high-volatility states. For example, Avinash et al. (2024) used hidden Markov guided deep learning models for nonlinear and non-stationary agricultural commodity price data, using technical indicators such as moving averages and frequency-based transformations.

External variables. Most of the reviewed price forecasting studies rely mainly on historical price behaviour or market-related indicators, but some recent work has started to include external variables. For example, Min et al. (2025) constructed a multivariate time-series dataset that combines wholesale prices of agricultural commodities with weather variables and week numbers, and compared neural network models for price prediction. Their study showed that weather information can be included in agricultural price forecasting. However, several recent studies rely on neural network architectures, while regional fresh production forecasting often has limited data availability at the crop-country level. This leaves room for research that combines price histories with regional supply-side indicators, while also considering models that are suitable to smaller tabular crop-country datasets.

Price forecasting takeaway. Taken together, the price forecasting literature shows that agricultural prices can be modelled using both historical price structure and more flexible machine learning approaches. However, the reviewed studies mainly evaluate forecast accuracy, not whether forecasts support sourcing decisions made before the season has unfolded. For Picnic's contract-versus-spot setting, this distinction matters because the buyer needs both an expected future price and a way to judge whether a contract offer is attractive, risky, or too uncertain relative to plausible future spot prices.

3.3 Feature Selection and Model Interpretability

Feature selection. Agricultural forecasting literature often uses high-dimensional feature sets that combine weather, soil, satellite, production, and market variables. This makes feature selection important, because too many variables can increase computational cost, add noise, and increase the risk of overfitting, while removing too many variables may discard useful predictive information (Gong et al., 2024). Several studies therefore use filter methods to reduce the feature space before model training. For example, Cheng et al. (2022) proposed an MI-VIF algorithm that combines mutual information with the variance inflation factor (VIF), allowing the selection of informative features while reducing multicollinearity. Similarly, Gong et al. (2024) developed a filter approach combining Pearson correlation with mutual information to capture both linear and nonlinear dependencies between variables. This thesis also uses several external data sources that create a large set of weather, satellite, soil, and market variables relative to the number of observations available per crop-country model, making these suggested feature selection methods an important part of the data processing.

Interpretability. Modelling with such a large set of features also raises questions about the interpretability of the model output. Tree-based models such as random forests and XGBoost provide feature importance measures, which can help identify which variables contribute most to the forecast (Breiman, 2001; Chen and Guestrin, 2016). Wang et al. (2024a) found that these built-in importance rankings from tree-based models can perform well during feature selection, while SHAP values add more computational cost when used only for that purpose. However, explainability remains important for operational agricultural decision-making, because model outputs need to be connected to factors that sourcing stakeholders can understand, such as weather conditions, vegetation indicators, and market history (Ryo, 2022; Paudel et al., 2023). For this thesis, interpretability is therefore not treated as a separate objective, but as a requirement for whether a forecast can be trusted and used in the buying process.

Interpretability takeaway. For this thesis, feature selection and interpretability are therefore not only technical steps. They determine whether a model can be explained in terms that are meaningful for the F&V buyer, such as seasonal price history, weather stress, vegetation conditions, and market volatility. This is important because a model that is slightly more accurate but difficult to interpret may be less useful in a negotiation context than a simpler model with a clearer rationale.

3.4 Prediction Intervals and Forecasting Uncertainty

Most crop forecasting research produces point predictions, meaning single expected values. However, contracting decisions depend not only on the expected price, but also on the range of plausible future prices. Prediction intervals are therefore relevant because these express forecast uncertainty around an expected value rather than reporting only one predicted outcome (Chatfield, 1993). In a contract-versus-spot buying setting, this matters because the same point forecast can lead to different buying interpretations depending on whether the surrounding uncertainty range is narrow or wide.

Empirical intervals. A simple way to construct prediction intervals is to use the historical forecast error distribution around a point forecast. When forecast errors are assumed to follow a normal distribution, this leads to a parametric Gaussian interval around the expected forecast value (Chatfield, 1993). However, fresh product prices can contain sharp spikes, skewness, and heavy-tailed deviations, making this assumption restrictive. A more direct alternative is to estimate interval bounds from empirical forecast residuals, for example by using historical lower and upper residual quantiles (Lee and Scholtes, 2014). This produces a transparent residual-based interval without assuming normality, although it does not adapt interval width to current market or production conditions.

Quantile-based intervals. Quantile regression forests (QRF) address forecast uncertainty by predicting the full conditional distribution of the response variable rather than just the mean, providing prediction intervals without requiring the data to follow a normal distribution (Meinshausen, 2006). Gyamerah et al. (2020) applied QRF to maize forecasting and used kernel density estimation to create smooth probability distributions, finding that this approach captured uncertainty more effectively than traditional parametric models. Similarly, Wang et al. (2024b) used QRF to quantify how weather variability affects cotton yields under different climate scenarios, demonstrating the model's ability to handle complex environmental inputs. Although these applications focus on yield rather than price, they show why QRF is relevant for agricultural settings with nonlinear environmental drivers and non-normal uncertainty.

Coverage calibration. While QRF can estimate conditional quantiles, the resulting prediction intervals are not guaranteed to achieve the intended coverage in finite samples. Conformalized quantile regression addresses this by adding a calibration step to quantile-based prediction intervals (Romano et al., 2019). The method uses a separate calibration set to measure how often observations fall outside the predicted quantile interval and then adjusts the interval width accordingly. This makes conformal calibration relevant in settings where uncertainty forecasts need to be evaluated not only by how narrow the interval is, but also by whether it covers the realized outcome sufficiently often. Alternative probabilistic approaches also exist, such as the model developed by Newlands et al. (2014), which propagates weather uncertainty through biological growth simulations. However, this approach remains dependent on the heavy calibration of rule-based models.

Literature gaps relevant to this thesis:

1. **Fresh production gap:** agricultural forecasting research focuses mainly on yield and staple crops, while fruit and vegetable price forecasting receives less attention.
2. **Supply-side integration gap:** price forecasting studies often rely on historical price behaviour or market indicators, while regional supply-side information such as weather, satellite, soil, and yield data is less often integrated.
3. **Decision-risk gap:** prediction interval forecasting remains limited in agricultural sourcing contexts, even though sourcing decisions depend on uncertainty as well as expected prices.

The feature-selection and interpretability literature adds one design requirement: the framework must remain explainable enough for the buyer to understand and challenge the model output.

Together, these gaps define the thesis design: monthly price forecasting is used to study fruit and vegetable price dynamics, external agricultural data are included as possible supply-side signals, annual yield is explored separately because of its limited data structure, and prediction intervals are evaluated because contract-versus-spot decisions depend on risk as well as expected prices.

4 Data Methodology

This chapter explains how the external data sources were turned into the modelling dataset used in the following chapters. Because the framework combines monthly price forecasts, annual yield outcomes, and regional environmental variables, the main data challenge is to align sources that differ in timing, geography, and completeness. The chapter therefore first describes the *data sources and collection strategy*, then discusses the main *data quality steps*, *exploratory analysis*, *feature engineering*, and *final dataset characteristics*.

4.1 Data Sources and Collection Strategy

The data for this research is based on multiple external data sources carefully joined into one dataset. The dataset combines meteorological observations, satellite vegetation indicators, soil characteristics, crop production statistics, and market prices. These sources were selected to capture both the production environment and the market outcomes relevant for price forecasting and yield forecasting. This methodology builds on the regional crop forecasting approach of Paudel et al. (2022), who demonstrated the value of combining weather, satellite, and soil-related information for crop forecasting across European regions. This thesis adapts that logic to the F&V sourcing context by focusing on cucumbers, strawberries, and peppers, and by adding monthly market prices as the main forecasting target. This differs from Paudel et al. (2022), who focused on staple cereals and root crops and did not model market prices.

The forecasting framework relies on external data for two main reasons. **[...Confidential information removed...]** External data therefore provide a market-level foundation for forecasting, while the translation to Picnic's operational buying process is evaluated separately in the validation chapter.

The data sources differ in temporal and spatial resolution. The research uses six primary data source categories, each capturing a different part of the agricultural production system. Table 4.1 shows an overview of these sources together with their temporal and spatial characteristics. As shown in the last two columns of this table, the temporal resolution and spatial coverage differ considerably between sources, calling for careful alignment into a unified dataset suitable for modelling. This is the **main data integration challenge** in this chapter: daily, monthly, annual, static, regional, and national observations all need to be combined without changing the meaning of the original data. This integration challenge was addressed as shown in the data processing pipeline illustrated in Figure 4.1.

Table 4.1: External data sources used for the price and yield forecasting dataset.

Data source	Provider	Type	Temporal coverage	Temporal resolution	Spatial coverage
Crop production	Eurostat	Production statistics	1970–2024	Annual	National
Meteorological data	JRC MARS / Agri4Cast	Weather observations	1979–2024	Daily	25km gridded
Satellite indices	Copernicus Global Land	Vegetation	1999–2024	Monthly	Gridded raster
Soil properties	ESDB (JRC)	Soil	Static	Static	NUTS2 regional
Market prices	Eurostat & national agencies	Prices (EUR/100 kg)	2002–2023	Monthly	National

Note. Data sources include (Eurostat, 2026; Toreti, 2026; Copernicus Land Monitoring Service, 2026; European Commission, Joint Research Centre, 2026; Statistics Netherlands (CBS), 2024; ISMEA, 2024).

Spatial alignment was performed at NUTS2 level. After collecting the external data sources, the first harmonization step was spatial alignment, as shown in Figure 4.1. This was needed because the sources were not reported in the same geographical format. As shown in Table 4.1, weather observations were available as station coordinates, satellite indicators as gridded raster data, soil characteristics at regional level, and crop production statistics and market prices mainly at national level. To combine these sources, NUTS2 regions were used as the common regional layer where possible.

NUTS, the Nomenclature of Territorial Units for Statistics, is the European Union's regional classification system (Eurostat, 2016). NUTS2 was selected as the target spatial resolution because it is the finest scale at which agricultural statistics were consistently available across the target countries, while still capturing meaningful agroclimatic variation between production areas. It also fits Picnic's sourcing context, where decisions are made at the regional to national level rather than at the individual farm or pixel level. As an example, the Netherlands consists of 12 NUTS2 regions, which correspond to the Dutch provinces.

The geographic focus of this research is on three countries representing distinct European production systems: the Netherlands for intensive greenhouse production, Spain for Mediterranean warm-climate production, and Italy for more seasonal and partly open-field production systems. Weather stations and satellite rasters were spatially matched and aggregated through coordinates to NUTS2 regions. Crop production statistics and market prices were not disaggregated below national level, because this would require unsupported allocation assumptions. Instead, national values were assigned to all retained NUTS2 regions within the same country. As a result, regional variation enters the model through explanatory variables such as weather, satellite, and soil characteristics, while price and yield targets remain national-level outcomes. This distinction is important when interpreting the model results.

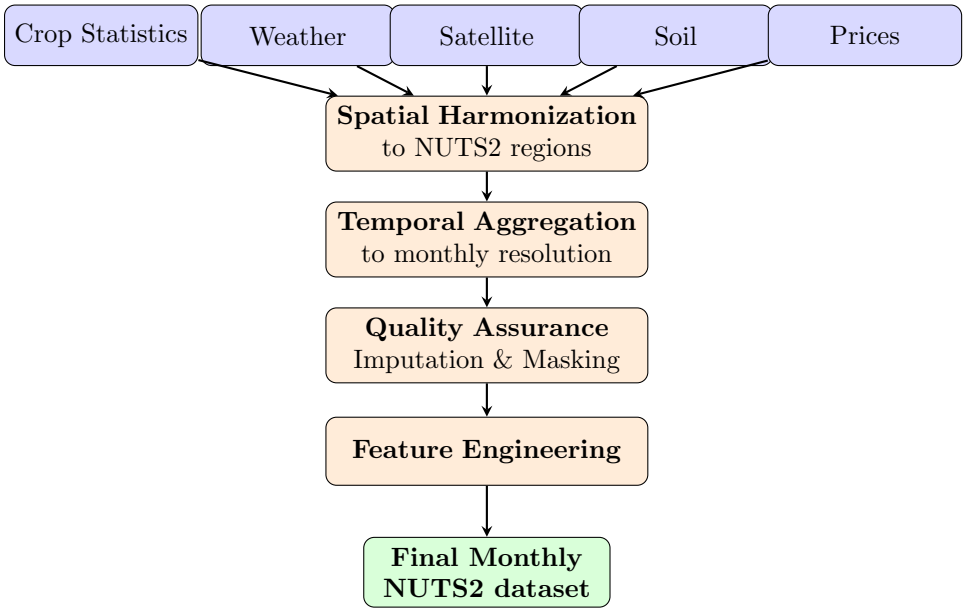

Figure 4.1: Data integration pipeline from external source tables to the final monthly NUTS2 modelling dataset.

Temporal alignment was performed to a monthly frequency. Next to spatial harmonization, the sources also differ in temporal resolution. Weather data are available daily, while most other sources operate at monthly, annual, or static frequency. The dataset was therefore aligned to monthly resolution, because monthly market prices are the main target variable in the price forecasting task. Daily weather observations were aggregated to monthly features, while annual crop production statistics were retained at their yearly frequency and linked to the monthly rows within the corresponding year.

Daily weather dynamics were summarised into monthly features in a way that preserves information on short-term extreme weather events. Temperature features include monthly means, percentiles, minima, maxima, frost-day counts, and heat-stress-day counts. Precipitation features include monthly totals, intensity distributions, and heavy-rainfall-day counts. This is important because agricultural impacts often result from extreme events rather than monthly averages (Hatfield and Prueger, 2015). For example, a single frost night can severely damage strawberry production even if the monthly mean temperature remains moderate, while a few days of extreme heat can affect fruit development regardless of overall monthly warmth.

Annual crop production statistics were not disaggregated across months, because production is not uniformly distributed across the year and such a split would require additional assumptions. This preserves the original meaning of the annual yield target, but it also creates an important limitation: yield can only be evaluated as an annual national outcome, not as a monthly sourcing signal. This constraint is addressed

in the modelling strategy discussed in Section 5.4. After spatial and temporal harmonization, the sources were integrated using DuckDB, an analytical database system suited to large-scale aggregation workflows (Raasveldt and Mühleisen, 2019). Figure 4.1 summarises the resulting pipeline from raw source tables to the analysis-ready monthly NUTS2 modelling dataset.

4.2 Data Quality Assurance

After the external sources were joined, several data quality steps were applied to make the dataset complete enough for modelling while keeping it focused on relevant production regions and months. The main checks concerned *missing values*, *outliers*, and *regional and seasonal masking*.

Missing Value Analysis. Data completeness varied across sources. Weather data had minimal missing values, and satellite vegetation indices from Copernicus showed strong post-2000 coverage. The largest data gaps appeared in market prices from Eurostat. Eurostat mandates price data submission only when a crop exceeds a production significance threshold (European Parliament and Council of the European Union, 2022). Consequently, Spanish crop prices are more consistently available for several products, while Dutch and Italian crop prices more often fall below the reporting threshold in some months and are therefore absent from Eurostat databases. To improve coverage, market prices gathered from national statistical sources were added, including CBS for the Netherlands and ISMEA for Italy (ISMEA, 2024; Statistics Netherlands (CBS), 2024). Where national series overlapped with Eurostat, a scaling factor was calculated as the median ratio between the two sources and used to impute missing Eurostat values. This assumes that both series follow the same underlying price movement during overlap periods, while allowing for a stable level difference.

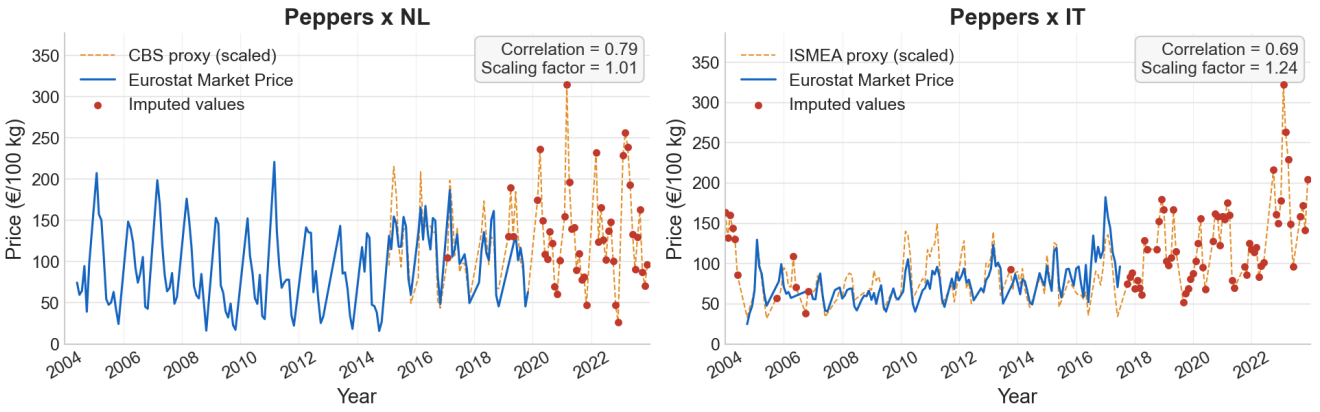

Figure 4.2: Price gap imputation for peppers using national proxy series aligned to Eurostat market prices during overlapping periods. Left: Netherlands peppers using CBS data as proxy. Right: Italy peppers using ISMEA data as proxy.

Figure 4.2 illustrates this procedure for two pepper series: one relatively stable alignment case for the Netherlands and one more uncertain imputation case for Italy. For peppers from the Netherlands, the correlation between Eurostat and CBS prices was 0.79 across 44 overlapping months, with a scaling factor of 1.01. This indicates that the CBS proxy follows the Eurostat series closely during the overlap period, so the imputed values mainly fill missing months without strongly changing the price level. For Italian peppers, the ISMEA-Eurostat overlap correlation was lower, at 0.69 across 112 months, with a scaling factor of 1.24. Imputed observations account for 37% of the Italian pepper price series, so this group should be interpreted with caution. The main risk is that the imputation may understate Italian pepper price volatility, because the scaled proxy assumes a stable level difference between the two sources. Unusual price movements that appear in one source but not the other during non-overlapping periods would therefore not be fully reflected in the imputed values.

Outlier Detection. Outlier identification was done using two complementary statistical methods: **Z-score analysis**, which identifies values more than three standard deviations from the mean, and **Interquartile Range (IQR) analysis**, which flags values falling beyond 1.5 times the IQR from the first

or third quartile. Together, these methods provide both a mean-based and a distribution-based diagnostic for unusual observations (Leys et al., 2013). The analysis revealed that weather and satellite variables showed minimal outliers, suggesting no broad data quality problem in the environmental measurements. Outlier flags were used as a diagnostic step rather than as an automatic removal rule, because extreme values can represent real agricultural stress events rather than data errors. Values were only removed or corrected when they could be traced to clear placeholder values or impossible measurements. No major weather or satellite observations were removed solely because they were statistically extreme.

Regional and Seasonal Masking. An important data quality step was the application of regional and seasonal masks to focus the dataset on relevant production regions and months. Not all NUTS2 regions produce each target crop, and even regions that do produce certain crops may not be active throughout the full year. Including region-month combinations with zero or negligible production would add noise and could create misleading relationships between environmental variables and national market outcomes. The masks were developed in consultation with the F&V buyer and category manager at Picnic, who have extensive domain knowledge about European production geographies. Regions without documented annual production were excluded, together with months outside the relevant production season. For example, northern Spanish regions were excluded for cucumbers and peppers, which are concentrated in southern Mediterranean production zones, while Dutch cucumber and pepper production is mainly excluded in winter due to insufficient sun hours. These masks created a more focused dataset where weather and satellite observations align temporally with relevant production activity.

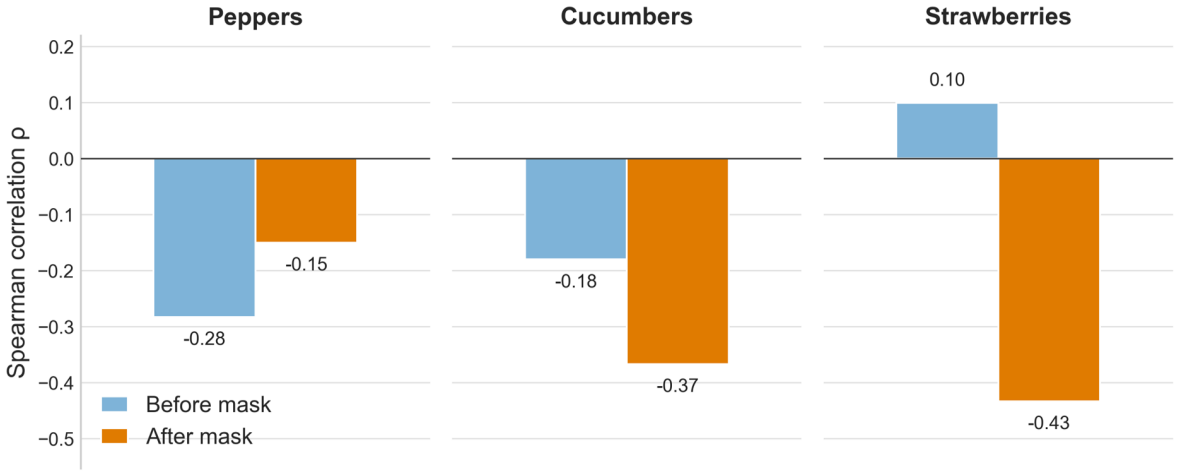

Figure 4.3: Correlations between Leaf Area Index (low vegetation) and the monthly price target, shown per crop before and after applying the regional and seasonal mask.

As a result, for peppers, the mask reduced potential observations by 30.8%, for cucumbers by 22.2%, and for strawberries by 39.2%. Figure 4.3 shows how the masking step affects the relationship between the satellite vegetation indicator leaf area index (LAI) and the monthly price target. The figure compares Spearman correlations before and after applying the regional and seasonal mask. For cucumbers, the relationship becomes more strongly negative after masking. For strawberries, the relationship changes from weakly positive to more strongly negative, while for peppers the absolute correlation becomes weaker. This shows that masking does not uniformly increase correlations across crops, but changes the observed signal by removing region-month combinations that are not relevant for actual production. Therefore, the mask is interpreted as a methodological relevance filter rather than as a guaranteed accuracy-improving transformation.

4.3 Exploratory Data Analysis

The exploratory analysis uses the 2004-2018 training period for feature inspection and correlation analysis, so that modelling choices are not informed by outcomes from the validation or test periods. This follows the temporal split used throughout the thesis, where 2004-2018 is used for training, 2019-2020 for validation, and 2021-2023 for final testing. The data split is discussed in more detail in Section 4.5. Full

price series are shown to provide context on the later evaluation period, but feature relationships and selection diagnostics are based on the training data only. Across the crop-country combinations, **three patterns** were most relevant for the modelling strategy: recurring seasonal price movements, strong short-term price persistence, and weaker but still interpretable relationships between environmental variables and prices.

Figure 4.4 shows the monthly price development for peppers in the Netherlands, Spain, and Italy. Peppers are used as the main illustrative example because the three country series show clear seasonality as well as strong cross-country differences in volatility, which are central challenges for the later modelling task. The 12-month rolling trend indicates that prices shift over longer periods as well as within the annual cycle. The validation and test periods are highlighted to show that the 2021-2023 test period falls in a more volatile part of the price history for several series. This visual pattern is consistent with the interpretation of this test period as a market-stress period, discussed in more detail in Section 4.5.

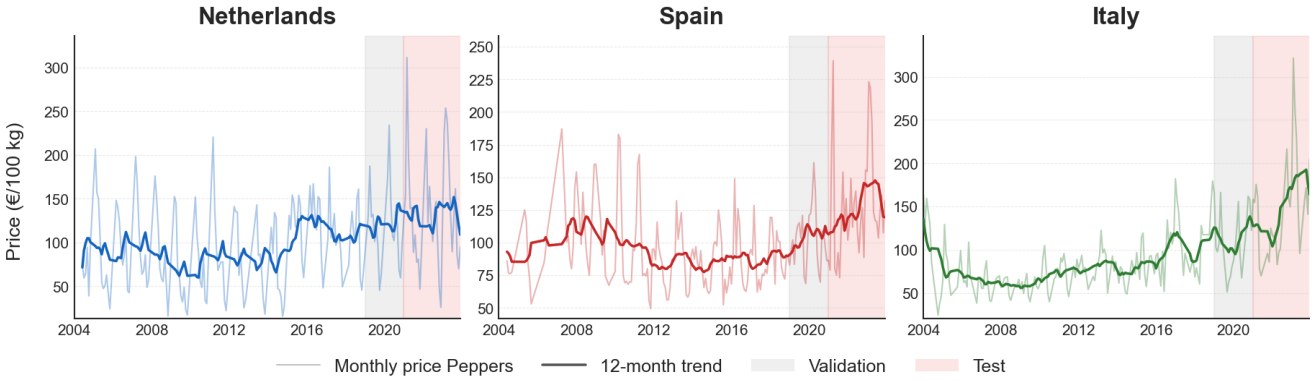

Figure 4.4: Monthly pepper prices in the Netherlands, Spain, and Italy, with a 12-month rolling trend and the validation and test periods highlighted.

To inspect the strongest direct feature-target relationships, Spearman correlations were calculated between selected features and the next-month price target. These correlations were computed only on the training period. Figure 4.5 shows the average strongest correlations across the nine crop-country combinations. Price-based features show the strongest positive correlations, especially short-term exponential moving averages, the 12-month price lag, and the 1-month price lag. This confirms that recent price history and seasonal price history contain the clearest direct signal for future prices.

Weather and satellite variables show weaker and more mixed correlations than price-based features. Frost days have a positive average correlation with next-month price, while average temperature and average daily radiation show negative average correlations. This is consistent with the seasonal structure of fresh production markets, where warmer and brighter production periods can coincide with higher supply and lower prices. However, variation across crop-country combinations is large, so these correlations should not be interpreted as simple causal relationships. Weather effects are nonlinear: moderate warmth may coincide with higher supply and lower prices, while extreme events such as heat stress can increase price pressure. Satellite indicators such as the fraction of absorbed photosynthetically active radiation (FAPAR) and LAI show weaker direct correlations on average, but still describe production conditions that may be useful in combination with crop, country, and seasonal information.

Overall, the EDA indicates that price dynamics are the dominant short-term signal, while weather and satellite variables provide additional context on production conditions. This supports the feature engineering strategy in which raw environmental observations are transformed into cumulative, lagged, and difference-to-normal features before being used in the models. The full monthly price series for all crop-country combinations and the detailed crop-country Spearman correlation heatmap are included in Appendix Figures 10.1 and 10.2.

4.4 Feature Engineering and Selection

Following the exploratory analysis, features were engineered to represent seasonal, biological, and market-related patterns in a form suitable for model training. These features were included as candidate predictors based on biological reasoning and observed training-period patterns, but were not assumed to be predic-

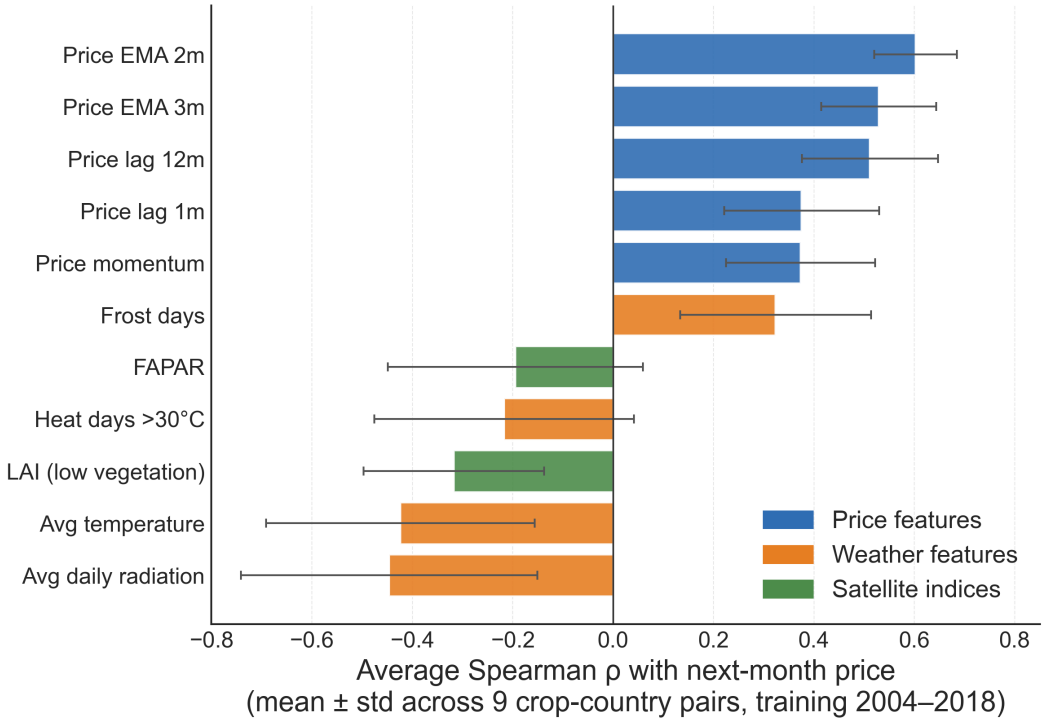

Figure 4.5: Average Spearman correlations between selected features and next-month price across the nine crop-country combinations, computed on the training period 2004-2018.

tive beforehand. The later feature selection process determines which variables contain useful signal for each crop-country combination.

Cumulative Weather Features. Crop development depends on accumulated energy and resource inputs over time rather than only on conditions in a single month (McMaster and Wilhelm, 1997). To capture this, cumulative features were constructed, such as three-month accumulated solar radiation and accumulated temperature indicators. **Growing Degree Days (GDD)**, a standard phenological indicator in agricultural science (McMaster and Wilhelm, 1997), was added to summarize the thermal time above a base temperature threshold:

$$\mathrm{GDD} = \sum_{i=1}^{n} \max\left(0,\ \frac{T_{\max,i} + T_{\min,i}}{2} - T_{\mathrm{base}}\right), \tag{4.1}$$

where $T_{\max,i}$ and $T_{\min,i}$ are the maximum and minimum temperature on day i, T_{base} is the base temperature below which crop development slows down, and the summation occurs over the chosen aggregation window. A shared base temperature of 10°C was used across the three products. This is a simplifying assumption, because crop-specific base temperatures can differ, but it provides a consistent thermal accumulation indicator within the monthly dataset.

Lag Features. Temporal lag features were created to capture both seasonal patterns and short-term persistence in this time-series dataset. Twelve-month price lags provide a seasonal reference, reflecting the annual cyclicity in agricultural markets where prices in a given month tend to relate to prices in the same month of the previous year (Manogna et al., 2025; Avinash et al., 2024). Shorter price lags capture recent market momentum. Lagged weather variables were also constructed to account for delayed biological responses between growing conditions and market-observable outcomes. In addition, exponential moving averages (EMA) of prices were created as smoothed trend indicators, and rolling price volatility was calculated over six-month windows to capture market instability. Cross-commodity price lags were included for related crops, based on exploratory evidence that peppers and cucumbers partially co-move. All lag, moving-average, and rolling-volatility features were calculated using only past observations available at the forecast origin.

Difference-to-Normal Features. Difference-to-normal features were created to capture whether current conditions deviate from what is typical for the same region and period. For each feature, the standardised deviation was calculated as $\frac{X_t - \mu_{\text{hist}}}{\sigma_{\text{hist}}}$, where X_t is the current observation and μ_{hist} and σ_{hist} are the historical mean and standard deviation calculated from the training data. This prevents information from the validation or test periods from entering the feature-building process. These features are used to identify unusual weather or vegetation conditions relative to the historical pattern.

Extreme Event Counts. Agricultural production can respond nonlinearly to extreme conditions, where a small number of severe weather days can have a larger effect than monthly averages suggest (Hatfield and Prueger, 2015). Therefore, daily weather observations were also converted into monthly counts of extreme events. These include frost days, heat-stress days, and heavy-rainfall days. These features allow the model to distinguish between a month with moderate average conditions and a month with short but potentially damaging extreme events.

Feature Selection. The feature engineering process generated a large set of candidate predictors across weather, satellite, soil, production, and market-related categories. A multi-stage selection pipeline was applied to reduce this set to a stable and interpretable feature set for each crop-country combination, combining correlation-based screening, mutual information, multicollinearity filtering, and model-based importance ranking (Cheng et al., 2022; Gong et al., 2024). Near-zero variance features were removed first. The remaining predictors were then screened using Spearman correlation to remove features with little monotonic association with the target ($|r| \leq 0.05$). Features removed at this stage were re-evaluated using mutual information, so that predictors with nonlinear signal could still be retained. VIF filtering was then applied to reduce multicollinearity. Finally, Random Forest permutation importance was used separately for each crop-country combination to retain the strongest predictors, resulting in 26 to 28 features per crop-country model. All feature selection steps were fitted on the training data only. Validation performance was used only during later model tuning, not to refit the feature selection filters. A complete overview of the feature count after each pipeline step is provided in Appendix Table 10.2.

4.5 Final Dataset Characteristics

This final section summarises the dataset that results from the data integration, quality assurance, exploratory analysis, and feature-selection steps described above. The resulting dataset spans 2004-2023, providing 20 years of monthly observations selected to balance data quality and the available temporal depth. After regional and seasonal masking, the dataset consists of 7,560 region-month-product observations across 14 NUTS2 regions: 5 in the Netherlands, 3 in Spain, and 6 in Italy. The masking process reduced the number of potential observations by 22-39%, depending on the crop-country combination, and concentrates the analysis on regions and months with documented production activity.

Table 4.2: Final dataset summary statistics

Attribute	**Value**
Temporal Coverage	2004-2023 (20 years, 240 months)
Temporal Resolution	Monthly
Spatial Coverage	14 NUTS2 regions (NL: 5, ES: 3, IT: 6)
Products	Cucumbers, Strawberries, Peppers
Total Observations	7,560 region-month-product combinations
Feature Categories	Weather, satellite, soil, production, and temporal features
Candidate Predictors	165 for price models and 166 for yield models
Selected Features	26-28 per crop-country model after feature selection
Train / Validation / Test	2004-2018 / 2019-2020 / 2021-2023
Split Proportions	Approximately 75% / 10% / 15%

The effective sample size differs from the total number of dataset rows. It is important to distinguish between the total number of rows in the integrated dataset and the effective number of target observations available to each model. The models are trained separately per crop-country combination,

so the aggregate total of 7,560 region-month-product observations does not represent the sample size of one model. For price forecasting, each crop-country model is trained on region-month rows, but the monthly price target is observed at the national level and is therefore shared across retained NUTS2 regions within the same country. The number of unique monthly price targets available to each price model ranges approximately from 53 to 180 across crop-country groups, with the full breakdown provided in Appendix Table 10.1. This variation is mainly caused by seasonal masking, which removes inactive production months, and by differences in the starting year of available price data. Yield forecasting is more constrained because yield is reported annually at national level. The repeated monthly NUTS2 rows therefore do not create independent yield outcomes. After accounting for this, each crop-country yield model contains only approximately 14 to 16 distinct annual target values in the training period. This difference is central to the modelling strategy, because price forecasting can be evaluated as a monthly time series, while yield forecasting remains exploratory.

Time-series forecasting requires temporal split strategies that respect the order of observations to prevent data leakage, where future information would inappropriately influence model training (Tashman, 2000). The dataset was therefore split into training data from 2004-2018, validation data from 2019-2020, and test data from 2021-2023. All exploratory analyses, feature-engineering parameters that require historical baselines, and feature-selection procedures were fitted using the training data only. This ensures that no information from the validation or test periods enters the model development process.

The 2021-2023 test period represents a market-stress window. This period overlaps with the extraordinary market behaviour following the COVID-19 pandemic and the Russia-Ukraine war, during which EU food price inflation was driven mainly by agricultural production-cost increases (Kornher et al., 2024). For the Dutch vegetable sector specifically, 2022 also showed lower production volumes and higher fresh vegetable output prices (Statistics Netherlands (CBS) and Wageningen Economic Research, 2022). The final test results should therefore be read as model performance under stressed market conditions rather than under stable historical price dynamics.

5 Mathematical Modeling

The goal of the modelling process is to forecast monthly market prices and annual yields in a way that supports Picnic's contract-versus-spot buying decisions. Since this decision does not only depend on the expected price level, but also on the uncertainty around that expectation, the modelling process includes both point forecasts and prediction intervals. The approach is exploratory, as multiple model families are evaluated against the same benchmarks to assess whether additional model complexity and external agricultural features add predictive value over simpler seasonal baselines. The chapter first introduces the baseline models that are used to test how much predictive signal is already captured by simple historical price patterns. It then describes the six-month forecast window and the point forecasting models used for monthly price prediction. After that, the chapter explains how prediction intervals are constructed around the best selected point forecast. Finally, yield forecasting is discussed separately at the end of the chapter, because the annual yield target creates a structurally different modelling problem.

5.1 Baselines

Before fitting any machine learning model, three baseline models were created to measure how much predictive information is already given in the historical price series itself. Any machine learning model that cannot outperform these baselines would not justify the added complexity **[...Confidential information removed...]** The three baselines are:

1. **Historical Mean.** The Historical Mean baseline predicts the unconditional average of the target variable from the training period for all future time points (Hyndman and Athanasopoulos, 2021, Sec. 5.2). It ignores all seasonal fluctuations in the data and is therefore expected to provide a weak comparison point.

2. **Naive Seasonal.** For monthly price forecasting, the Naive Seasonal baseline predicts the price observed in the same calendar month one year prior (Hyndman and Athanasopoulos, 2021, Sec. 5.2). So, the predicted price for a month is the observed price in that month from last year. For annual yield forecasting, the same logic is applied at yearly frequency by using the previous year's yield. This baseline is especially relevant because it reflects the strong seasonality in fruit and vegetable markets and resembles a transparent decision process based primarily on historical seasonal prices.

3. **Linear Regression.** The last baseline is a Linear Regression model fitted on the five most important features identified using training-period feature rankings. The Linear Regression baseline is included as a simple feature-based comparison. It provides a low-complexity test of whether the engineered feature set contains useful signal before moving to more flexible machine learning models.

Together, these baselines provide reference points for comparing simple historical, seasonal, and low-complexity feature-based forecasts. Their performance on the 2021-2023 test set is discussed in Section 7.1.

Forecast window setup. The price forecasting setup is designed to simulate Picnic's contract pricing process. In practice, the buyer needs a view on the expected market price over an upcoming buying period rather than only for one isolated future month. A six-month forecast window was therefore used as the main operational setup in this thesis. This horizon is long enough to cover a typical seasonal buying period and the weeks before the season in which contract prices are evaluated. The forecast is implemented as a *rolling six-month window*. Each year, a forecast is made at the end of December predicting monthly prices for January through June, corresponding to horizons $h = 1$ to $h = 6$. A second forecast made at the end of June predicts monthly prices for July through December. Repeating this procedure over the 2021-2023 test period creates a sequence of forecast windows that simulates the information needed in the contract pricing process. This setup ensures that every forecast only uses information that would have been available at the *forecast origin*, when a contract offer is evaluated by the F&V buyer.

5.2 Point Prediction Models and Model Selection

Four model types were evaluated for point price forecasting: Long Short-Term Memory (LSTM) networks, Random Forest (RF), XGBoost, and Holt-Winters exponential smoothing (HW). These models represent different forecasting logics. LSTM was included as a deep learning approach for sequential data, RF and XGBoost as feature-based machine learning models that can use external agricultural and market features, and HW as a classical time-series model that estimates level, trend, and seasonality directly from the historical price series.

All models were trained separately for each available crop-country price series. For example, one model was trained for Spanish cucumbers and another model for Dutch cucumbers. This architecture was chosen because Picnic's contract-versus-spot buying decision is made at crop-country level, where Dutch cucumbers and Spanish cucumbers are separate buying contexts with different seasonal calendars, supply regions, and price dynamics. The combination of all nine markets into a single model with dummy variables for crop and country was also tested, but produced equal or worse accuracy because the heterogeneity between markets was too large to be captured by the shared model with the available data depth. The chosen per-crop-country setup also keeps the results more interpretable, since model outcomes and feature importance findings can be read as statements about a specific market rather than as averages over all markets.

LSTM. LSTM networks were evaluated against the baseline models, as this deep learning approach is designed to learn temporal patterns from sequential data. However, deep learning models generally require large datasets to learn such patterns effectively and are often limited in interpretability due to their black-box nature, which reduces their suitability for decision-making contexts where explainability is required (Huber et al., 2022). In this study, the per-crop-country datasets were relatively small, which limited the ability of the LSTM models to learn price dynamics with sufficient accuracy. Appendix Figure 10.3 illustrates this for peppers in the Netherlands in the six-month forecast setting. The LSTM forecast follows part of the broader seasonal movement, but remains too smooth and does not reproduce the magnitude of the observed price peaks and fluctuations. Moreover, even if predictive performance were competitive, the lack of interpretability on LSTM outcomes would limit practical applicability within Picnic, where contracting decisions must be transparent and justifiable.

LSTM is retained only as an exploratory model. As the example shown in Appendix Figure 10.3, forecasts were too smooth for the sharp price movements relevant to the contract pricing process, and their limited interpretability makes it less suitable for operational use at Picnic.

Random Forest and XGBoost. Given the weak validation performance and limited interpretability of the LSTM model, more interpretable machine learning techniques were explored. Random Forest (RF) was included because tree-based ensemble methods have performed well in agricultural forecasting studies with tabular weather, soil, satellite, and historical production features (Jeong et al., 2016; Paudel et al., 2022; Huber et al., 2022). RF constructs multiple decision trees on bootstrap samples of the training data and aggregates their predictions, typically by averaging (Breiman, 2001). This bagging mechanism reduces variance and improves generalization performance, making RF relatively robust to overfitting. This is particularly relevant given the limited number of unique monthly price observations per crop-country model, as reported in Appendix Table 10.1.

In contrast to RF, XGBoost is based on gradient boosting, where decision trees are built sequentially and each new tree is trained to correct the residual errors of the previous ensemble (Chen and Guestrin, 2016). This sequential learning process enables the model to capture complex, non-linear interactions in the data. However, this increased flexibility requires careful tuning of the hyperparameters to achieve optimal performance and avoid overfitting (Chen and Guestrin, 2016). Within the context of crop-related forecasting problems, XGBoost has been shown to achieve competitive performance compared to deep learning approaches, while offering improved interpretability through feature importance analysis (Huber et al., 2022). For RF and XGBoost, hyperparameters were tuned using a 60-iteration randomised search on the validation period, while keeping the test set unseen. The full search spaces are reported in Appendix Table 10.3.

Holt-Winters. Next to these feature-based machine learning models, HW exponential smoothing was evaluated as a classical time-series model. This method estimates each price series through a level component, a trend component, and a seasonal component (Sabu and Kumar, 2020). In this thesis, the seasonal component is treated multiplicatively, because the size of seasonal price movements can depend on the underlying price level so should scale with changing price levels. In simplified notation, the h-step-ahead forecast can be written as

$$\hat{p}_{t+h|t} = (\ell_t + h \cdot b_t) \cdot s_{t+h}, \tag{5.1}$$

where $\hat{p}_{t+h|t}$ is the price forecast for future month $t + h$ based on information available at time t, ℓ_t is the estimated level, b_t is the estimated trend per time step, s_{t+h} is the seasonal factor for the forecasted month, and h is the forecast horizon. The formula shows that HW does not simply repeat last year's price, but combines the current estimated price level with an extrapolated trend and a seasonal adjustment.

HW is closely related to the Naive Seasonal baseline, because both methods rely only on historical prices and do not use external agricultural features. However, HW is less rigid than Naive Seasonal. Instead of copying the price from the same month in the previous year directly, it smooths past observations and updates the estimated level, trend, and seasonal structure over time. Figure 5.1 illustrates this difference where it can be seen that Naive Seasonal follows last year's observed seasonal pattern directly, while HW adjusts the seasonal forecast upward or downward when the recent level and trend of the series have changed. Therefore, HW tests whether a smoothed seasonal-trend structure can add value over a strict last-year seasonal benchmark. This is especially relevant in markets where the price level shifts after abnormal market conditions, because a direct last-year comparison may then be too rigid. At the same time, HW remains transparent because it uses only historical price information.

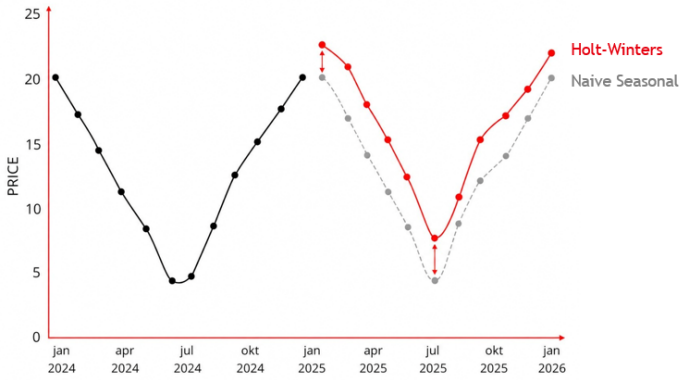

Figure 5.1: Difference between Naive Seasonal and Holt-Winters forecasting. Naive Seasonal copies the same calendar month from the previous year, while Holt-Winters updates the seasonal pattern using the estimated level and trend.

Model comparison logic. The point forecasting comparison follows the modelling logic introduced above. The baselines test how much future price variation can already be explained by simple historical and seasonal patterns. HW then tests whether this historical price logic improves when the seasonal pattern is allowed to adapt through level and trend updates. RF and XGBoost test whether external agricultural and market features add predictive information beyond historical price structure. This comparison determines which model is most suitable as the point-forecast anchor for the prediction interval prediction methods. The full point-forecast results are reported in Section 7.1.

However, a point forecast alone is not sufficient for Picnic's contract-versus-spot buying decision. **[...Confidential information removed...]**

The point forecast provides the expected price level, but not the buying risk around that expectation. Because even moderate forecast errors can translate into substantial margin impact for high-volume products, the selected point forecast is used as the anchor for the prediction intervals in Section 5.3.

5.3 Prediction Interval Forecasting

The point forecasting comparison provides an expected future price level, but it does not show how uncertain that expectation is. This motivates the prediction interval forecasting step. In the main interval framework, HW is used as the point-forecast anchor, so the prediction intervals are constructed around the errors that remain after the HW forecast. The key quantity is therefore the **HW residual**,

$$\varepsilon_{t+h} = p_{t+h} - \hat{p}^{\mathrm{HW}}_{t+h|t}, \tag{5.2}$$

where p_{t+h} is the realized market price at future month $t+h$, and $\hat{p}^{\mathrm{HW}}_{t+h|t}$ is the HW forecast for that month made using information available at time t. The residual ε_{t+h} represents the remaining forecast error after accounting for the estimated level, trend, and seasonal pattern in the historical price series. The prediction interval forecasting task is therefore to estimate a plausible range for this residual and apply that range around the HW point forecast.

As an initial benchmark, a Gaussian residual interval was considered, which is a common parametric approach for constructing prediction intervals around forecast errors (Chatfield, 1993). However, this approach assumes that forecast residuals are approximately normally distributed. Figure 5.2 shows a Q-Q plot of the pooled HW residuals across all crop-country combinations in the training period. The deviations from the normal reference line, especially in both tails, indicate that the residual distribution is not well approximated by a Gaussian distribution. This suggests that large positive and negative price deviations occur more often than a Gaussian model would imply. This is important in the contract pricing process, because these extreme deviations are precisely the cases where uncertainty estimates matter most for the contract-versus-spot buying decision. The Shapiro-Wilk test also rejects normality for the pooled residuals ($W = 0.779$, $p < 0.001$) (Shapiro and Wilk, 1965). Moving forward, the analysis therefore focuses on prediction interval methods that do not impose a normality assumption.

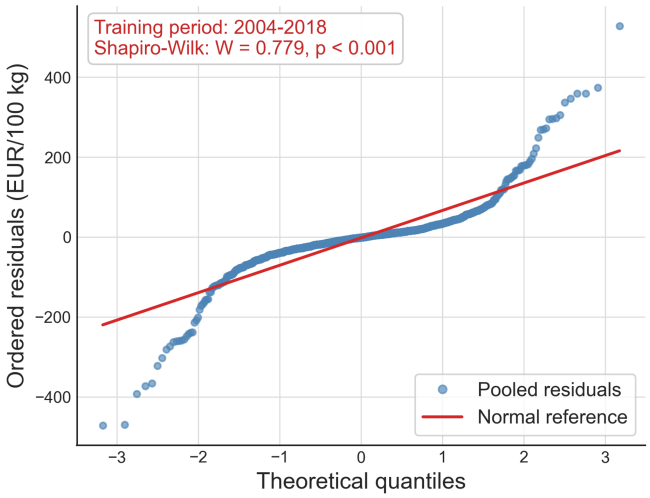

Figure 5.2: Q-Q plot of pooled HW residuals across all crop-country combinations in the 2004-2018 training period. The red line shows the normal reference distribution.

Fixed Empirical. The Fixed Empirical method constructs an interval directly from historically observed HW residuals, following the idea of empirical prediction intervals based on historical forecast errors (Lee and Scholtes, 2014). For each crop-country group and forecast origin, the 10th and 90th percentiles of the residuals available up to that origin are calculated. The resulting 80% prediction interval is defined as

$$\left[\hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}^{\mathrm{hist}}_{0.10}(t), \quad \hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}^{\mathrm{hist}}_{0.90}(t)\right], \tag{5.3}$$

where $\hat{q}^{\mathrm{hist}}_{0.10}(t)$ and $\hat{q}^{\mathrm{hist}}_{0.90}(t)$ are the historical 10th and 90th residual percentiles available at forecast origin t. This makes the method transparent and operationally simple: it asks how large HW forecast errors have been historically and applies the middle 80% of those errors around the new HW forecast.

As fixed empirical range is calculated at each forecast origin, the interval width is fixed within each six-month forecast window, because the same residual quantiles are applied to all six monthly forecasts

issued at the same origin. However, the width can change slightly across forecast origins as more historical residuals become available. Fixed Empirical is therefore a transparent, historical-error approach for creating prediction intervals, but it does not adapt within the forecast window to current market, weather, or volatility conditions.

Simple QRF. Simple QRF applies Quantile Regression Forests (QRF) directly to the future price target p_{t+h}, instead of modelling the residual around HW. QRF extends Random Forest by estimating conditional quantiles of the response variable rather than only the conditional mean (Meinshausen, 2006). In this direct-price setup, the model estimates the lower and upper price quantiles as

$$\left[\hat{q}_{0.10}(p_{t+h} \mid \mathbf{x}_t), \quad \hat{q}_{0.90}(p_{t+h} \mid \mathbf{x}_t)\right], \tag{5.4}$$

where $\mathbf{x}_t$ denotes the feature vector available at time t.

This method is explored to tests whether the future price distribution can be learned directly from the available features, without first anchoring the forecast to a seasonal time-series model. It is useful as a comparison because it separates the effect of QRF itself from the effect of modelling residuals around HW. However, with limited crop-country data, directly learning the full future price distribution is difficult. The model has to learn both the seasonal price level and the uncertainty around that level from the same small dataset. Simple QRF is therefore included in the method comparison to evaluate whether direct price quantiles are sufficient, or whether residual-based intervals around HW are more appropriate.

Residual QRF. Instead of directly forecasting the price level, residual QRF applies the QRF framework to the HW residual target ε_{t+h} rather than to the raw future price (Meinshausen, 2006). The model estimates residual quantiles $\hat{q}_\tau(\varepsilon \mid \mathbf{x}_t)$ conditioned on the current feature vector $\mathbf{x}_t$. These residual quantiles are then added to the HW point forecast:

$$\left[\hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}_{0.10}(\varepsilon \mid \mathbf{x}_t), \quad \hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}_{0.90}(\varepsilon \mid \mathbf{x}_t)\right]. \tag{5.5}$$

Residual QRF tests whether the engineered features can explain the *conditional residual distribution*: when HW is likely to be too high, too low, or more uncertain than usual. In contrast to Fixed Empirical, the interval width of residual QRF can vary with the available market, weather, production, and satellite features. However, its reliability depends on whether the training data contain enough comparable residual patterns. Extreme events that are not represented in the historical data cannot be learned and predicted reliably. Residual QRF is therefore important as the adaptive residual model, but its intervals may still be miscalibrated when the model is overconfident.

QRF with conformal calibration (QRF+CQR). To reduce the risk of miscalibrated prediction intervals, this research also evaluates conformal calibration. Specifically, QRF+CQR builds on Residual QRF by adding a conformal calibration step (Romano et al., 2019). Specifically, QRF+CQR builds on residual QRF by adding a conformal calibration step. Residual QRF produces conditional quantile estimates, but these estimates are not guaranteed to achieve the intended 80% coverage in finite samples. Conformalized Quantile Regression (CQR) addresses this by applying a post-hoc calibration step that adjusts the interval width based on calibration errors (Romano et al., 2019). For each observation in the *calibration set*, a non-conformity score is computed as

$$s_i = \max\left(\hat{q}_{0.10}(\varepsilon \mid \mathbf{x}_i) - \varepsilon_i,\ \varepsilon_i - \hat{q}_{0.90}(\varepsilon \mid \mathbf{x}_i)\right), \tag{5.6}$$

which measures how far the observed residual falls outside the predicted residual interval. A correction term δ is then derived from these scores and applied to both interval bounds:

$$\left[\hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}_{0.10}(\varepsilon \mid \mathbf{x}_t) - \delta, \quad \hat{p}^{\mathrm{HW}}_{t+h|t} + \hat{q}_{0.90}(\varepsilon \mid \mathbf{x}_t) + \delta\right]. \tag{5.7}$$

This correction widens the interval when the calibration residuals show that the original Residual QRF interval is too narrow. The formal coverage guarantee relies on *exchangeability*, meaning that calibration and test observations should be drawn from the same underlying distribution. In this thesis, that assumption should be interpreted carefully because the 2021-2023 test period contains unusual price

dynamics that are only partly represented in earlier years as seen in Figure 10.2. QRF+CQR is therefore useful because it can correct overconfident residual intervals, but the correction often comes at the cost of wider intervals. This makes QRF+CQR the main adaptive alternative to compare to the simple, transparent Fixed Empirical method in the results.

The four interval methods differ in two main levels: whether they are anchored around the HW residuals, and whether the interval width is fixed or feature-adaptive. Fixed Empirical represents the transparent historical-error approach, Simple QRF estimates price quantiles directly, Residual QRF estimates feature-conditioned residual quantiles, and QRF+CQR adds a calibration step to those residual intervals. Their empirical performance is compared in Section 7.5. Building the distributional framework around HW residuals also changes the role of environmental and market features. In the direct price setting, these features may help explain the overall seasonal price level. In the residual setting, the main seasonal-trend structure is already captured by HW. The features therefore only add value if they help explain the remaining forecast errors, such as unusually large deviations caused by weather stress, market volatility, or supply pressure. Whether the monthly feature set captures these residual risk patterns sufficiently is discussed further in the prediction interval forecasting results in Section 7.5.

5.4 Yield Forecasting

Yield forecasting is treated separately because the structure of the available data creates constraints that do not apply to the monthly price forecast. Yield in tonnes per hectare is published by Eurostat as an *annual national statistic*, which means that when it is assigned to the monthly feature dataset, every month within a given calendar year receives the same yield label. The dataset contains up to 180 monthly rows per crop-country group, but these rows do not represent 180 independent annual outcomes. Because yield data is only available at national level while the feature dataset operates at NUTS2 regional level, there is also only one distinct yield value per country per year regardless of how many regions are included.

After accounting for both constraints, each crop-country model has approximately 14 to 16 *distinct annual yield observations* in the training period. The variation within this range is caused by differences in data availability across crop-country combinations and by the exclusion of incomplete early years. The exact number of distinct annual yield observations per crop-country combination is reported in Appendix 10.1. This small number of distinct target observations is the main limitation for yield forecasting in this thesis, especially for flexible machine learning models such as RF and XGBoost.

Yield forecasting implication. Yield is evaluated as an exploratory supply-side signal, not as a direct operational forecast for the contract pricing process. **[...Confidential information removed...]**

The same point forecasting models were applied to yield as for price where feasible, and results are reported in Section 7.4. Given the sample size, distributional forecasting is not feasible for yield, because prediction intervals estimated from so few annual outcomes would be unstable and difficult to evaluate reliably. Yield is therefore excluded from the prediction interval evaluation. A more reliable yield-based risk signal would require more granular and forward-looking production information than is available in the current dataset. This limitation is discussed further in Section 8.2.

6 Evaluation Setup

This chapter defines the metrics used to evaluate the monthly price forecasts, annual yield forecasts, price prediction interval estimates, and the validation procedure used to connect external forecasts to Picnic's internal buying prices. This study covers nine crop-country groups with different price levels and volatility patterns, so a metric that is useful within one group may not be suitable for comparisons across groups. In addition, prediction interval metrics capture different aspects of interval quality and should therefore be interpreted jointly. The evaluation setup combines multiple metrics instead of selecting one single accuracy measure, because the buying decision depends not only on average forecast accuracy, but also on whether large price deviations and uncertainty are captured in a useful way.

6.1 Point Forecast Metrics

For point forecasts, multiple error metrics are reported because each metric captures a different aspect of forecast performance. This is relevant because the crop-country combinations differ strongly in price level and volatility. An error expressed in EUR per 100 kilograms, which is the unit of the monthly market price target, is useful for understanding model performance within one crop-country group, but it is less suitable for comparing groups with different price scales. Therefore, percentage-based and normalized metrics are also included.

For each evaluated observation i, y_i denotes the realized value, $\hat{y}_i$ denotes the forecast, and n denotes the number of evaluated observations. The primary point forecast metric is the weighted absolute percentage error (WAPE), defined as

$$\text{WAPE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i|} \times 100\%. \tag{6.1}$$

WAPE measures the total absolute forecast error relative to the total realized target level across the evaluated test observations. This makes the metric useful for comparing forecast performance across crop-country groups with different price levels, while still keeping the interpretation close to the price level faced in the contract-versus-spot buying decision. Since it summarizes aggregate absolute error over the evaluated observations, WAPE is treated as the main point forecast metric in the results in Section 7.1.

Table 6.1: Additional point forecast metrics reported in the results. Here, y_i denotes the realized value, $\hat{y}_i$ the forecast, n the number of evaluated observations, and $\bar{y}$ the mean realized value for the crop-country group.

Metric	Formula	Interpretation
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Penalizes large errors quadratically and is expressed in the original unit of the target. Relevant for identifying large misses, such as price peaks or dips that could affect the contract-versus-spot buying decision.
NRMSE	$\frac{\text{RMSE}}{\bar{y}} \times 100\%$	Normalizes RMSE by the mean realized target value, making it more suitable for cross-group comparison. Because it is based on RMSE, it remains sensitive to a small number of extreme observations.
MAE	$\frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$	Measures average absolute error in the original unit. Comparing MAE with RMSE helps indicate whether errors are driven by a few large misses or are more evenly distributed.
MAPE	$\frac{1}{n}\sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert \times 100\%$	Expresses average error as a percentage of the realized value and is easy to interpret. However, Hyndman and Koehler (2006) noted that it can be inflated when realized prices are low, because the denominator becomes small.

Table 6.1 summarizes the additional point forecast metrics reported in the results. These metrics are not used as separate model-selection rules, but help interpret the structure of the forecast errors. A model with low WAPE but high RMSE may perform well on aggregate while still missing important price peaks or dips. Differences between RMSE and MAE indicate whether errors are driven by a few large misses or are more evenly distributed, while differences between WAPE and MAPE can reveal whether low-price months strongly affect the percentage-error interpretation. For this reason, WAPE is used as

the primary decision-oriented metric, while the remaining metrics are used to interpret the type of error a model makes.

6.2 Prediction Interval Metrics

For prediction interval estimates, multiple metrics are used because interval quality depends on both *calibration* and *sharpness*. Calibration measures whether realized prices fall inside the interval as often as intended, while sharpness measures whether the interval remains narrow enough to be useful in decision-making. The prediction interval coverage probability (PICP) measures how often forecasted intervals contain the realized outcome:

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[y_i \in [L_i, U_i]], \tag{6.2}$$

where $[L_i, U_i]$ is the predicted lower and upper bound for observation i. The target coverage level is 0.80, corresponding to the 80% prediction interval used throughout this study. The 80% level was chosen because it balances calibration and actionability in the Picnic buying context. A 95% interval would be more conservative, but would also tend to be wider and less useful for distinguishing between clearly favourable, uncertain, and risky contract offers.

When PICP is below 0.80, the intervals under-cover the realized prices. When it is above 0.80, the intervals may be overly conservative. However, PICP evaluates calibration only and says nothing about interval width. An interval spanning nearly the full observable price range could achieve the target coverage while still providing little decision-relevant information. For this reason, interval width is used to measure *sharpness*:

$$\text{Width} = \frac{1}{n}\sum_{i=1}^{n}(U_i - L_i). \tag{6.3}$$

Because the crop-country groups differ strongly in price level, the results primarily report relative interval width, calculated as the average interval width divided by the average realized price level of the corresponding crop-country group:

$$\text{Relative Width} = \frac{\text{Width}}{\bar{y}} \times 100\%. \tag{6.4}$$

This prevents high-price products, such as strawberries, from appearing less sharp only because their absolute price scale is larger. Width is important in the contract pricing process because it indicates how uncertain the model is about the upcoming price level. However, narrow intervals are only useful when they are also well calibrated, so width should always be interpreted together with PICP.

Furthermore, the Winkler Score is included because it combines interval width and coverage performance into a single penalty score. For a central $(1-\alpha)$ prediction interval, the score is defined as

$$W_\alpha(L, U, y) = (U - L) + \frac{2}{\alpha} \cdot \max(L - y, 0) + \frac{2}{\alpha} \cdot \max(y - U, 0), \tag{6.5}$$

where α is the nominal miscoverage rate, so $\alpha = 0.20$ for an 80% interval (Winkler, 1972). Lower values are better. When the observation falls inside the interval, the score equals the interval width and directly penalizes unnecessary width. When the observation falls outside the interval, an additional penalty proportional to the distance from the nearest interval bound is added. The score is therefore useful for comparing intervals that trade off sharpness and coverage differently.

Finally, the continuous ranked probability score (CRPS) evaluates the full predictive distribution rather than a single interval:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left[F(z) - \mathbf{1}(z \geq y)\right]^2 \, dz, \tag{6.6}$$

where F is the predicted cumulative distribution function and y is the observed outcome. CRPS is a proper scoring rule, meaning that the expected score is minimized when the forecaster reports the true predictive distribution honestly (Gneiting and Raftery, 2007). Lower values indicate better distributional accuracy. In contrast to the Winkler Score, which evaluates one specific prediction interval, CRPS evaluates the full predictive distribution implied by the interval method. **So prediction interval quality is interpreted as a combination of coverage, width, interval score, and distributional accuracy rather than through one metric alone.**

6.3 Evaluation Procedure

The evaluation uses the fixed temporal split described in Section 4.5, with training on 2004-2018, validation on 2019-2020, and testing on 2021-2023. All reported final metrics are computed on the *held-out test set* only. The validation period is used for hyperparameter tuning and, where relevant, calibration steps such as conformal adjustment. Metrics are reported per crop-country group and as averages across groups with sufficient valid predictions.

For monthly price forecasting and prediction interval estimation, metrics are computed on the rolling six-month forecast windows introduced in Section 5.1. Each end-of-December forecast contributes target months from January through June, and each end-of-June forecast contributes target months from July through December. Over the 2021–2023 test period, this creates six operational test windows. Within each window, h denotes the number of months between the *forecast origin* and the target month, with $h = 1$ for the first month and $h = 6$ for the last month. The main results focus on this six-month window because it approximates the timing of the contract pricing process. The horizon sensitivity analysis in Section 7.3 applies the same evaluation logic to alternative horizons, since product seasons and contract timings do not always map exactly onto a six-month window. Yield forecasting is evaluated separately at annual frequency, because the yield target is annual rather than monthly.

As discussed in Section 4.5, the 2021-2023 test period is interpreted as a *market-stress window*. The reported metrics therefore evaluate model performance under unusually difficult market conditions rather than as performance under a fully stable price regime. This is especially relevant for prediction interval methods, since intervals calibrated on more stable historical years may fail to cover realized prices during unusually volatile periods.

6.4 Picnic Validation Methodology

[...Confidential information removed...]

7 Results & Validation

This chapter presents the results and practical value of this research for Picnic. The chapter first evaluates the monthly price point forecasts and examines whether the main conclusions hold across crop-country groups and contract-window lengths. It also discusses the yield forecasting results separately, because yield is evaluated as an exploratory annual supply-side signal rather than as a monthly forecast. Furthermore, the chapter evaluates the prediction interval methods, focusing on the trade-off between coverage, width, and interpretability. Finally, the external market forecasting results are connected to Picnic's internal data to assess how the results can support the contract-versus-spot buying decision in practice.

7.1 Price Point Forecast Performance

The price point forecast results are evaluated in the six-month contract-window setting defined in Section 6.3, corresponding to $H = 6$. In this forecasting setup, each December a forecast is made for next year January to June and then in June for July to December and so on. As explained in Section 6.1, WAPE is used as the primary point forecast metric because it expresses total absolute forecast error relative to the realized price level, making it suitable for comparison between crop-country groups with different price scales. RMSE, NRMSE, MAE, and MAPE are also reported to assess whether the model ranking changes when large errors, scale normalization, or percentage errors are emphasized. The presented results are based on the held-out 2021-2023 test period, which is interpreted as a demanding market-stress period as discussed in Section 4.5. Strawberries × Italy is excluded from the model comparison because too few valid test observations remain after horizon alignment, consistent with the limited observation count for this group reported in Appendix Table 10.1. Including this group would make the average model ranking sensitive to only a small number of observations.

Moreover, LSTM models are not included in this result comparison because the validation results showed overly smooth predictions that did not reproduce sharp price peaks and dips sufficiently well, as discussed in Section 5.2. Since these sharp deviations are the highest margin risk periods for the contract-versus-spot buying decision, the main comparison focuses on *Naive Seasonal*, *Holt-Winters (HW)*, *Random Forest (RF)*, and *XGBoost*.

Table 7.1: Mean test-period forecast errors in the six-month contract-window setting, averaged across all evaluated crop × country groups. Strawberries × Italy is excluded due to insufficient valid test observations after horizon alignment. **Bold** indicates the lowest error.

Model	WAPE	NRMSE	RMSE	MAE	MAPE
Unit	%	%	€/100 kg	€/100 kg	%
Naive Seasonal	27.6	35.4	67.89	53.32	28.0
HW	**24.5**	**32.3**	**62.44**	**48.22**	**23.6**
RF	31.7	41.5	78.98	61.51	28.2
XGBoost	31.7	41.3	75.08	58.92	28.7

Average model ranking. Table 7.1 shows that HW achieves the lowest average error across all reported point forecast metrics. Its WAPE is 24.5%, compared with 27.6% for Naive Seasonal and 31.7% for both RF and XGBoost. The same conclusion holds under NRMSE, RMSE, MAE, and MAPE, showing that the superior HW performance is not driven by one specific metric definition. Even when large absolute errors are emphasized through RMSE, or when errors are expressed as average percentage errors through MAPE, HW remains the strongest point forecast model on average across crop-country combinations.

Historical price-based logic. [...Confidential information removed...]

The difference between these two time-series methods is especially relevant in the 2021-2023 market-stress period. Naive Seasonal copies the price from the same month in the previous year and therefore

assumes that last year's seasonal pattern is directly representative for the new contract window. This can work well when seasonal peaks repeat, but it becomes less reliable when the price level shifts across years. During a period with abnormal market conditions, such as the energy crisis in 2022, the previous year may be either too low or too high as a reference. HW can adjust more gradually because it updates the estimated level, trend, and seasonal structure over time. This likely explains why HW improves on Naive Seasonal on average in this setup, because it keeps the historical seasonal signal, but avoids relying on one previous year too mechanically. The full visual comparison between Naive Seasonal and HW is provided in Appendix Figure 10.4.

Conclusion. HW is the strongest average point forecast model in the six-month contract-window setting. The result supports the use of historical price-based forecasting in the contract pricing process, but also suggests that updating level, trend, and seasonality improves on a strict last-year seasonal reference.

Feature-based models. The feature-based machine learning models, RF and XGBoost, perform worse than both historical price-based models on average. This does not mean that the external variables are irrelevant, but it does show that they do not provide a stable point-forecast improvement in this setup. As reported in Appendix Table 10.1, the available training samples per crop-country model range from 53 to 180 monthly observations. This is a limited amount of observations for models that need to learn both seasonal structure and deviations from that structure.

This point forecasting setup asks RF and XGBoost to learn the full future price level, while much of that level is already explained by recurring seasonal price patterns. In this setting, the added feature complexity of machine learning does not translate into a consistent improvement over HW or Naive Seasonal. Appendix Figure 10.6 illustrates this comparison between HW and RF specifically: while RF follows parts of the price movement in some markets, it does not consistently capture the level and amplitude of the realized price series better than HW. The external variables may still contain useful information, but the results suggest that these are more promising for explaining deviations around a strong historical price anchor than for replacing that anchor as the point forecast itself.

Holt-Winters. The results identify HW as the best selected point-forecast anchor. Figure 7.1 illustrates what this means visually for cucumbers across the Netherlands, Spain, and Italy. The figure shows that HW captures recurring seasonal structure, especially in markets with repeated seasonal patterns such as Cucumbers × Italy. At the same time, the Cucumbers × Spain panel shows that HW still misses sharp price movements in the test period. This illustrates why the point forecast should be interpreted as an *expected price level rather than as a deterministic buying rule.*

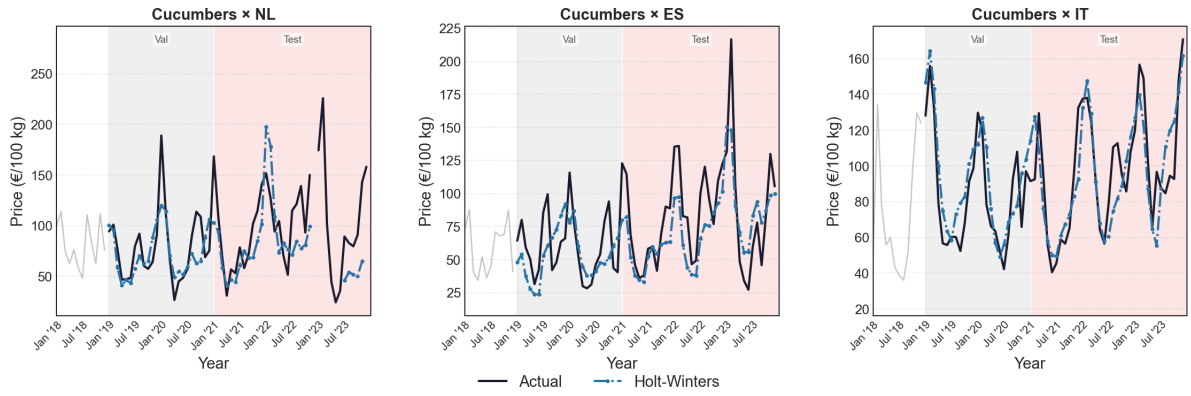

Figure 7.1: HW price forecasts for cucumbers in the six-month contract-window setting, illustrating the selected point-forecast anchor for one product across the three evaluated countries.

Contract buying decision. For the contract-versus-spot buying decision, this means that the HW point forecast should be used as *reference-level support, not as a final buying rule.* If a contract offer is clearly below the HW forecast, the model can support the interpretation that the offer may be attractive. If the offer is clearly above the forecast, it can support caution. However, the missed price movements

in Figure 7.1 show that the point forecast remains imperfect. When the contract offer is close to the forecasted price level, the remaining forecast error is too large to rely on the point estimate alone. In those cases, the relevant question is not only what the expected price is, but also how uncertain that expected price is, so how close to the expected price is to close to give certainty on the contract decision.

Overall, the average point forecast results show that market prices for the evaluated F&V products are strongly driven by historical seasonal price patterns. HW provides the strongest average point forecast because it keeps this historical price structure, while allowing level, trend, and seasonality to update over time. However, Figure 7.1 shows that even this point forecast can miss sharp price movements. The selected HW forecast is therefore used later in this chapter as the anchor for prediction interval estimation. Before moving to those intervals, the next section examines how robust this conclusion is and whether the average result hides important differences across crop-country groups.

7.2 Differences Across Crop-Country Groups

As mentioned, the average results in Table 7.1 can hide important differences between the different evaluated crop-country groups. Therefore, Figure 7.2 moves the analysis to the crop-country level by showing WAPE in the six-month contract-window setting for each evaluated group. Strawberries × Italy is again excluded because too few valid test observations remain after horizon alignment. **[...Confidential information removed...]**

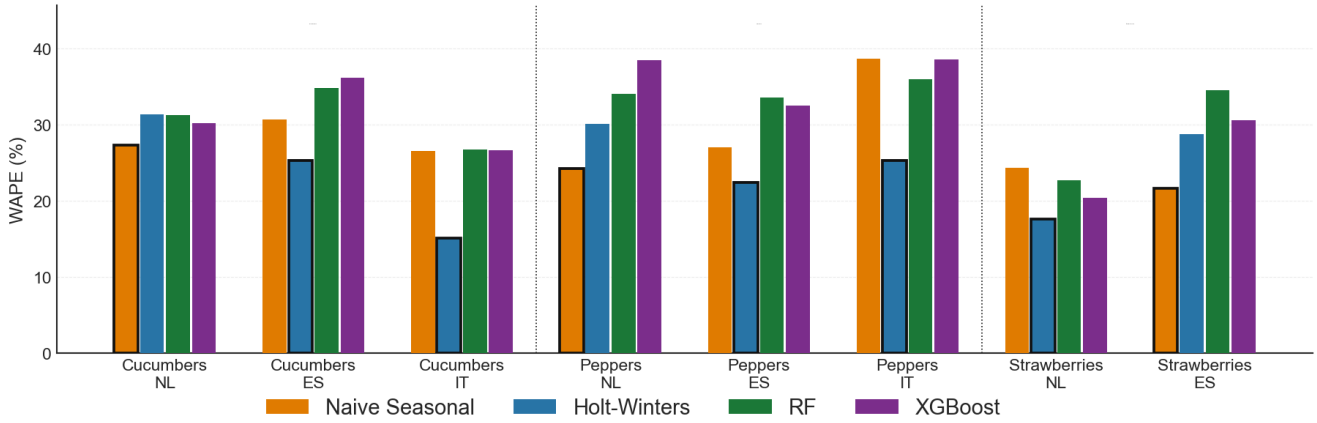

Figure 7.2: Price forecast WAPE by crop-country group in the six-month contract-window setting during the 2021-2023 test period. The black border indicates the method with the lowest WAPE within each crop-country group.

Where Holt-Winters is strongest. Figure 7.2 shows that HW is the strongest model in most, but not all, evaluated crop-country groups. Under WAPE, HW performs best for five out of eight groups. The full visual comparison in Appendix Figure 10.4 helps explain this pattern. In several of these markets, HW keeps the recurring seasonal structure but avoids copying previous year shocks too literally. This is visible most clearly for Cucumbers × IT and Strawberries × NL, where HW follows the repeated seasonal movement closely. For Peppers × ES and Peppers × IT, the realized series contains sharp spikes, but Naive Seasonal sometimes copies earlier spikes into periods where those do not repeat. HW still misses some extremes, but its smoother level and trend updates keep it closer on average across the full contract window.

Where Naive Seasonal remains competitive. Naive Seasonal performs best for the remaining three groups: Cucumbers × NL, Peppers × NL, and Strawberries × ES. These wins show that the strict last-year seasonal reference remains useful in selected markets. The visual comparison in Appendix Figure 10.4 suggests two suspected reasons for this. First, in the sparser series, especially Peppers × NL and Strawberries × ES, the available observations occur in relatively distinct seasonal episodes. This may be related to both missing price observations and the seasonal masking step described in Section 4.2. In such cases, HW has less stable information to update a smooth level, trend, and seasonal component,

while Naive Seasonal directly carries forward the most recent same-month price pattern. If that previous-year pattern is representative, this simple copy mechanism can perform surprisingly well. However, this also makes the result fragile, because the same mechanism can become misleading when the previous year was abnormal.

Cucumbers × NL shows a different pattern, where the performance gap between Naive Seasonal and HW is smaller. This group is less about sparse observations and more about the amplitude of short price movements. The realized series contains sharp peaks and drops that HW tends to smooth, while Naive Seasonal can preserve more of the previous year's peak structure when these movements repeat. The Naive Seasonal wins are therefore best interpreted as cases where direct last-year seasonality happened to be informative for the test window, but the robustness of this year-on-year copying approach needs to be interpreted with caution.

Why RF and XGBoost do not become crop-country winners. RF and XGBoost do not achieve the lowest WAPE in any evaluated crop-country group. Appendix Figure 10.6 illustrates this for RF. The RF forecasts often stay within a compressed middle range and do not reproduce the full amplitude of the realized price series. This is visible, for example, in Peppers × ES and Peppers × IT, where RF does not follow the sharp upward movements in the test period. This suggests that the feature-based models learn some average price structure, but do not capture the magnitude and timing of the most commercially relevant peaks better than HW.

This crop-country evidence supports the interpretation from Section 7.1. The limitation of RF and XGBoost is not only visible in the average table, but also in the individual markets. The models have to learn both the seasonal price level and deviations from that level from *relatively small crop-country datasets*. In the current setup, that historical price structure remains more reliable than the direct feature-based price forecasts.

Robustness across error metrics. The crop-country conclusions discussed are not specific to the WAPE metric. Appendix Figure 10.5 shows that the best model per crop-country group is consistent across WAPE, NRMSE, MAE, RMSE, and MAPE. This strengthens the result: HW remains the most frequent winner across groups and metrics, while the feature-based models do not show a stable advantage. At the same time, Naive Seasonal remains visible as a relevant benchmark because it also wins on other metrics in the groups where the previous-year seasonal reference is strongest. The additional metrics therefore do not change the main conclusion, but they clarify that the difference between HW and Naive Seasonal is market-specific rather than random metric noise.

Conclusion. The crop-country results show that HW is the strongest point-forecast anchor overall, but not in every individual market. Naive Seasonal remains useful as a benchmark in markets where the previous year's pattern repeats strongly. RF and XGBoost do not show a stable crop-country advantage, mainly because their forecasts do not consistently capture the magnitude of sharp price movements.

For Picnic, the practical conclusion is that crop-country results should be used as diagnostics rather than as a separate model-selection rules for every group. Selecting a different point forecast model for each crop-country group would make the forecasting framework less transparent and could make the results too dependent on the relatively short 2021-2023 test period. Instead, HW is retained as the selected point-forecast anchor because it performs best on average, wins most crop-country groups. **[...Confidential information removed...]** Naive Seasonal remains useful as a comparison model that shows where direct last-year seasonality is more informative. The next section validates whether the same point-forecast conclusion holds when the evaluated contract-window length changes.

7.3 Contract-Window Length Sensitivity

The previous sections focused on the six-month contract-window setting, corresponding to $H = 6$. To assess whether the point forecast conclusion depends on this window length, the same model comparison was repeated for alternative contract-window lengths. Here, H denotes the number of monthly targets included in the evaluated contract window, rather than one isolated forecast horizon. Increasing H changes the amount of uncertainty included in the evaluation. Months near the end of a contract window are

further away from the last observed price information at the forecast origin, so larger forecast errors are expected to become more likely as the window extends. However, the WAPE reported here is averaged over all months included in the window. Therefore, the relationship between H and WAPE does not have to be strictly increasing: the score *also depends both on the additional months included and on the seasonal difficulty of those months.*

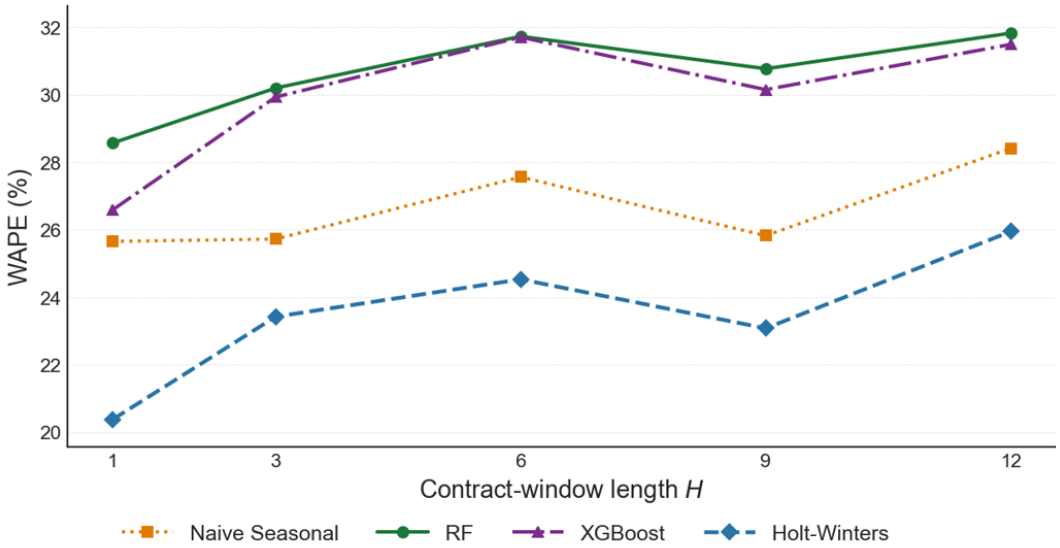

Figure 7.3: Average WAPE by contract-window length H during the 2021-2023 test period. Results are averaged across crop-country groups with sufficient valid observations at each window length. Lower values indicate better point forecast accuracy.

Figure 7.3 shows that the main model conclusion is robust to the evaluated contract-window length. HW achieves the lowest average WAPE for all tested values of H, while Naive Seasonal remains the closest alternative. RF and XGBoost remain consistently higher, indicating that changing the contract-window length does not make the feature-based models more competitive in the point forecast setting. The pattern shows that HW is not only strong at the selected six-month setting. This supports the interpretation from the previous sections that the most reliable point-forecast signal comes from historical price structure, especially when level, trend, and seasonality are updated over time.

Conclusion. Changing the contract-window length does not change the main point forecast conclusion. HW remains the most reliable point-forecast anchor across the evaluated window lengths, but meaningful forecast errors remain, especially as the evaluation includes months further away from the forecast origin.

In the contract pricing process, product seasons and contract timings do not always map exactly onto one six-month window. A model that only performs well for one specific window length would therefore be less useful operationally. The sensitivity analysis suggests that HW remains the strongest historical price-based anchor when the evaluated contract window is shortened or extended. HW therefore provides the most stable point forecast reference across the evaluated settings, but the WAPE values show that point forecasts remain imperfect. This is especially relevant toward the later months of a contract window, where the forecast is further removed from the last observed market information. The following sections therefore shift from selecting the point forecast anchor to evaluating prediction interval methods that quantify uncertainty around that anchor.

7.4 Yield Forecasting Results

Table 7.2 reports the average test-period errors for the yield forecasting models. Compared with the price forecasting results in Table 7.1, the yield WAPE values are much lower. Naive Seasonal achieves the lowest overall error on all reported metrics, with a WAPE of 4.5%. HW, RF, and XGBoost all perform worse on average, although their percentage error levels remain lower compared with the monthly price forecasts.

However, this lower error level should be interpreted carefully. It does not mean that yield forecasting is more useful for Picnic's contract-versus-spot buying decision than price forecasting. Rather, annual national yield is a much smoother target than monthly market prices. Yield is observed once per year and aggregated at country level, while prices are evaluated monthly and are more directly exposed to short-term market volatility. The lower WAPE therefore mainly reflects the smoother structure of the yield target, *not necessarily a stronger operational forecasting signal.*

Table 7.2: Mean yield forecast errors during the 2021-2023 test period, averaged across all evaluated crop × country groups. RF and XGBoost are trained on monthly observations from 2004-2018 and evaluated after aggregation to annual yield predictions. **Bold** indicates the lowest error.

Model	WAPE	NRMSE	RMSE	MAE	MAPE
Unit	%	%	tonnes/ha	tonnes/ha	%
Naive Seasonal	**4.5**	**5.4**	**8.52**	**6.91**	**4.5**
HW	6.9	7.5	12.97	12.29	6.9
RF	7.8	8.2	12.48	11.94	7.7
XGBoost	9.3	10.1	15.35	14.35	9.2

Model interpretation. The strong performance of Naive Seasonal suggests that annual yield is largely driven by persistent crop-country level differences and recent historical levels. This is also visible in Appendix Figure 10.7, where the yield series are generally smoother than the monthly price series. Consequently, this helps explain why the simple benchmark performs well despite its limited modelling structure.

HW performs worse than Naive Seasonal on average. This is consistent with the structure of the yield target. Unlike monthly prices, yield is reported annually and therefore does not contain within-year seasonal variation and trends that a this time-series model can exploit. The model can only smooth a short annual series, which limits the value of estimating separate level, trend, and seasonal components. Both machine learning methods also fail to improve upon Naive Seasonal. Although RF and XGBoost use monthly regional features, the target variable remains annual. This created a high-dimensional learning problem with only a small number of independent yield outcomes, making it difficult to learn a stable relationship between the engineered features and future national yield.

The group-level WAPE comparison in Appendix Figure 10.8 shows that Naive Seasonal is the strongest average model, but not the best model in every crop-country group. RF performs relatively well in selected groups, while HW is competitive in some others. However, these group-level differences should be not be overinterpreted, because the test period contains only three annual target values per group. One unusual harvest year can strongly affect which model appears best. The group-level results are therefore useful as diagnostics, but not as a basis for selecting separate yield models per crop-country group.

Beyond these mentioned statistical limitations, the yield forecast also has an operational limitation. The models predict annual national yield, not yield for a specific sourcing season, harvest window, supplier region, or quality class. For Picnic's contract-versus-spot buying decision, this is too coarse. Even an accurate annual national yield forecast would not directly indicate whether supply pressure is expected during the months in which contracts are negotiated or fulfilled. What Picnic would need instead is a signal that indicates supply risk during the relevant buying window. The current yield target does not provide the *timing or granularity* needed for this decision.

Conclusion. The yield forecasting target is too coarse, and the number of independent test observations too small to effectively support Picnic's contract-versus-spot buying decision.

Overall, the yield results do not justify further yield prediction interval estimation in this research. Next to all other discussed yield forecasting limitations resulted in too unstable model rankings to justify selecting one yield approach. Yield remains relevant as contextual information for understanding supply conditions, but is *not developed further as a standalone operational forecasting target.* The remainder of

the results focuses on price uncertainty, where the monthly price target structure is better aligned with the contract-versus-spot buying decision.

7.5 Prediction Interval Performance

The point forecast results identified that HW was the strongest point-forecast, but also showed that the remaining forecast error is too large for the contract-versus-spot buying decision to rely on the point estimate alone. This section therefore evaluates whether prediction intervals around the HW forecast can quantify the remaining uncertainty in a useful way. All interval methods are evaluated in the six-month contract-window setting, corresponding to $H = 6$. The target is an 80% prediction interval, meaning that realized prices should fall inside the interval in approximately 80% of the evaluated cases. The interval metrics introduced in Section 6.2 are interpreted jointly in the research context, because coverage is only useful if the resulting interval remains narrow, transparent, and actionable enough to support the buying decision.

Average interval performance. Table 7.3 reports the average prediction interval performance across the evaluated crop-country groups during the 2021-2023 test period. Relative width is included to make interval width comparable across products with different price levels. None of the methods reaches the nominal 0.80 coverage target on average, which shows that realized price uncertainty in the test period remains difficult to capture, even after anchoring the intervals around the HW point forecast. However, this could be caused by the demanding test period in this research.

Table 7.3: Prediction interval performance in the six-month ahead contract-window forecast, averaged across all evaluated crop × country groups during the 2021-2023 test period. Lower values are better for relative width, Winkler Score, and CRPS. For PICP, the coverage target is 0.80. **Bold** indicates the best value per metric.

Method	PICP	Rel. Width	Winkler Score	CRPS
Unit	Proportion	%	€/100 kg	€/100 kg
Fixed Empirical	0.683	57.9	**230.7**	**35.1**
Residual QRF	0.456	**47.6**	280.6	43.2
QRF+CQR	**0.712**	73.1	252.0	43.2
Simple QRF	0.607	49.2	284.4	41.8

Table 7.3 shows a *coverage-width trade-off* rather than one method dominating all metrics. QRF+CQR comes closest to the 0.80 coverage target with a PICP of 0.712, while Residual QRF reaches only 0.456. Since QRF+CQR builds directly on Residual QRF, this difference shows that the conformal calibration step increases coverage as intended. However, it does so mainly by widening the interval substantially.

This additional width is the main reason why QRF+CQR is not the strongest overall interval method in all metrics. Although it has the highest PICP, Fixed Empirical achieves the lowest Winkler Score and CRPS. The Winkler Score is especially relevant here because it penalizes both unnecessarily wide intervals and under-coverage. The lower Winkler Score therefore indicates that Fixed Empirical provides the better average balance between coverage and sharpness. Fixed empirical's lower CRPS supports the same conclusion from a distributional perspective, as it suggests that the predictive distribution is closer to the realized prices overall than in other methods explored.

Residual QRF and Simple QRF provide useful comparisons, but neither is reliable enough as an operational interval method. Their intervals are sharper, but the coverage is too low, indicating that they are too confident in the uncertainty estimates. Residual QRF shows that feature-conditioned residual intervals around HW are too narrow without calibration. On the other hand, Simple QRF shows that modelling the future price distribution directly does not solve the interval problem either, as it results in the highest Winkler Score. This supports the choice to keep *HW as the central point-forecast anchor* and to model uncertainty around its residuals.

Conclusion. Fixed Empirical provides the strongest overall prediction interval result in the current setup. QRF+CQR achieves the highest coverage, but its wider intervals do not improve the overall scoring metrics. The other tested QRF variants under-cover too strongly for operational use.

Temporal robustness of interval scores. The average results in Table 7.3 show that Fixed Empirical performs best on the combined scoring metrics, but they do not show whether this result is stable across the test period. Figure 7.4 therefore reports the Winkler Score and CRPS separately for each test year.

Figure 7.4: Average Winkler Score and CRPS by test year for 80% prediction intervals in the six-month contract-window setting. Lower values indicate better interval performance. Values are averaged across crop-country groups with sufficient valid observations.

The yearly scores confirm that the main conclusion is *not driven by one specific test year*. Fixed Empirical achieves the lowest Winkler Score and CRPS in each year, which supports its robustness in the current setup. The difference is clearest in 2021 and 2022, where Fixed Empirical combines relatively moderate interval width with fewer severe misses than the sharper QRF variants. In 2023, the scores increase for all methods, indicating that the interval forecasting task becomes more difficult toward the end of the test period. This pattern is consistent with the broader price-series in Appendix Figure 10.2, where several crop-country groups show more irregular price movements during the later test years.

The yearly pattern also helps explain why QRF+CQR is not selected as the best performing model, despite having the highest average coverage. The conformal calibration step makes the interval safer than Residual QRF, but the additional width does not lead to better Winkler or CRPS performance in any test year. Appendix Figure 10.9 shows the underlying coverage-width pattern: QRF+CQR has the widest intervals in each year, while Fixed Empirical even reaches higher coverage in 2021. The yearly results therefore reinforce the same conclusion as the average table that QRF+CQR improves coverage, but Fixed Empirical gives stronger overall interval performance.

Visual Peppers×ES example. The previously discussed average results explain the method ranking, but do not show how the intervals behave around actual price movements. Figure 7.5 therefore zooms in on Peppers × ES, a volatile series with several sharp price movements during the test period. This example visualizes the differences between the interval methods. Fixed Empirical produces a relatively stable band around the HW forecast. The interval is transparent and easy to interpret as a historical residual range, but it is *stable rather than expressive* as its width does not change much over time. As a result, the method does not clearly signal when the coming period is expected to be more or less uncertain.

QRF+CQR shows the opposite behaviour. Its interval is much wider, but also more expressive. The band changes more over time and is not always centered around the HW forecast. This means that the method can indicate not only that a period is more uncertain, but also whether the conditional residual distribution suggests more upside or downside risk around the HW point forecast. In this example, QRF+CQR also has the highest coverage of 0.81, just above the nominal target. However, it also raises the interval width compared to Fixed Empirical. The method therefore provides a *richer uncertainty signal* for Spanish Peppers, but at the cost of a broader range.

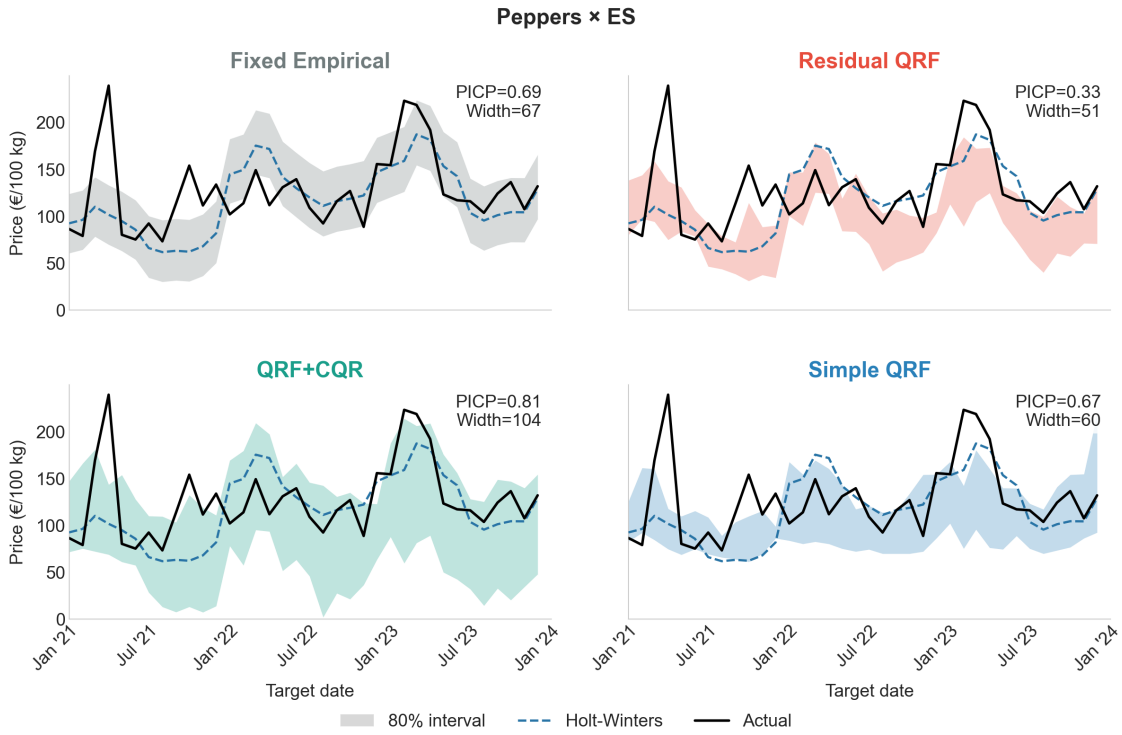

Figure 7.5: Prediction intervals for Peppers × ES during the 2021-2023 test period in the six-month contract-window setting. The panels compare Fixed Empirical, Residual QRF, QRF+CQR, and Simple QRF around the HW point forecast.

Figure 7.5 also shows what the conformal calibration step adds when comparing Residual QRF with QRF+CQR. Residual QRF uses feature-conditioned residuals around HW, but without calibration the interval remains too narrow for this volatile price series. Its coverage is only 0.33, far below the 0.80 target. QRF+CQR corrects this overconfidence by widening the residual interval after calibration. The improvement in coverage is therefore clear, but the figure also shows that this improvement mainly comes from adding width rather than from capturing every individual price spike of Spanish Peppers.

Lastly, Simple QRF provides a useful comparison because it does not build the interval around HW residuals, but models the future price distribution directly. In this example, Simple QRF performs reasonably visually, with a narrower interval than Fixed Empirical and coverage close to Fixed Empirical. However, the crop-country diagnostics in Appendix Figure 10.10 show that Simple QRF does not improve the coverage-width trade-off relative to the HW-based interval methods. This supports the choice to keep HW as the central point-forecast anchor and to model uncertainty around its residuals. The full 3-by-3 fan-chart results for all interval methods can be found in Appendix Figures 10.13-10.16.

Crop-country specific performance. The crop-country results add nuance to the average results in Table 7.3. Appendix Figure 10.10 gives the full method comparison by crop-country group. This figure shows two patterns. **First**, QRF+CQR improves coverage relative to Residual QRF in most crop-country groups, which confirms that the conformal calibration step generally corrects part of the under-coverage of the uncalibrated residual intervals. However, this correction is accompanied by an increase in interval width. The pattern is strongest in the pepper markets, where Residual QRF is visibly too narrow for the realized test-period movements.

Second, Appendix Figure 10.11 then isolates the most relevant operational comparison: Fixed Empirical versus QRF+CQR. This comparison explains why higher average coverage alone is not enough to select QRF+CQR. In some groups, such as Peppers × ES and Cucumbers × IT, QRF+CQR improves coverage relative to Fixed Empirical. In several other groups, however, the gain in coverage is small or absent while the interval becomes substantially wider. Cucumbers × ES is a clear example as QRF+CQR does not improve coverage compared with Fixed Empirical, but it does require a wider interval. Peppers × NL shows a similar pattern. These cases explain why Fixed Empirical can win on Winkler Score and CRPS even though QRF+CQR reaches the highest average PICP.

The strawberry groups should be interpreted more cautiously. Strawberries × NL has a strong seasonal

pattern, so Fixed Empirical remains relatively competitive without needing a highly adaptive interval. Strawberries × ES is more fragile because fewer test observations are available, making the estimated interval performance sensitive to a small number of misses. For these sparse groups, the QRF-based methods are less stable.

Overall, the crop-country specific results do not overturn the average conclusions made before, but do explain where these come from. QRF+CQR is most valuable where Residual QRF is often overconfident and where extra width improves coverage. Fixed Empirical is strongest where the HW forecast already captures the broad seasonal structure and where a historical residual range is representative enough of future price uncertainty. Since this pattern appears more consistently across the evaluated groups, and because the method remains easier to explain, Fixed Empirical remains the recommended prediction interval method for Picnic's contract pricing process.

Contract-window length sensitivity. The final robustness check evaluates whether the preferred interval method changes when the contract-window length H is shortened or extended. Figure 7.6 reports the Winkler Score across alternative window lengths. This score is more useful than coverage alone for this comparison because it summarizes the full interval trade-off: a method has to cover enough realized prices without becoming unnecessarily wide.

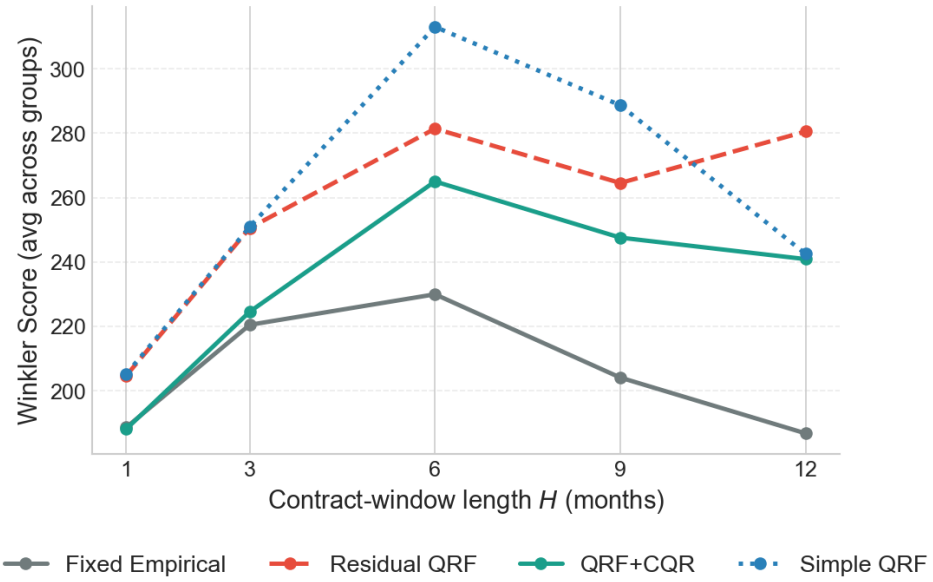

Figure 7.6: Average Winkler Score by contract-window length H during the 2021-2023 test period. Lower values indicate better interval performance.

Figure 7.6 illustrates that Fixed Empirical remains the preferred method across the evaluated window lengths. This supports the main conclusion, because the selected interval method is not only strongest at the main six-month setting but also remains competitive when the evaluated contract window is shortened or extended. The Winkler Score increases from the shortest windows toward the six-month window for most methods, including Fixed Empirical. This was expected because a longer contract window includes target months that are further removed from the forecast origin and therefore further from the last observed market information. These later months are harder to place inside a narrow interval, especially when the window crosses seasonal transitions or contains months with more irregular price movements, like Dutch winter for example. At the same time, the Wrinkler Score of the methods is not strictly increasing. This is because H is a contract-window length, not one isolated forecast horizon. A larger H means that more target months are averaged into the score. When the window becomes longer, the score is affected not only by distance from the forecast origin, but also by the mix of seasonal months included in the evaluation. The lower score again at $H = 12$ may therefore reflect that a full-year window includes a more complete seasonal cycle, so difficult months are averaged together with more regular months. It should therefore *not* be interpreted as evidence that twelve-month forecasts are actually easier, but just as a comparison between methods. QRF+CQR performs close to Fixed Empirical at the shortest window length, suggesting that the adaptive residual interval can be useful for near-term uncertainty.

Fixed Empirical is more stable across window lengths, which supports its selection as the best performing interval method in this thesis.

Operational interpretation. [...Confidential information removed...]

Conclusion. The selected interval method is Fixed Empirical around the HW point forecast. It is not the highest-coverage method, but it provides the best overall trade-off between interval accuracy, width, transparency, and usability in the contract pricing process.

7.6 Picnic Validation

[...Confidential information removed...]

8 Conclusion & Discussion

8.1 Conclusion

This thesis investigated to what extend a multi-output forecasting framework, combining historical market prices and external agricultural features, can support risk-aware contract-versus-spot buying decisions for fruit and vegetables at Picnic. **[...Confidential information removed...]** To study this decision context, external market prices, production statistics, weather observations, satellite vegetation indices, and soil characteristics were combined into a harmonised dataset for cucumbers, strawberries, and bell peppers across the Netherlands, Spain, and Italy. The final modelling dataset covers 2004-2023, with 2021-2023 used as the held-out test period. This test period is important for the interpretation of the results, because it represents a market-stress window rather than a stable evaluation period.

The results show that the full framework does not yet provide a general decision-support tool for Picnic's contract pricing process. The feature-based models also did not consistently improve price forecasts, and the annual yield forecasts from the multi-output framework were too coarse to become a useful operational supply-side signal. The strongest and most useful part of the framework was therefore narrower than the setup tested in this research: *a price-focused forecasting approach based on historical market prices to support Picnic's contract pricing process.*

More specifically, the results support using historical price-based forecasting as a quantitative risk layer in the buying process for specific products. HW can provide an expected future price level, while Fixed Empirical prediction intervals is used to show the historical uncertainty around that forecast. This does not replace the buyer's judgement or the existing contract pricing process. Instead, it formalises part of the reasoning that is already present in the process: comparing contract offers with expected seasonal market prices and assessing whether the offer lies within a plausible expected price range. The resulting forecasting framework approach is therefore most useful as structured risk and negotiation support for selected product groups, especially when the external market price can be linked reliably to Picnic's internal buying price.

Data harmonisation. *Multi-source agricultural and market data can be harmonized into one forecasting dataset, but only by preserving the original granularity limits of each source.* Weather, satellite, soil, production, and market data were combined into a monthly crop-country forecasting dataset by aligning the sources to a common NUTS2-based modelling structure. Weather and satellite variables could be aggregated to monthly regional observations, while market prices and yield targets had to remain national-level outcomes because more detailed external target data were not consistently available. The data was split for modelling per crop-country group, because the same product from different countries represents a different buying context, with different seasonal calendars, production systems, and price dynamics.

This harmonisation made it possible to evaluate price and yield forecasts across crop-country combinations, but only after making explicit choices about spatial and temporal aggregation. The key methodological choice was not to force all sources into the same level of detail when this would change the meaning of the data. Annual national yield was therefore not disaggregated into monthly or regional yield observations, because that would imply a level of precision that the original data do not contain. Similarly, monthly national price series were assigned to the relevant crop-country models, but they were not treated as regional price observations. As a result, regional variation enters the research through explanatory variables such as weather, satellite, and soil conditions, while the forecast targets remain national-level price and yield outcomes.

This creates a *robust external forecasting dataset*, but one that should be interpreted carefully due to some misalignment in granularity between target variables and explanatory variables. It is suitable for testing whether external agricultural and market data contain forecasting signal across crop-country combinations. On the other hand, this mismatch remains an important limitation when interpreting the model results, especially for the yield target.

Point price forecasting. *Feature-based machine learning did not improve the researched price forecasts over historical price-based time-series models.* The strongest point-forecast signal came from the price

series itself. In the six-month contract-window setting, HW achieved the lowest average error across all reported point forecast metrics, with a WAPE of 24.5%. This was lower than the Naive Seasonal benchmark, with a WAPE of 27.6%, and clearly lower than RF and XGBoost, that both reached 31.7% WAPE on average. This same pattern persisted when changing the contract-window length as the ranking of the models stayed consistent with the 6 month ahead general outcome.

This result shows that the seasonal time-series structure is more informative than the external feature set in the current setup. HW is especially useful because it keeps the historical seasonal logic that is already being used in the buying process, but updates the level, trend, and seasonal component over time. It is therefore less rigid than Naive Seasonal, which simply copies the price from the same month last year. The feature-based models, in contrast, had to learn from many input variables but relatively few independent monthly price targets per crop-country group. Once the data were split into separate crop-country models, the effective target variable history became limited. Consequently, RF and XGBoost did not learn deviations from seasonality reliably enough to outperform the historical price-based models in this data setting. However, the split-out crop-country results add nuance to the time-series results, because the size of the HW advantage differs across products and countries, but these differences do not change the broader finding that the strongest price signal comes from historical seasonal price behaviour. So either from HW or Naive Seasonal performed best, but the machine learning models did not attain the lowest WAPE for any crop-country group. **[...Confidential information removed...]**

Yield forecasting. *The annual yield forecast resulted in relatively low percentage errors, but does not provide a useful supply-side signal for Picnic's contract-versus-spot buying decision.* The yield results show a clear difference between statistical prediction performance and operational usefulness. Naive Seasonal achieved the lowest average yield error, with a WAPE of 4.5%, which is lower than previously seen in the monthly price forecasts. However, this does not mean that annual yield forecasting is more useful for Picnic's contract-versus-spot buying process.

The main limitation is the structure of the yield variable. Yield is reported annually and nationally per product,**[...Confidential information removed...]**. A national annual yield forecast does not show whether supply pressure will occur during the months in which Picnic evaluates a contract offer. It also does not indicate whether that pressure affects the specific supplier regions or product specifications that Picnic uses. The low percentage error therefore mainly reflects the smoother year-over-year structure of the annual yield target that may hide seasonal variations. As a result, yield is best interpreted as contextual supply-side information. In this research, it is not reliable enough as a standalone operational forecasting target. More granular regional, monthly, or supplier-level supply data could make this component more valuable in future work.

Prediction intervals. *Prediction intervals add value by turning the HW point forecast into a price-risk range, but the simplest historical-error interval was found to be the most useful method.* Prediction intervals address the remaining point forecasting error by showing a range of plausible future prices around the expected market price level. Fixed Empirical was selected as the strongest interval method in this thesis. The split-out interval results show that this choice is not equally strong for every crop-country group, but Fixed Empirical was the most useful overall method because it remained transparent and stable. Fixed empirical did not achieve the highest average coverage across groups, but still resulted in the best overall trade-off between interpretability, interval width, Winkler Score, and CRPS. QRF+CQR achieved higher coverage and remains a useful adaptive benchmark, because it can adjust interval width depending on within- period uncertainty using feature-conditioned residual risk. However, the improved coverage of QRF+CQR came mainly through wider intervals, making this less attractive as the default operational method in the current data setting. Therefore, the interval results therefore reinforce the broader conclusion from the point forecasts: *the strongest signal comes from historical price behaviour where HW provides the expected price anchor, while Fixed Empirical uses the historical error distribution around that anchor to create a transparent risk range.*

For the contract-versus-spot buying decision, this uncertainty layer is limitedly useful as it can help identify contract offers that are clearly outside the historically plausible price range. However, the intervals often still have a relative width that is too high to support precise price negotiation when a contract offer lies close to the expected price level. The interval will therefore support buyer judgement by making risk visible, not replace the buyer's commercial assessment of the contract offer.

Picnic validation. [...Confidential information removed...]

Overall conclusion. *The full multi-output framework does not provide a general decision-support tool for Picnic, but the research does identify a narrower approach that can offer support for selected products.* In the current data setting, the useful direction is not feature-based complexity but a historical price-based forecast with a transparent uncertainty layer, applied only where the external market price can be reliably mapped to Picnic's internal buying price. The split-out crop-country results reinforce this point: the framework should be applied selectively, because forecast performance and internal price matching differ strongly by product and country. [...Confidential information removed...]

The contribution of this thesis is therefore not that a complex multi-output model outperforms the current buying logic. The contribution is that it shows where complexity does and does not add value in a fresh fruit and vegetable buying context. It also shows that forecast accuracy alone is not enough in this setting, because the forecast must be interpretable, uncertainty-aware, and translatable to Picnic's contract pricing process. With the current external data depth, historical price-based forecasting remains the strongest foundation. Prediction intervals can add risk context around that foundation. Yield and external agricultural features remain promising for future development, but these require more granular and decision-aligned data before becoming useful operational signals.

8.2 Discussion

This discussion reflects on the main limitations and operational implications of the forecasting approach evaluated in this thesis. The conclusion showed that historical price-based forecasting with a transparent price uncertainty range can support selected contract-versus-spot buying decisions, but is not yet ready to support every crop-country decision in Picnic's contract pricing process. This section discusses why that boundary appears, which data and modelling choices affect the interpretation of the results, and what would be needed to use the approach more broadly in Picnic's buying process.

Historical price dependence under market stress. *Historical price-based time-series models are useful because these capture the strongest signal in the data, but remain backward-looking.* The point forecast results show that most of the usable price signal comes from the historical price series itself. HW and Naive Seasonal both build on this logic, although HW updates level, trend, and seasonality instead of copying the previous year observations directly. This makes these methods strong and interpretable in Picnic's contract pricing process.

However, this comes with the limitation that these models can only extend patterns that are already visible in the price history. They cannot anticipate a new storm, heat wave, energy-price shock, or sudden supply shortage before such information becomes visible in market prices. If the recent price history was stable, the forecast may underestimate the risk of a new shock. But if prices in recent years were unusually volatile, the forecast may continue to reflect part of that volatility even when the market starts to normalise. This thesis specifically uses 2021-2023 as the test period, which includes years with abnormal market pressure following COVID-19, energy-price increases, and wider agricultural input-cost disruptions (Kornher et al., 2024; Statistics Netherlands (CBS) and Wageningen Economic Research, 2022). The models were therefore tested under difficult conditions, which presents forecasting periods in which risk support is most relevant for buying decisions. However, future robustness checks should also test on additional stable years to see if that would impact HW performance relative to machine learning.

External features and forward-looking risk. *The external agricultural features used in this thesis mainly describe observed conditions, not future shocks.* The limited contribution of feature-based machine learning in this thesis should not be seen as evidence that weather, satellite, soil, or production information is irrelevant for fruit and vegetable markets. Rather, the current feature set was not sufficiently aligned with the risk that matters at the contract decision moment. Most environmental variables describe conditions observed up to the forecast origin, such as accumulated temperature, precipitation, frost days, and vegetation indicators. They can describe the growing conditions that have already occurred, but not directly the extreme weather, supply, or cost shocks that may occur between the forecast origin and the target months. In Picnic's setting, the future risk is not limited to the contract period itself, but also includes the remaining growing and supply period before and during the sourcing window. As a result,

the machine learning models had to explain price deviations without observing part of the future risk that may cause those deviations.

This timing mismatch helps explain why RF and XGBoost did not add enough predictive value in the current setup. A useful extension would therefore be to include more forward-looking inputs, such as seasonal weather forecasts and expected energy costs. However, adding these variables would only be valuable if enough historical examples are available to learn how such signals translate into price deviations for each crop-country group. The hypothesis that machine learning can identify deviations from historical price and yield patterns therefore remains plausible, but requires clearer forward-looking inputs and longer time-series data than were available in this thesis.

Price data completeness and reporting lag. *The forecasts depend strongly on complete and up-to-date historical data on target variables, which makes missing prices and external data reporting lags important limitations.* Missing observations are not equally problematic for every model type. Feature-based tree models, like RF and XGBoost, can deal relatively flexibly with missing explanatory values. However, historical time-series models were found to be the strongest models and depend directly on the observed target price series. If this time-series is incomplete, HW has less information to estimate level, trend, and seasonality, and Fixed Empirical has fewer forecast residuals from which to construct the uncertainty range.

During preprocessing, missing external price observations were imputed where possible by using national statistical sources and scaling them to the Eurostat price series during overlap periods. This was a cautious solution, because the imputed values are still based on observed prices from another source rather than on simple interpolation. Nevertheless, the **imputation step** remains a limitation. If the national source does not fully capture the same market movements or product definitions as the Eurostat series, the resulting price history may contain bias or misrepresent part of the true price volatility. A useful robustness extension would therefore be to compare the main results with less-imputed price variants, both for point forecast accuracy and for prediction interval performance.

A second limitation of relying on externally available data is **reporting lag**. Recent price information turned out to be one of the most important inputs for the price forecasts, while the yield forecasts also rely heavily on the most recent annual yield level. If external price or yield data only become available with a delay of a few months, the real forecast used in the contract pricing process would have to rely on older information than assumed in the modelling setup. This effectively increases the distance between the buying moment and the latest observed market information, which is likely to reduce forecast accuracy. This limitation is difficult to remove when relying only on public external data, because the most recent market information is not always available at the moment the buyer needs to evaluate a contract offer.

Regional and seasonal masking. *The dataset was filtered to improve agricultural relevance, but the seasonal mask also reduced the available price history observations.* The regional mask is relatively easy to justify, because it prevents the model from linking weather conditions in non-producing regions to national market prices. A Dutch, Spanish, or Italian region without relevant production should not contribute weather information to a crop-country forecast. In that sense, the regional mask makes the external feature set more consistent with the production reality of the product, but does not affect the national target variables.

The seasonal mask improves agricultural relevance by focusing on production months, but also removes price observations from the target series in other months. This poses a limitation as the seasonal mask not only results in a reduction in explanatory data, but also changes the price history on which the models learn. For HW, fewer price observations mean less information to estimate on. Similarly, for Fixed Empirical, fewer price observations lead to fewer residuals for estimating the historical uncertainty range. This is especially relevant for less established sourcing countries, where the number of usable price observations is already limited.

Future research could therefore test softer seasonal masking strategies. For example, the full price series could be retained while environmental variables from inactive production months are masked or explicitly flagged. This would preserve the target variable history while still preventing irrelevant production signals from dominating the feature set. Such a robustness check would help determine whether the current seasonal mask improved the modelling setup overall, or whether it made some crop-country groups too sparse. However, because the best-selected forecasting approach relies on historical prices

rather than external features in the current setup, this makes the other explanatory variables and masking redundant when operationalizing this research.

Yield granularity and supply-side forecasting. *Annual national yield is too coarse for Picnic's buying decision, but supply-side forecasting remains relevant.* Annual national yield forecasting was included to test whether external agricultural data could provide a supply-side signal next to price forecasting. The results show that annual yield can be forecast with relatively low percentage error, but that it is not operationally useful for Picnic's contract pricing process because the target variable is too coarse. Annual national yield predictions do not show whether supply pressure will occur in March, June, or October, nor whether that pressure affects the supplier regions or product specifications Picnic actually uses. The low percentage forecasting errors mainly reflect that annual national yield is a smoother target than monthly market prices.

[...Confidential information removed...]

Fixed Empirical as a risk layer. *Fixed Empirical is useful as a transparent price-risk flag, but it does not model changing risk within the contract window.* Fixed Empirical was selected because it provides a transparent and stable uncertainty range around the HW forecast. This is attractive in the contract pricing process, because this is an interpretable interval made up of a historical error range around the selected expected price level. However, the same simplicity also limits the method's support to the buyer in the contract pricing process. Since the interval is based on historical residual quantiles, its width does not adapt much within a forecast window. It shows how uncertain HW has been historically, but it does not clearly indicate whether one upcoming period is expected to be more volatile than another.

This limitation changes the use of the resulting interval, which is then most useful when a contract offer lies clearly outside the historically plausible price range. In that case, the forecasted price range can flag that the offer deserves closer attention. When an offer lies close to the expected price level, the Fixed Empirical range is less decisive, because it does not show whether the coming period is unusually risky or relatively stable.

Adaptive interval methods such as QRF+CQR therefore remain relevant for future work, but not as the current setup. QRF+CQR attempts to vary the interval width using feature-conditioned residual risk, but in this thesis that adaptivity came with additional interval width and complexity without improving the overall interval scores compared to Fixed Empirical. With deeper price history training and more forward-looking risk indicators, adaptive intervals could be tested again. Until then, Fixed Empirical is best interpreted as a transparent first risk layer, not as a complete model of changing market volatility.

Contract-window timing. **[...Confidential information removed...]**

Bibliography

Avinash, Chebrolu, Suresh Nalla, Deepak Gupta, Sanjeev Kumar, and Manoj Kumar (2024). "Hidden Markov guided deep learning models for forecasting highly volatile agricultural commodity prices". In: *Expert Systems with Applications* 238, p. 122257. DOI: 10.1016/j.eswa.2023.122257.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Chatfield, Chris (1993). "Calculating Interval Forecasts". In: *Journal of Business & Economic Statistics* 11.2, pp. 121–135. DOI: 10.1080/07350015.1993.10509938.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794. DOI: 10.1145/2939672.2939785.

Cheng, Junhui, Jianfei Sun, Kunshan Yao, Min Xu, and Yan Cao (2022). "A variable selection method based on mutual information and variance inflation factor". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 268, p. 120652. DOI: 10.1016/j.saa.2021.120652.

Copernicus Land Monitoring Service (2026). *Copernicus Global Land Service*. Available at https://land.copernicus.eu.

European Commission, Joint Research Centre (2026). *European Soil Database and Soil Properties*. Available at https://esdac.jrc.ec.europa.eu/resource-type/european-soil-database-soil-properties.

European Parliament and Council of the European Union (2022). *Regulation (EU) 2022/2379 on statistics on agricultural input and output*. Available at EUR-Lex: https://eur-lex.europa.eu/eli/reg/2022/2379/oj.

Eurostat (2016). *Nomenclature of Territorial Units for Statistics (NUTS)*. Available at https://ec.europa.eu/eurostat/web/nuts.

— (2026). *Agriculture Database*. Available at https://ec.europa.eu/eurostat/web/agriculture/database.

Gneiting, Tilmann and Adrian E. Raftery (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *Journal of the American Statistical Association* 102.477, pp. 359–378. DOI: 10.1198/016214506000001437.

Gong, Haipeng, Yaojin Li, Jian Zhang, Biao Zhang, and Xiaoyan Wang (2024). "A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information". In: *Engineering Applications of Artificial Intelligence* 131, p. 107865. DOI: 10.1016/j.engappai.2024.107865.

Guindani, Luana Gonçalves, Gilson Adamczuk Oliveira, Matheus Henrique Dal Molin Ribeiro, Gabriel Villarrubia Gonzalez, and José Donizetti de Lima (2024). "Exploring current trends in agricultural commodities forecasting methods through text mining: Developments in statistical and artificial intelligence methods". In: *Heliyon* 10.23, e40568. DOI: 10.1016/j.heliyon.2024.e40568.

Gyamerah, Samuel A., Philip Ngare, and Dennis Ikpe (2020). "Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov kernel function". In: *Agricultural and Forest Meteorology* 280, p. 107808. DOI: 10.1016/j.agrformet.2019.107808.

Hatfield, Jerry L. and John H. Prueger (2015). "Temperature extremes: Effect on plant growth and development". In: *Weather and Climate Extremes* 10, pp. 4–10. DOI: 10.1016/j.wace.2015.08.001.

Hernández Hernández, Guillermo C., Jorge Gómez Gómez, and Javier Jiménez-Cabas (2025). "Predictive models based on artificial intelligence to estimate crop yield: A literature review". In: *Agriculture* 15, p. 2438. DOI: 10.3390/agriculture15232438.

Huber, Florian, Artem Yushchenko, Benedikt Stratmann, and Volker Steinhage (2022). "Extreme gradient boosting for yield estimation compared with deep learning approaches". In: *Computers and Electronics in Agriculture* 202, p. 107346. DOI: 10.1016/j.compag.2022.107346.

Hyndman, Rob J. and George Athanasopoulos (2021). *Forecasting: Principles and Practice*. 3rd ed. Available at https://otexts.com/fpp3/. OTexts.

Hyndman, Rob J. and Anne B. Koehler (2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4, pp. 679–688. DOI: 10.1016/j.ijforecast.2006.03.001.

ISMEA (2024). *Agricultural Market Data*. Available at https://www.ismeamercati.it.

Jabed, Monjur and Masrah Azmi Murad (2024). "Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches". In: *IEEE Access*. In press.

Jeong, Jig Han, Jonathan P. Resop, Nathaniel D. Mueller, Dennis H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Dennis J. Timlin, Kyo-Moon Shim, James S. Gerber, Vangimalla R. Reddy, and Soo-Hyung Kim (2016). "Random forests for global and regional crop yield predictions". In: *PLoS ONE* 11.6, e0156571. DOI: 10.1371/journal.pone.0156571.

Khaki, Saeed and Lizhi Wang (2019). "Crop yield prediction using deep neural networks". In: *Frontiers in Plant Science* 10, p. 621. DOI: 10.3389/fpls.2019.00621.

Khaki, Saeed, Lizhi Wang, and Sotirios V. Archontoulis (2020). "A CNN-RNN framework for crop yield prediction". In: *Frontiers in Plant Science* 10, p. 1750. DOI: 10.3389/fpls.2019.01750.

Klompenburg, Thomas van, Ayalew Kassahun, and Cagatay Catal (2020). "Crop yield prediction using machine learning: A systematic literature review". In: *Computers and Electronics in Agriculture* 177, p. 105709. DOI: 10.1016/j.compag.2020.105709.

Kornher, Lukas, Tomas Baležentis, and Fabio Gaetano Santeramo (2024). "EU food price inflation amid global market turbulences during the COVID-19 pandemic and the Russia-Ukraine War". In: *Applied Economic Perspectives and Policy* 46.4, pp. 1563–1584. DOI: 10.1002/aepp.13483.

Lecerf, Romain, Andrej Ceglar, Raúl López-Lozano, Marijn Van Der Velde, and Bettina Baruth (2018). "Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe". In: *Agricultural Systems* 168, pp. 191–202. DOI: 10.1016/j.agsy.2018.03.002.

Lee, Yun Shin and Stefan Scholtes (2014). "Empirical prediction intervals revisited". In: *International Journal of Forecasting* 30.2, pp. 217–234. DOI: 10.1016/j.ijforecast.2013.07.018.

Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata (2013). "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". In: *Journal of Experimental Social Psychology* 49.4, pp. 764–766. DOI: 10.1016/j.jesp.2013.03.013.

Ling, Shuai, Can Zheng, and Dongmin Cho (2023). "How Brand Knowledge Affects Purchase Intentions in Fresh Food E-Commerce Platforms: The Serial Mediation Effect of Perceived Value and Brand Trust". In: *Sustainability* 15.5, p. 4385. DOI: 10.3390/su15054385.

López-Lozano, Raúl, Grégory Duveiller, Lorenzo Seguini, Michele Meroni, Sergio García-Condado, Jan Hooker, Olivier Leo, and Bettina Baruth (2015). "Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: Strengths and limitations at pan-European level". In: *Agricultural and Forest Meteorology* 206, pp. 12–32. DOI: 10.1016/j.agrformet.2015.02.021.

Maestrini, Bernardo, Gordan Mimic, Pepijn A. J. van Oort, Keiji Jindo, Sanja Brdar, Ioannis N. Athanasiadis, and Frits K. van Evert (2022). "Mixing process-based and data-driven approaches in yield prediction". In: *European Journal of Agronomy* 139, p. 126569. DOI: 10.1016/j.eja.2022.126569.

Manogna, R. L., Vijay Dharmaji, and S. Sarang (2025). "Enhancing agricultural commodity price forecasting with deep learning". In: *Scientific Reports* 15, p. 20903. DOI: 10.1038/s41598-025-05103-z.

McMaster, Gregory S. and W. W. Wilhelm (1997). "Growing degree-days: one equation, two interpretations". In: *Agricultural and Forest Meteorology* 87.4, pp. 291–300. DOI: 10.1016/S0168-1923(97)00027-0.

Meinshausen, Nicolai (2006). "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7, pp. 983–999.

Min, Youngho, Young Rock Kim, YunKyong Hyon, Taeyoung Ha, Sunju Lee, Jinwoo Hyun, and Mi Ra Lee (2025). "RNN and GNN based prediction of agricultural prices with multivariate time series and its short-term fluctuations smoothing effect". In: *Scientific Reports* 15, p. 13681. DOI: 10.1038/s41598-025-97724-7.

Newlands, Nathaniel K., David S. Zamar, Louis A. Kouadio, Yuan Zhang, Aston Chipanshi, Andries Potgieter, Serge Toure, and Harvey S. J. Hill (2014). "An integrated, probabilistic model for improved

seasonal forecasting of agricultural crop yield under environmental uncertainty". In: *Frontiers in Environmental Science* 2, p. 17. DOI: 10.3389/fenvs.2014.00017.

Oikonomidis, Alexandros, Cagatay Catal, and Ayalew Kassahun (2022). "Hybrid deep learning-based models for crop yield prediction". In: *Applied Artificial Intelligence* 36.1, e2031823. DOI: 10.1080/08839514.2022.2031823.

Pan, Z. and X. Zheng (2023). "Price volatility transmission of perishable agricultural products: evidence from China". In: *Economic Research-Ekonomska Istraživanja* 36.1. DOI: 10.1080/1331677X.2023.2180058.

Paudel, Dilli, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjef Osinga, Christos Pylianidis, and Ioannis N. Athanasiadis (2021). "Machine learning for large-scale crop yield forecasting". In: *Agricultural Systems* 187, p. 103016. DOI: 10.1016/j.agsy.2020.103016.

Paudel, Dilli, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N. Athanasiadis (2022). "Machine learning for regional crop yield forecasting in Europe". In: *Field Crops Research* 276, p. 108377. DOI: 10.1016/j.fcr.2021.108377.

Paudel, Dilli, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, and Ioannis N. Athanasiadis (2023). "Interpretability of deep learning models for crop yield forecasting". In: *Computers and Electronics in Agriculture* 206, p. 107663. DOI: 10.1016/j.compag.2023.107663.

Raasveldt, Mark and Hannes Mühleisen (2019). "DuckDB: An Embeddable Analytical Database". In: *Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*. ACM, pp. 1981–1984. DOI: 10.1145/3299869.3320212.

Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès (2019). "Conformalized Quantile Regression". In: *Advances in Neural Information Processing Systems* 32.

Ryo, Masahiro (2022). "Explainable artificial intelligence and interpretable machine learning for agricultural data analysis". In: *Artificial Intelligence in Agriculture* 6, pp. 257–265. DOI: 10.1016/j.aiia.2022.11.003.

Sabu, Kiran M. and T. K. Manoj Kumar (2020). "Predictive analytics in Agriculture: Forecasting prices of Arecanuts in Kerala". In: *Procedia Computer Science* 171, pp. 699–708. DOI: 10.1016/j.procs.2020.04.076.

Shapiro, Samuel S. and Martin B. Wilk (1965). "An Analysis of Variance Test for Normality (Complete Samples)". In: *Biometrika* 52.3/4, pp. 591–611. DOI: 10.2307/2333709.

Statistics Netherlands (CBS) (2024). *StatLine: Agricultural Prices*. Available at https://www.cbs.nl/en-gb/figures.

Statistics Netherlands (CBS) and Wageningen Economic Research (2022). *Agricultural sector records slightly higher income in 2022*. Published 19 December 2022. Available at https://www.cbs.nl/en-gb/news/2022/51/agricultural-sector-records-slightly-higher-income-in-2022.

Tashman, Leonard J. (2000). "Out-of-sample tests of forecasting accuracy: an analysis and review". In: *International Journal of Forecasting* 16.4, pp. 437–450. DOI: 10.1016/S0169-2070(00)00065-0.

Taylor, Sean J. and Benjamin Letham (2018). "Forecasting at Scale". In: *The American Statistician* 72.1, pp. 37–45. DOI: 10.1080/00031305.2017.1380080.

Toreti, Andrea (2026). *Gridded Agro-Meteorological Data in Europe*. European Commission, Joint Research Centre (JRC) dataset. Available at https://data.europa.eu/doi/10.2905/JRC.WG6925R. DOI: 10.2905/JRC.WG6925R.

Velde, Marijn van der and Luigi Nisini (2019). "Performance of the MARS-crop yield forecasting system for the European Union: Assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015". In: *Agricultural Systems* 168, pp. 203–212. DOI: 10.1016/j.agsy.2018.06.009.

Wang, Haoyan, Qianying Liang, Jeffrey T. Hancock, and Taghi M. Khoshgoftaar (2024a). "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods". In: *Journal of Big Data* 11.1, p. 44. DOI: 10.1186/s40537-024-00905-w.

Wang, Xingguo, Henry Bryant, and Chengcheng J. Fei (2024b). *Forecasting crop yield variability in response to weather information using quantile random forest regression - An application in non-irrigated upland cotton*. Tech. rep. 4686927. SSRN. DOI: 10.2139/ssrn.4686927.

Winkler, Robert L. (1972). "A Decision-Theoretic Approach to Interval Estimation". In: *Journal of the American Statistical Association* 67.337, pp. 187–191. DOI: 10.1080/01621459.1972.10481224.

10 Appendices

The appendices provide the additional visuals, tables, and robustness figures that support the main analysis. The main text focuses on the results that are most relevant for the contract-versus-spot buying decision and this chapter is intended to give more supporting detail behind those conclusions. This includes exploratory data checks, sample sizes, modelling settings, and the full set of visual results for the point forecasts, yield forecasts, and prediction intervals.

10.1 Data Methodology

10.1.1 Exploratory Data Analysis

The figures below give context for the data used throughout the forecasting framework. Figure 10.1 shows which engineered features are most strongly associated with future prices in the training period, while Figure 10.2 shows the underlying monthly price dynamics across all crop-country groups. Together, these figures illustrate why the forecasting task differs strongly by product and country as some series show repeated seasonal structure, while others contain more irregular or sparse price movements.

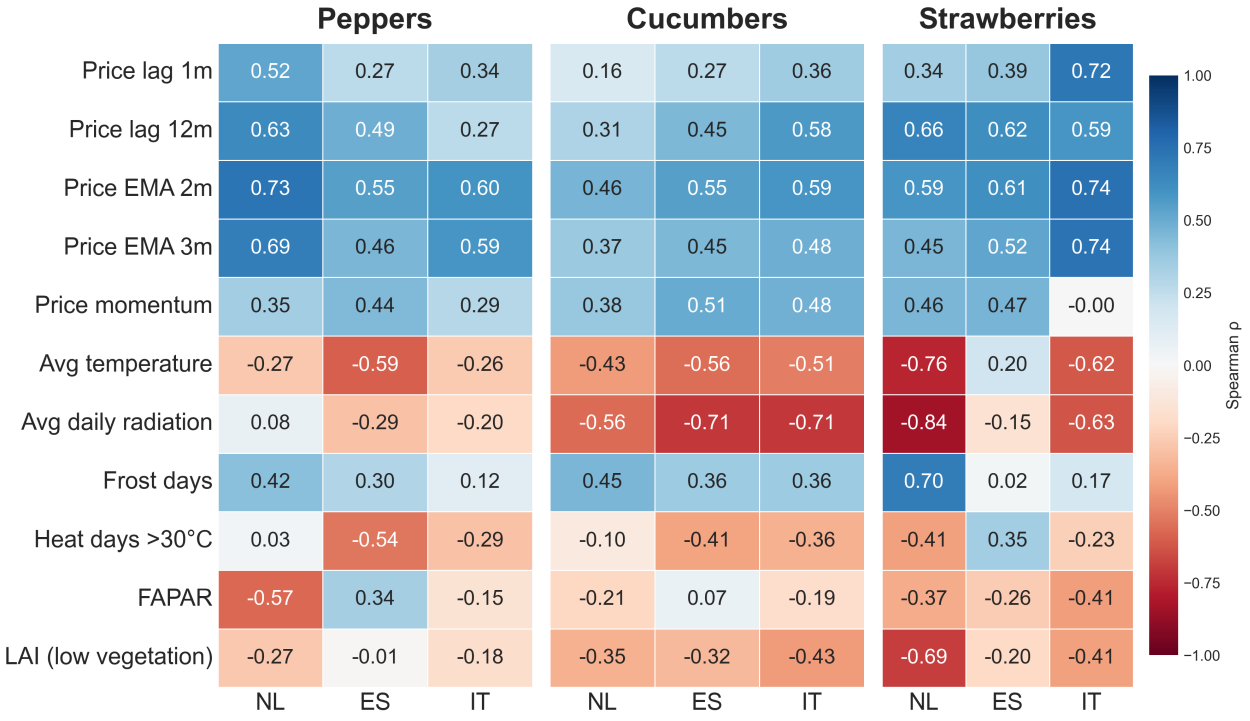

Figure 10.1: Strongest Spearman correlations between features and next-month price across the nine crop-country combinations, computed on the training period 2004-2018. Positive correlations are shown in blue and negative correlations in red.

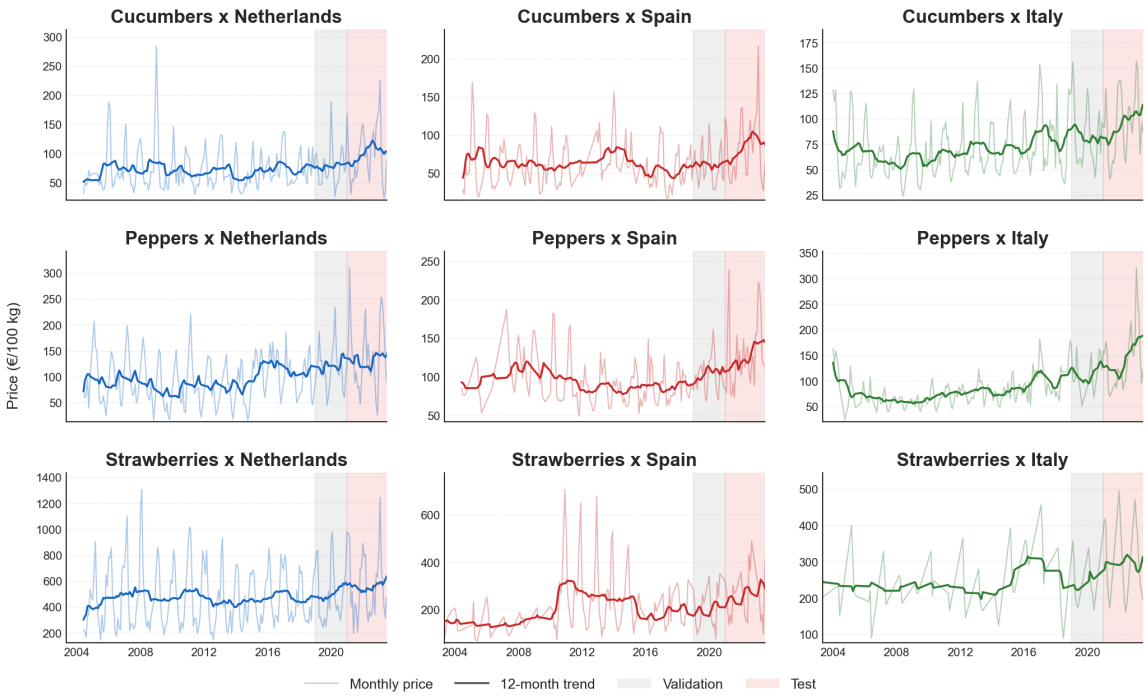

Figure 10.2: Monthly price series for all nine crop-country combinations, shown together with a 12-month rolling trend. The validation period (2019-2020) and test period (2021-2023) are highlighted.

10.1.2 Dataset Characteristics

Because the models are estimated separately by crop-country group, the available number of observations matters for interpreting the results. Table 10.1 gives the usable price and yield observations per group, and Table 10.2 summarizes how the feature set changes during preprocessing and feature selection.

Table 10.1: Price observation counts and annual yield observations per crop-country model, split by training (2004-2018), validation (2019-2020), and test period (2021-2023).

Crop	Country	Price months			Yield years		
		Train	Val	Test	Train	Val	Test
Cucumbers	ES	149	24	36	14	2	3
Cucumbers	IT	180	24	36	15	2	3
Cucumbers	NL	163	24	35	16	2	3
Peppers	ES	143	24	36	16	2	3
Peppers	IT	149	20	28	16	2	3
Peppers	NL	133	19	31	16	2	3
Strawberries	ES	133	18	27	16	2	3
Strawberries	IT	53	10	19	16	2	3
Strawberries	NL	174	24	36	16	2	3

Table 10.2: Step-by-step feature counts for the price and yield feature selection pipelines.

Selection step	Price	Yield
Initial candidates	165	166
After near-zero variance filter	162	163
After weak correlation removal ($\lvert r\rvert \leq 0.05$)	140	147
After MI rescue (MI ≥ 0.02)	166	166
After VIF reduction (VIF > 10)	126	126
After RF importance (per crop-country model)	26-28	26-28

10.2 Mathematical Modelling

The model comparison in the main text only reports final forecast performance. For reproducibility, Table 10.3 gives the hyperparameter search spaces used for Random Forest and XGBoost. The LSTM example in Figure 10.3 is included to show why LSTM was not retained for the final model comparison as the validation forecast was too smooth to capture the sharp price movements that matter for contract evaluation.

10.2.1 Hyperparameter Search Spaces

Hyperparameters for Random Forest and XGBoost were selected using 60-iteration randomised search with a predefined validation split on the 2019-2020 period, optimising for RMSE. Tables 10.3a and 10.3b report the full search spaces.

Table 10.3: Hyperparameter search spaces for the tree-based models.

(a) Random Forest

Parameter	Values searched
`n_estimators`	100, 200, 400
`max_depth`	5, 8, 10, 15, `None`‡
`min_samples_leaf`	3, 5, 10
`max_features`	`sqrt`†, 0.5, 0.8

(b) XGBoost

Parameter	Values searched
`n_estimators`	100, 200, 300, 500
`max_depth`	3, 5, 7, 10
`learning_rate`	0.01, 0.03, 0.05, 0.1
`subsample` / `colsample_bytree`	0.6, 0.7, 0.8, 0.9

† `sqrt`: at each split, a random subset of approximately $\sqrt{p}$ features is considered, where p is the total number of features.
‡ `None`: no maximum tree-depth constraint so trees are grown until the leaf-size stopping criterion is reached.

10.2.2 Deep Learning Modelling

This figure shows an example of the LSTM results and visualizes that LSTM resulted in a forecast that was too smooth and undertrained to retain in the rest of the research.

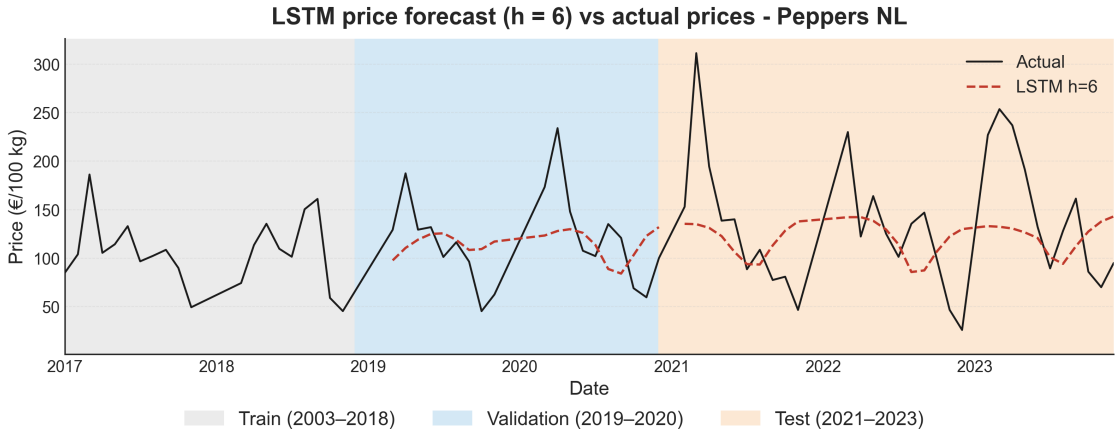

Figure 10.3: Exploratory LSTM price forecast example for Peppers × NL, illustrating the overly smooth behaviour that motivated excluding LSTM from the main model comparison.

10.3 Evaluation Setup

[...Confidential information removed...]

10.4 Results & Validation

The result figures in this appendix extend the main results chapter to give the fuller set of crop-country diagnostics behind the choices made.

10.4.1 Point Price Forecast

The point forecast appendix figures show how the selected HW anchor compares with the main benchmarks across the evaluated crop-country groups. Figure 10.4 compares HW with the Naive Seasonal baseline, Figure 10.5 checks whether the winning model changes across error metrics, and Figure 10.6 shows why RF does not replace HW as the point-forecast anchor.

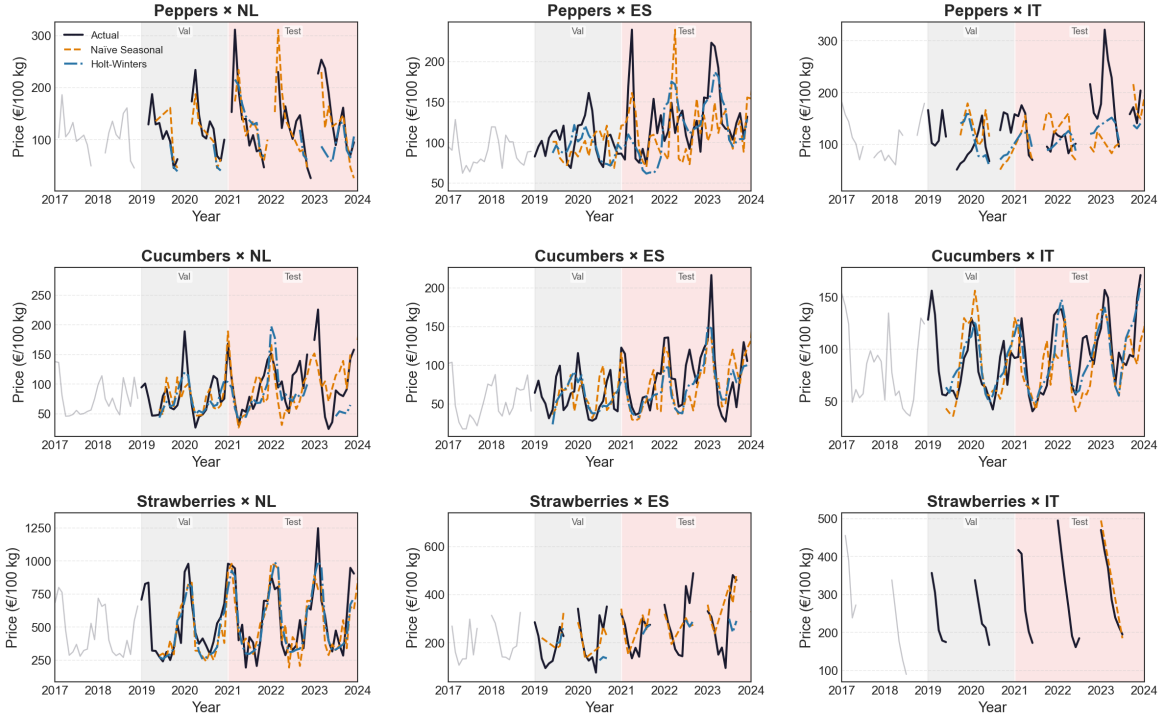

Figure 10.4: Six-month-ahead price forecasts from Naive Seasonal and HW compared with realized prices across the crop-country groups. Strawberries × Italy should be interpreted with caution because few valid test observations remain after horizon alignment.

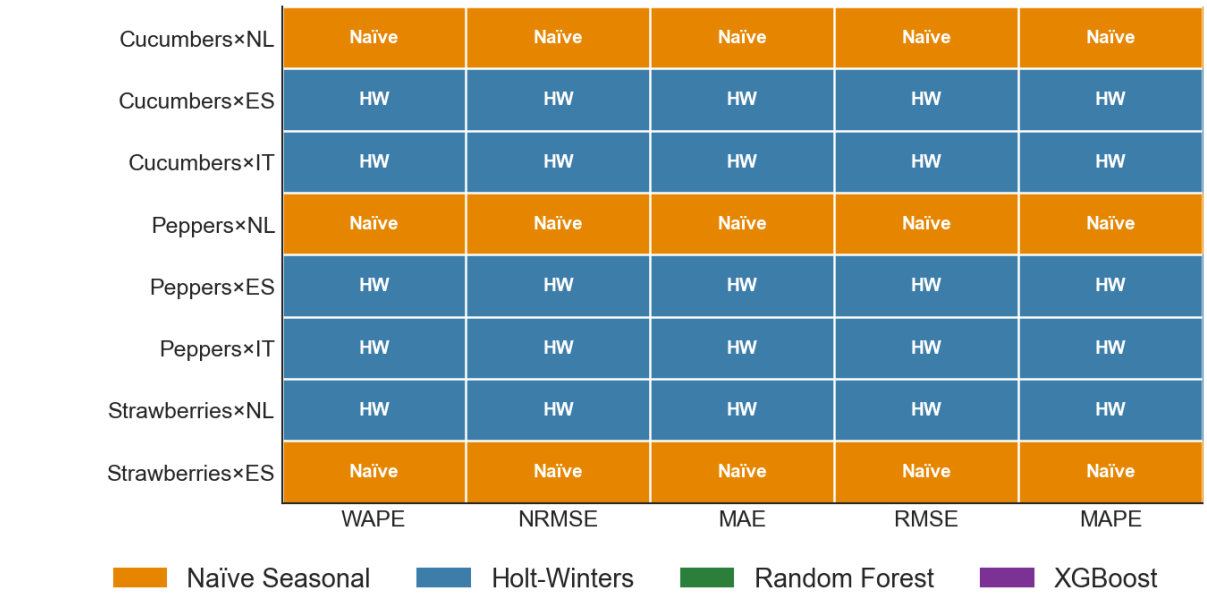

Figure 10.5: Best-performing point forecast model per crop-country group and evaluation metric in the six-month contract-window setting during the 2021-2023 test period. Strawberries × Italy is excluded because too few valid test observations remain after horizon alignment.

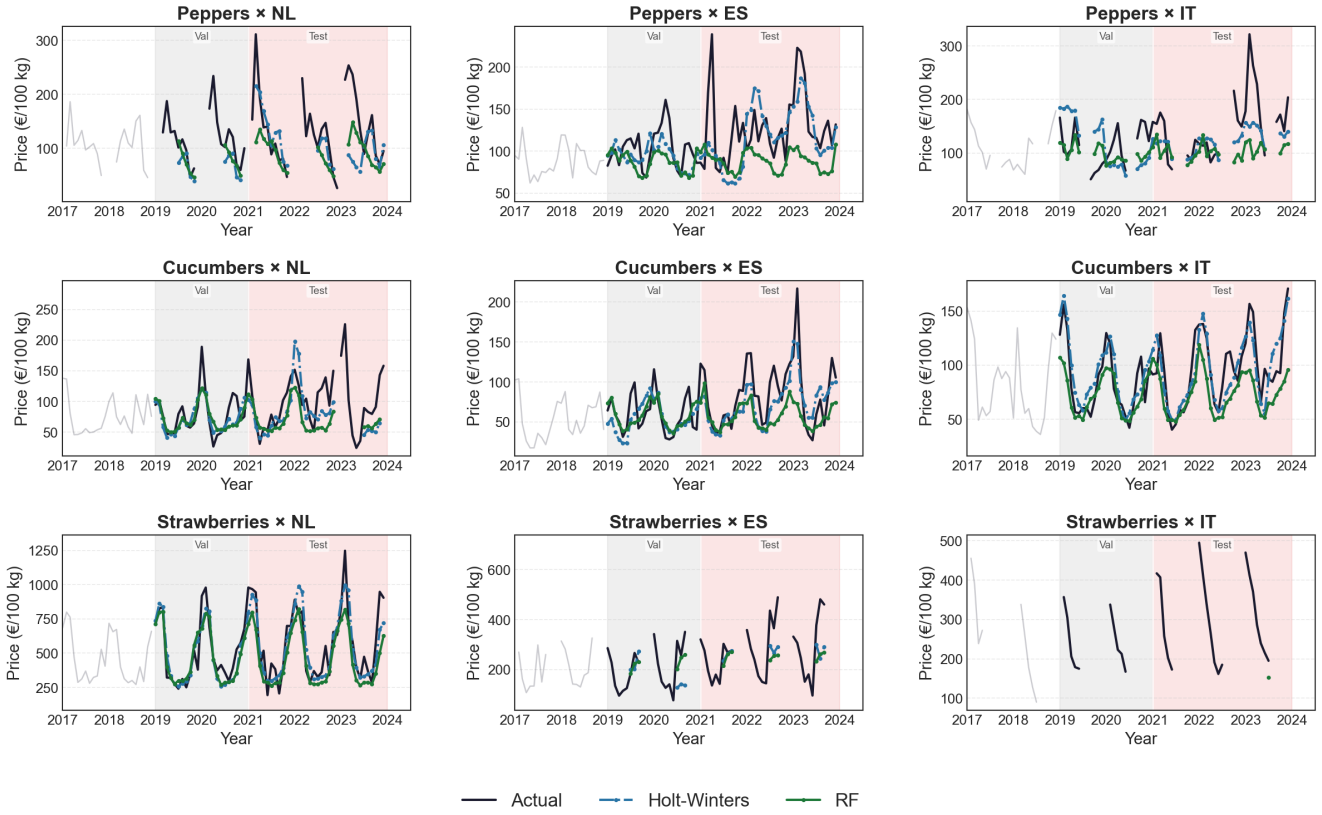

Figure 10.6: Six-month contract-window price forecasts from Random Forest and HW compared with realized prices across the crop-country groups. Strawberries × Italy should be interpreted with caution because few valid test observations remain after horizon alignment.

10.4.2 Yield Forecast

The following yield appendix figures show the annual yield forecasts behind the exploratory yield results. These are included to make the group-level patterns visible, but should be interpreted cautiously because each crop-country group contains only three annual test observations. The figures therefore support the conclusion that yield is useful as contextual supply information, but not robust enough as a standalone operational forecasting target.

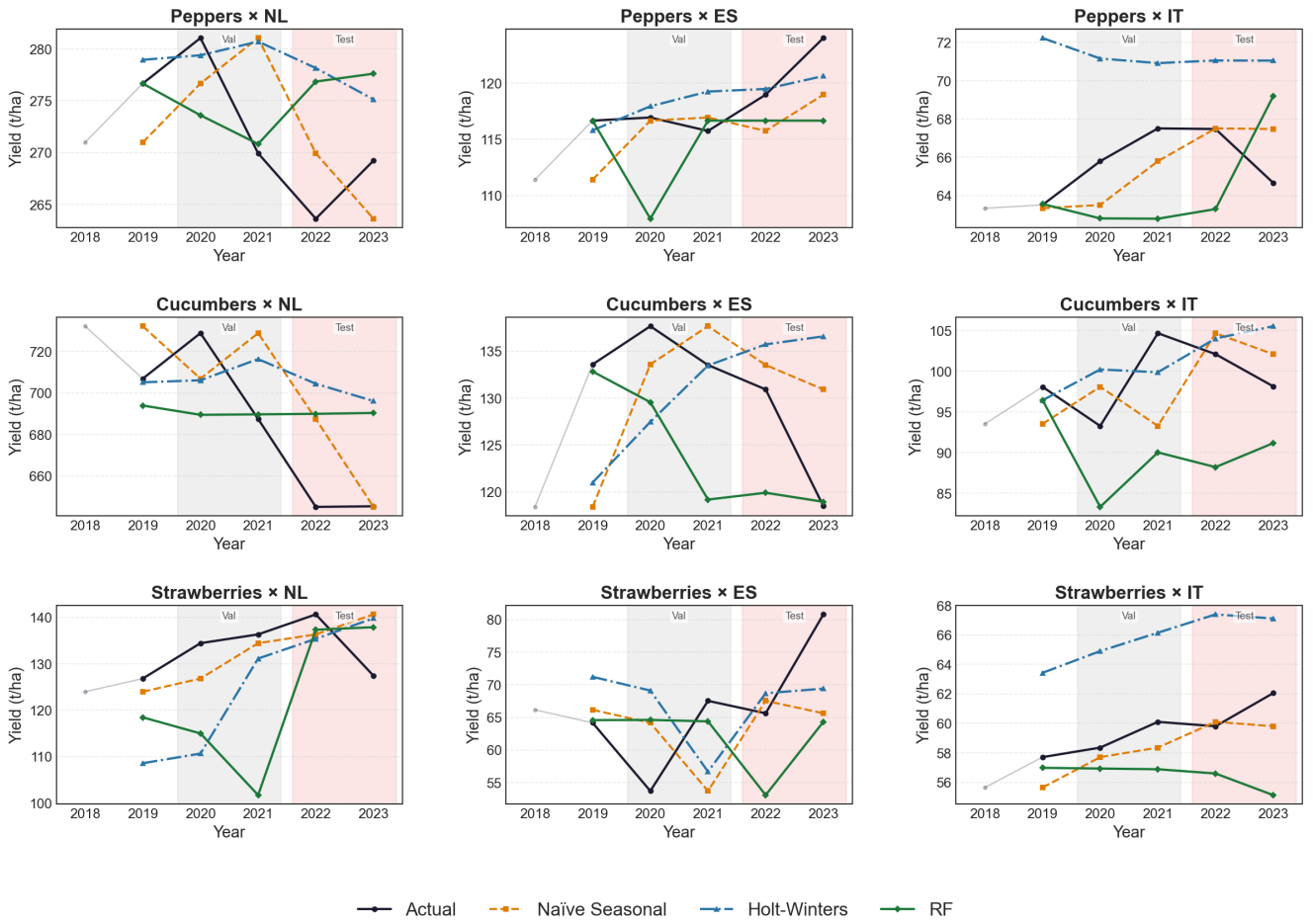

Figure 10.7: One-year-ahead yield forecasts from Naive Seasonal, HW, and Random Forest compared with realized annual yield across the crop-country groups. The validation and test periods are highlighted.

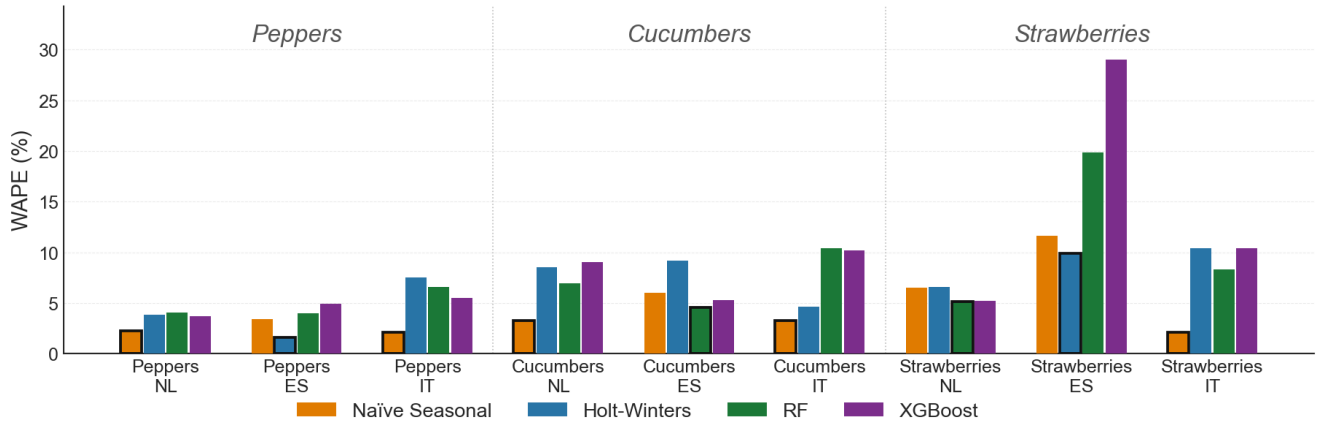

Figure 10.8: Yield forecast WAPE by crop-country group during the 2021-2023 test period. The black border indicates the lowest-error model within each crop-country group. The figure is used as directional evidence only, because each crop-country group contains three annual test observations.

10.4.3 Prediction Interval Results

The prediction interval appendix figures provide the detailed diagnostics behind the interval-method selection. This appendix adds to the main text with the full coverage-width comparisons by year, by crop-country group, and by country, followed by the full 3-by-3 fan charts for each interval method. These figures show where QRF+CQR adds coverage through wider intervals, and why Fixed Empirical remains the more consistent and interpretable default across the full evaluation.

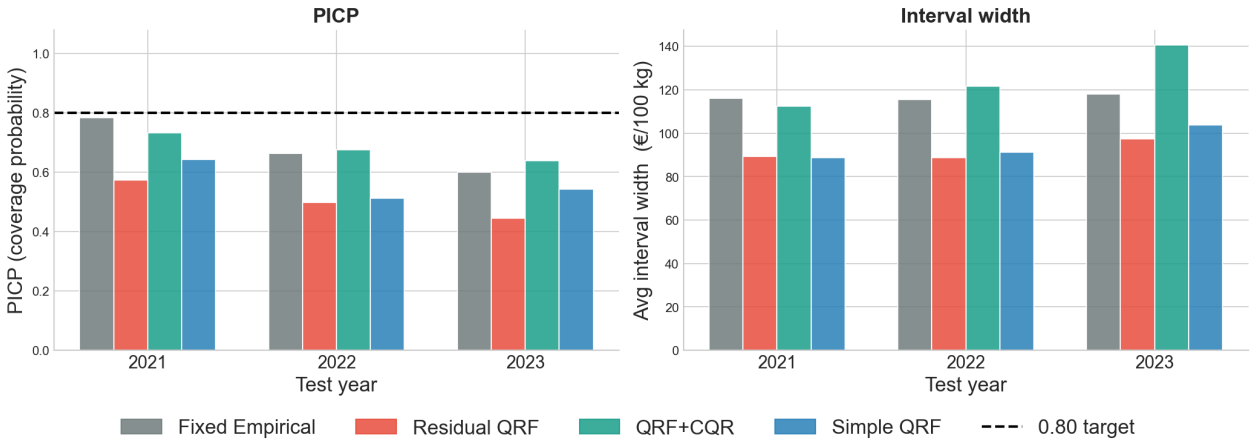

Figure 10.9: Coverage probability (PICP) and relative interval width by test year for 80% prediction intervals in the six-month contract-window setting. The dashed line indicates the nominal 0.80 coverage target. Values are averaged across crop-country groups with sufficient valid observations.

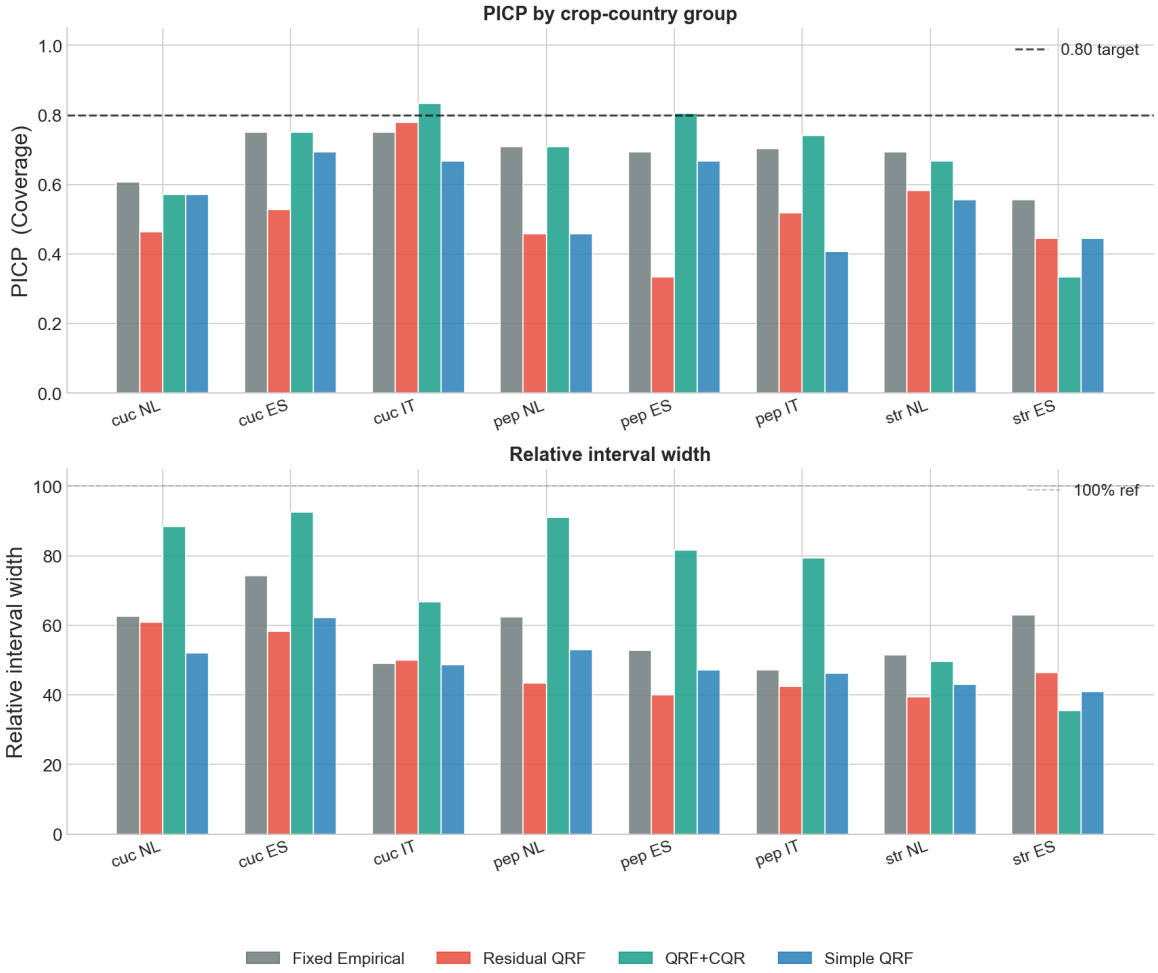

Figure 10.10: Coverage probability (PICP) and relative interval width by crop-country group and interval method in the six-month contract-window setting. The dashed line in the PICP panel indicates the nominal 0.80 coverage target.

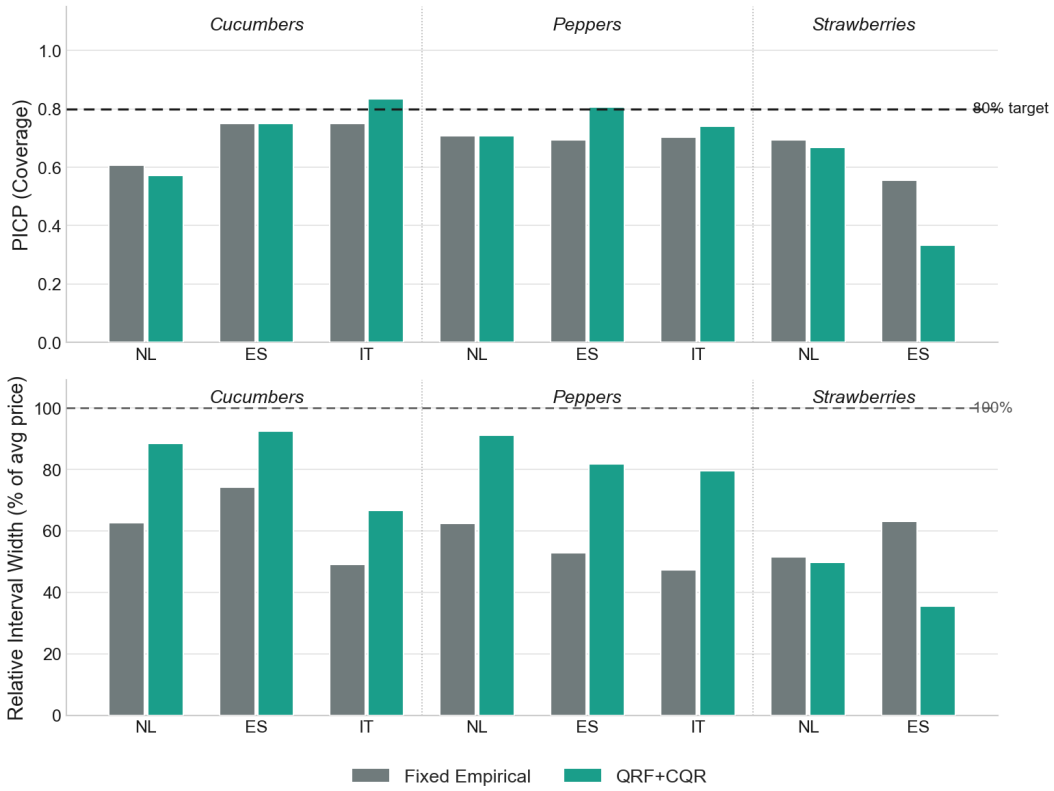

Figure 10.11: Crop-country comparison of Fixed Empirical and QRF+CQR in terms of coverage probability (PICP) and relative interval width. The figure isolates the main operational comparison between the transparent historical-residual interval and the adaptive calibrated residual QRF interval.

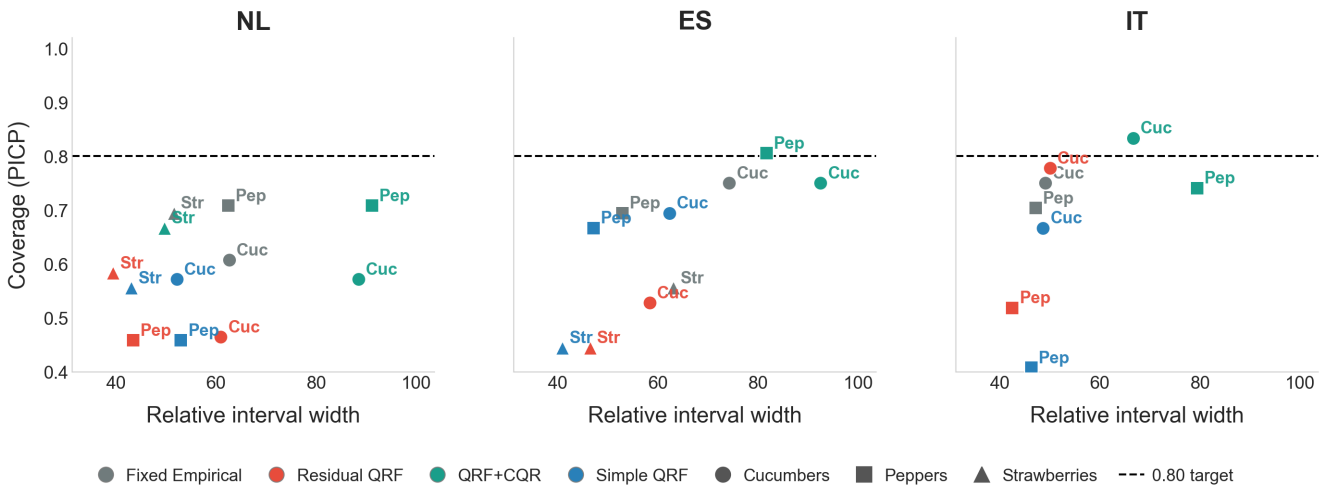

Figure 10.12: Coverage-width trade-off for all prediction interval methods in the six-month contract-window setting, split by sourcing country. Each point represents one method's performance for one crop-country group.

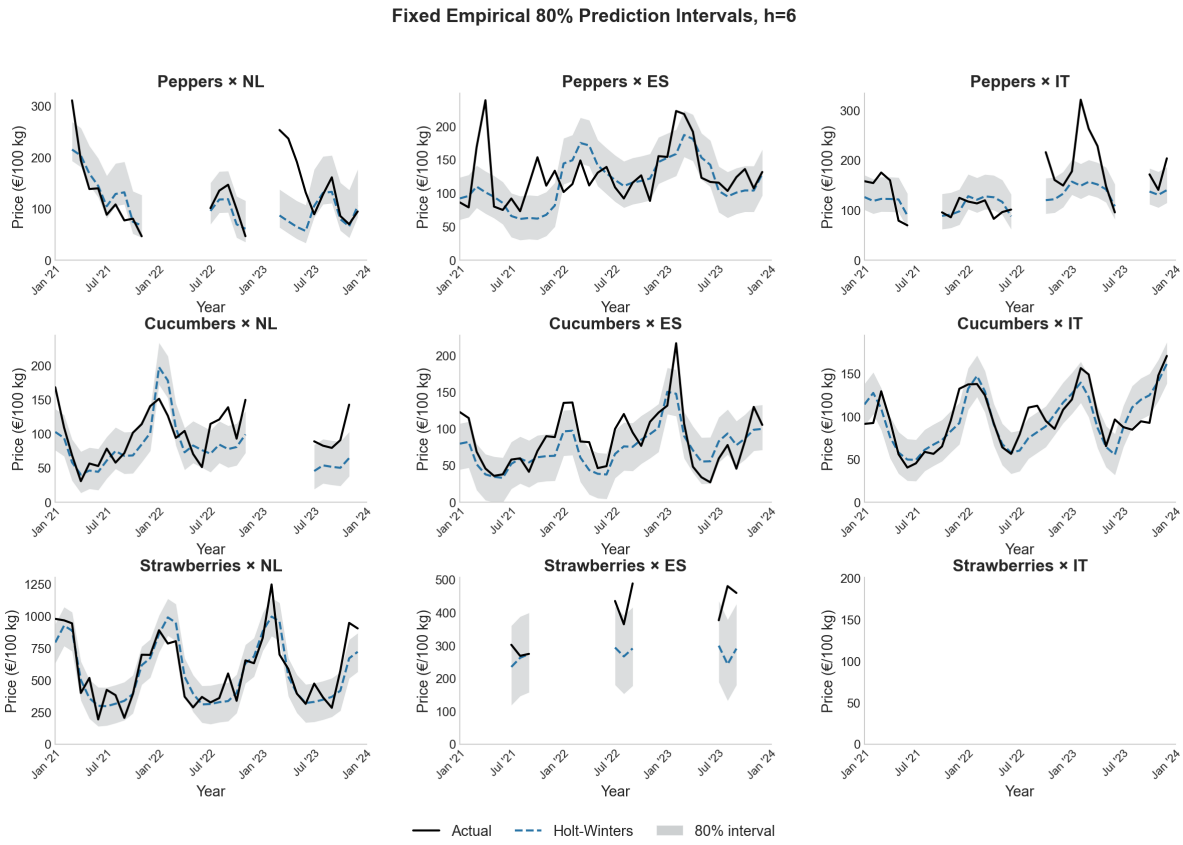

Figure 10.13: Fixed Empirical 80% prediction intervals in the six-month contract-window setting across the evaluated crop-country groups. The method creates a historical residual band around the HW forecast and therefore provides a stable but non-adaptive uncertainty benchmark.

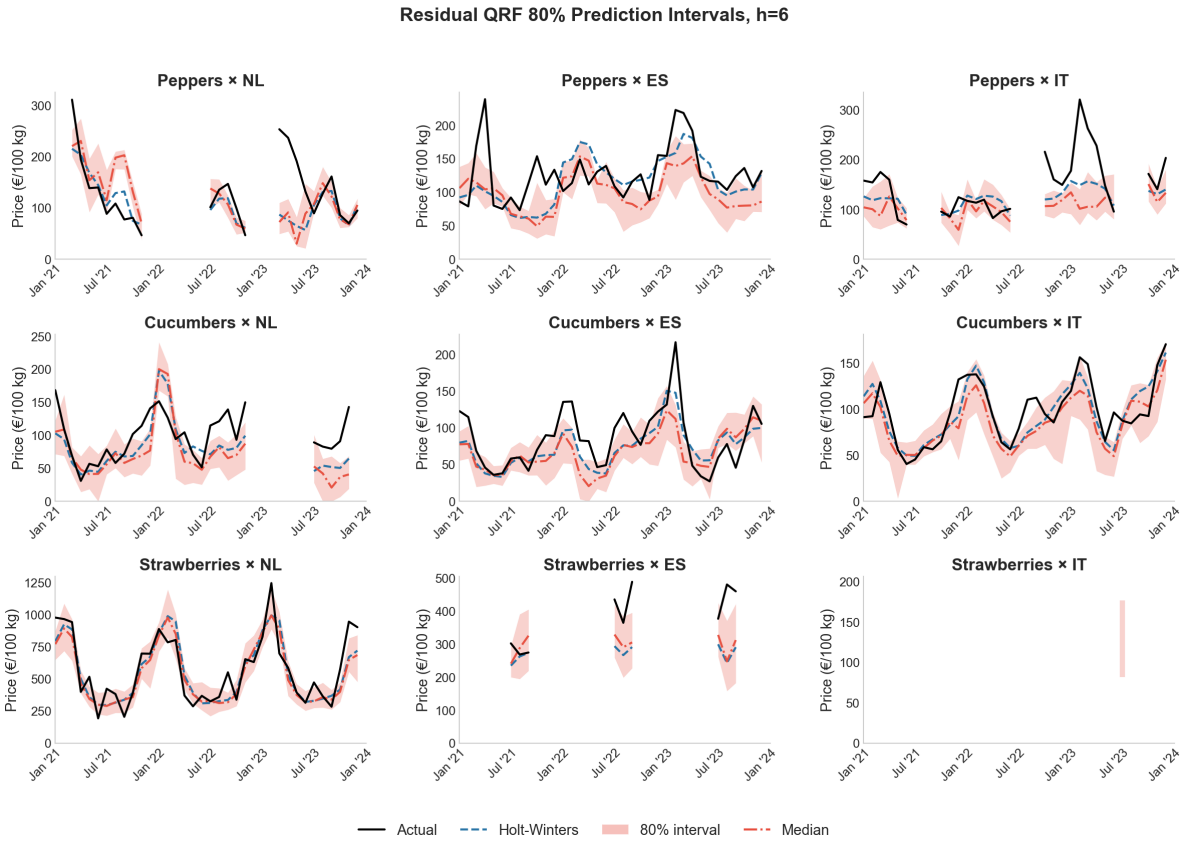

Figure 10.14: Residual QRF 80% prediction intervals in the six-month contract-window setting across the evaluated crop-country groups. The method estimates feature-conditioned residual quantiles around the HW point forecast.

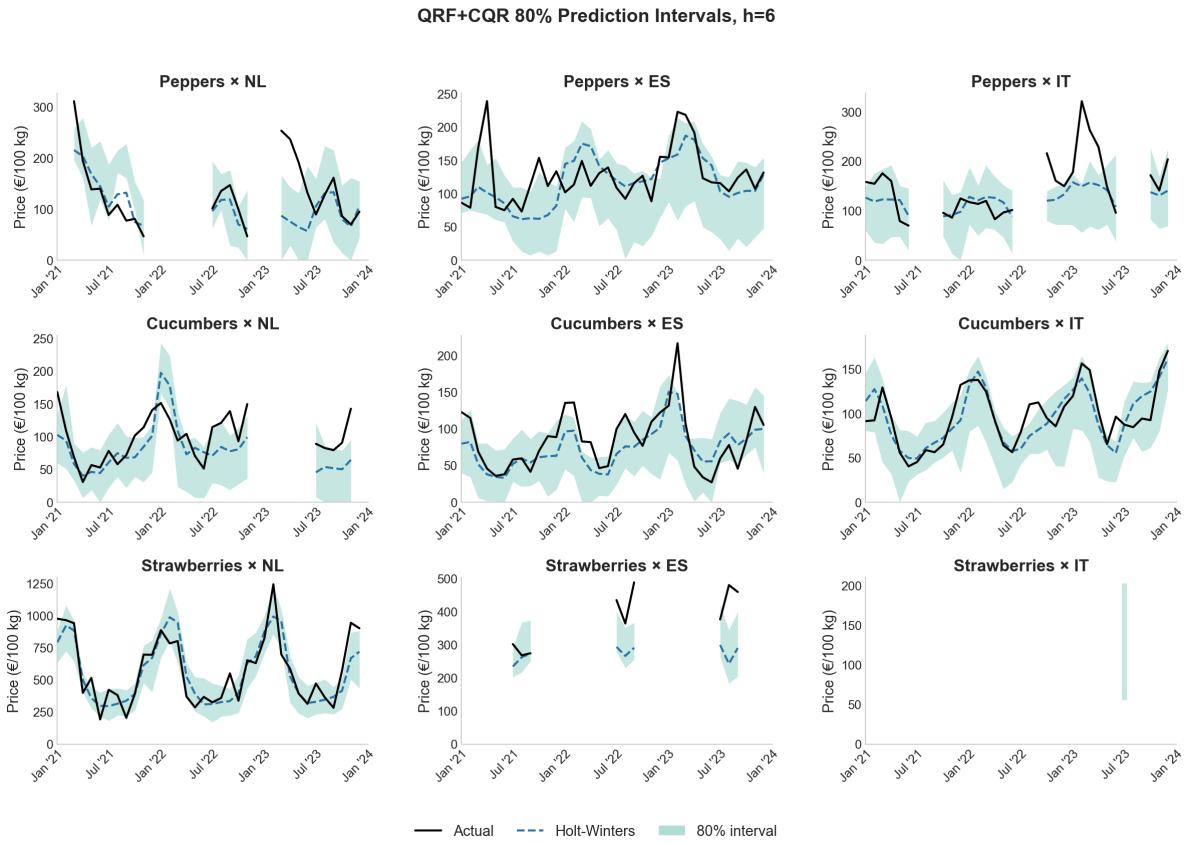

Figure 10.15: QRF+CQR 80% prediction intervals in the six-month contract-window setting across the evaluated crop-country groups. The method applies conformal calibration to the Residual QRF intervals, which generally improves coverage but increases interval width.

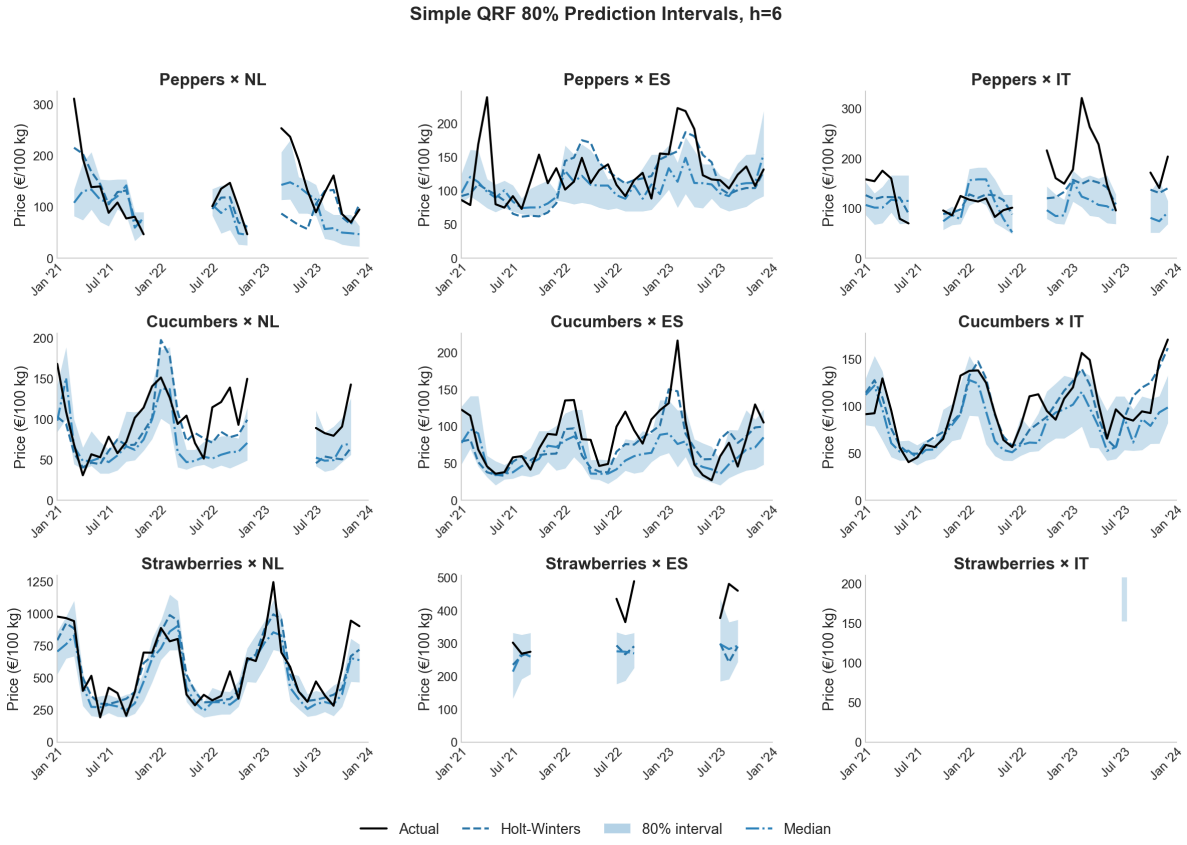

Figure 10.16: Simple QRF 80% prediction intervals in the six-month contract-window setting across the evaluated crop-country groups. Unlike the residual-based methods, Simple QRF models the future price distribution directly.