

Article

Hybrid Supply Chain Model for Wheat Market

Yulia Otmakhova ¹, Dmitry Devyatkin ^{2,*} and He Zhou ^{3,4}¹ Central Economic Mathematical Institute, Russian Academy of Sciences, Moscow 117418, Russia² Federal Research Center “Computer Science and Control”, Russian Academy of Sciences, Moscow 119333, Russia³ School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China; hezhou@ucas.ac.cn⁴ MOE Social Science Laboratory of Digital Economic Forecasts and Policy Simulation, University of Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: devyatkin@isa.ru

Abstract

Accurate modeling of wheat supply chains is of great importance. The methods for forecasting them can be utilized as strategic planning tools to manage sustainable and balanced supply chains, ensuring a high level of food security, economic growth, and social development. In this paper, we focus on wheat international trade indicators, and a regression model is a crucial component for the chain modeling. Trade indicators in the wheat market are inherently complex and exhibit significant stochasticity and non-stationarity due to the intricate interplay of various trade flows and factors, which pose challenges for accurate market forecasting. We proposed a novel hybrid recurrent and graph-transformer-based model to tackle these challenges. We collected and combined data from international providers such as UN FAOSTAT and UN Comtrade for all the world's wheat exporters. The experiments show that the proposed model can accurately predict wheat export levels. We have also analyzed how the proposed model can be utilized to predict exports in the case of some pre-defined trade limitations. In the future, the proposed model could be naturally extended to various derivative products of wheat, supporting real-world grain chain models. Our forecasting methods could be used to create an analytical tool to support strategic decision-making in cognitive situation centers, taking into account the national interests and priorities of actors in the international wheat market.

Keywords: wheat supply chain; wheat trade forecasting; graph transformers; recurrent networks; encoder-decoder architecture; strategic decision making; situational centers



Academic Editor: Ben Clegg

Received: 11 September 2025

Revised: 6 November 2025

Accepted: 14 November 2025

Published: 17 November 2025

Citation: Otmakhova, Y.; Devyatkin, D.; Zhou, H. Hybrid Supply Chain Model for Wheat Market. *Systems* **2025**, *13*, 1026. <https://doi.org/10.3390/systems13111026>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The global food market is a complex network of interacting factors that shape the balance of supply and demand through the integration of global supply chains. Food trade on the world market performs not only a commercial purpose but also solves the most important task of ensuring global food security. Wheat occupies a special place among the goods on the world market; it is one of the most traded commodities. In recent years, there has been a surge in research on the global trade system as a network. In the case of the food market, the reliance of diets on a small number of crops threatens global food security if the production or trade of one or more of these crops is curtailed for any reason. Wheat, maize, and rice comprise approximately half of all human diets and represent 86% of all cereal exports [1]. Global production of these staple crops is highly concentrated in regions known as breadbaskets. At least 72% of the global output of each of the four staple crops

occurs in five countries [2]. Given the oversized role breadbaskets play in food production, understanding their stability under anticipated warming is critical for global food security. Applied to the study of food systems, network analysis indicates that crop trade networks, the volume of trade, and the number of links between countries have increased during the last decades [3–5]. Interdependence between countries leads to a high degree of complexity in international trade [6]. It not only increases the possibility for countries to be influenced by the global market [7] but also promotes the diversification of global supply risks [8].

At the same time, food supply chains should be considered within a larger scope as food systems, involving not only the different (possibly international) actors from farmer to consumer but also including all processes and infrastructure involved in ensuring a population is fed, such as providing inputs, growing, harvesting, processing, transporting, selling, consuming, and producing waste, within environmental, social, economic, and political contexts, and dependent on human labor. Moreover, the spatial boundaries of these systems are not always clear, and they are under stress from an increasing scarcity of inputs and available land, soil degradation, and climate change [9–11]. The primary results of previous research [12] point to a highly concentrated cereal network; therefore, the dependency on a small number of exporters and a low diversification of imports lead to a global cereal system vulnerable to shocks, such as weather extremes, economic shocks, pandemics, and political uproar.

Wheat supply chains are large-scale complex adaptive systems containing multiple countries/regions, multinational corporations, investors, and intermediaries [13]. The methods used to model and forecast these chains can be used as strategic planning tools to manage sustainable and balanced agri-food supply chains to ensure a high level of food security, economic growth, and social development. To model and simulate such large chains, one requires a framework that integrates a large-scale complex adaptive system containing multiple countries/regions, multinational corporations, investors, and intermediaries. To model and simulate such large chains, we require a framework that integrates advanced multi-agent modeling tools and accurate trade-forecasting models.

Agri-food international trade indicators are inherently complex and exhibit significant stochasticity and non-stationarity due to the intricate interplay of various trade flows, factors that pose challenges for accurate market forecasting. In addition to this problem, trade wars and sanctions increase the relevance of trade modeling, provided certain pre-defined limitations are met, similar to a multi-agent setting [14]. The sustainability of food systems requires the diversification of both production centers and trade routes, including the development of re-export hubs in Asia and Africa. However, modern challenges are exacerbating imbalances in global supply chains, which can lead to food shortages and other problems.

Considering these complexities, our primary focus is on developing a solid theoretical framework for building forecasting models. Namely, we investigated various autoregressive forecasting techniques on fixed and limited global trade data to develop a reliable wheat export forecasting model. Other factors, such as geopolitical risks, macroeconomic shocks, or trade policy interventions, were not specifically considered in the study, as they were deemed to be outside the scope of the research; therefore, they can be tested in the future. The main contributions of our paper are as follows (Experimental code and data are available at https://github.com/masterdoors/agro_graph_transformer (accessed on 5 November 2025)):

- We propose a hybrid model that combines a graph transformer and recurrent network architectures to tackle the problems above. This model utilizes the transformer to catch the interdependence between wheat export quantities in different countries. As a recurrent network, the proposed model also generates hidden embeddings for each

country and export direction and utilizes these embeddings from the previous step to forecast exports. This way, the model summarizes the trading history via the hidden embeddings and uses them to perform accurate export predictions.

- We show how the proposed model can be applied to implement *if-then* scenarios in a multi-agent-like setting.

This paper is structured as follows: Section 2 discusses the current state of the art in international food trade forecasting. Section 3 describes the dataset we utilized to perform the experiments. Section 4 contains the proposed hybrid wheat export forecasting model. Section 5 provides the experimental results. Finally, Sections 6 and 7 contain the discussion and conclusions.

2. Related Work

Much of the work in the forecasting of food production and trade flows uses linear regression and autoregressive models such as seasonal ARIMA, which employ a limited set of indicators, including basic climatic features (temperature, precipitation) and trade indicators. This approach was used, for example, in [15]. Often, in addition to climatic and trade indicators, Earth remote sensing data is also used, which helps forecast production accurately. For example, in previous research [16], data from the Sentinel-2 database [17] showed that remote sensing-based indicators can improve the accuracy of forecasts compared to a basic model that considers only the planting area. One of the most established methods for ensuring trade network topology in market forecasting is the VAR (Vector Autoregression) model or its variants [18]. In VAR models, each output variable depends on other variables, which reflect the modeled trade indicators of market players. In VAR models, all the variables are treated as interdependent, which is the primary limitation of the model. The original VAR model ensures a stationary time series, though there are some ad hoc modifications mostly based on pre-defined linear or polynomial trend terms [19], or even recurrent neural networks (LSTM) [20]. Another option here is the VECM (vector error correction model) if the variables are cointegrated. For example, in a previous paper [21], such a model is used to forecast price transmission.

In our opinion, the primary drawbacks of the VAR models are that they ensure only linear dependencies between variables. Moreover, as the trade network and time lag grow (which is very relevant for worldwide trade modeling), the model becomes very large and tends to overfit. SC-VAR [22] tackles that problem by introducing sparsity to the model parameters, though the explicit introduction of sparsity constraints might lead to a decrease in modeling performance. It is worth noting that our Recurrent Graph-Transformer model tackles the problem out of the box with an attention mechanism, which naturally makes the dependencies between variables sparser.

PanelVAR and PanelVECM approaches introduce a cross-sectional dimension to the modeling. A panel VAR is similar to large-scale VARs, which allow for both dynamic and static interdependencies. For example, in a previous paper [23], countries are grouped into four stages. They applied a panel VAR model to study the effects of the market shocks to domestic product and fish production on fish consumption at the global and cluster levels.

In the case of medium-term and long-term food market forecasting, it is necessary to consider a large number of interrelated characteristics, some of which are nonlinearly related to the predicted value [24]. Therefore, nonlinear models are actively used. The most common way to create such models is to use machine learning and neural network models. For example, previous papers [25,26] presented several methods to forecast various crop productions. They tested several methods, such as Ridge Regression, Gradient Boosting Regression, and the Long-Short Term Memory recurrent network (LSTM). They noted the relationship between the amount of accumulated historical data on crop production in the

region and the accuracy of forecasts. Recent advances in modeling complex non-linear dependencies are related to Kolmogorov–Arnold networks (KANs) [27] and their modifications to analyze time series (TKANs) [28]. KAN uses spline-parametrized univariate functions in place of traditional linear weights, enabling them to dynamically learn activation patterns and significantly enhancing interpretability. An interesting approach is also proposed in [29], where they combine KANs with the Transformer architecture, which helps consider long-term dependencies between indicators. Although this approach seems to be strong in catching the dependencies, the model size grows with the time-series length, which is a major drawback in the case of limited training data. Thus, in this paper, we employ recurrent models to catch those dependencies.

In another paper [30], TKAN (EleKAN) was applied to forecast the Australian New South Wales electricity market, which includes various features, including the total demand, demand delay, power consumption, and temporal factors such as weekdays, holidays, and time intervals. The experimental results demonstrated that EleKAN significantly improves the precision and dependability of power demand estimates. Le et al. employed TKAN [31] to forecast the world gold market. The results indicated that the proposed model consistently outperforms other models, achieving the lowest forecasting errors across various feature combinations.

Considering the complex non-linear dependencies between different features, which affect food supply chains, much research focuses on utilizing Transformer-based neural networks. These networks were initially designed to capture complex dependencies between text features. Recently, different approaches to make the Transformers applicable to modeling continuous time series have been developed. For example, a previous paper [32] proposed the time-series transformer to forecast crop production. Time-series transformers closely follow the original transformer architecture adapted for time-series data. Since the range of the time-series values is continuous, the time-series transformer uses a swappable distribution head as its final layer. Namely, it models the parameters of a continuous distribution of the target values and is trained by minimizing its corresponding negative log-likelihood loss. Another paper [33] introduced a framework that integrates trade forecasting with time-series decomposition to split trade value sequences into seasonal, trend, and residual elements, identifying a potential economic cycle. The framework utilizes a Transformer-based model with a spatiotemporal encoder to process these components. The experiments showed that the proposed approach overcomes other models in terms of accuracy.

One of the novel frameworks to forecast highly stochastic trade flows is to enable transformers to model the Fourier spectrum of the target values. Another paper [34] proposed a Fourier neural operator, which enables efficient learning of mappings between infinite-dimensional spaces using input–output pairs. The motivation here stems from the image-processing domain [35]. The signal-to-spectrum conversion can be used to suppress redundant or noisy frequency components and retain only those features that are beneficial for the modeling [36]. Another paper [37] introduced a modification of the method from reference [34] called the transformer-based neural operator. This model comprises several multi-head attention mechanisms with a Fourier-decomposition-based neural spectral regressor module. The results showed that the proposed model achieves better accuracy than the original spectrum regression models.

Although the models above provide a helpful tool for dealing with interconnected trade flows, they still lack a mechanism for utilizing information from past steps. A previous paper [38] proposed a graph Transformer architecture for temporal data. First, they learnt the structure of the graph that represents the spatial dependency between the data. Then, they sparsified the standard transformer based on that structure. This approach

significantly outperformed other models on the taxi ride-hailing demand task. However, this approach leads to too large models, which are easily overfit for large graphs. Moreover, the model topology is data-dependent; therefore, it cannot be utilized to process data with a different dependency structure. In another paper [39], researchers combined graph convolution and recurrent networks, which made it possible to model time series with a rather compact model, though the recurrent network may suffer from gradient vanishing effects on large time series. To deal with long-term dependencies, another paper [40] proposed combining the graph convolution network with the self-attention mechanism of transformers. They used a sliding window to catch temporary dependencies; therefore, extending the size of the window for large time series leads to proportionate growth of the model size. More studies [41,42] developed that approach with the STGformer architecture. STGformer combines the attention mechanism with graph-convolution networks to build spatial-temporal attention layers and consider both global and local data patterns. Unlike other architectures that require multiple attention layers, this attention layer captures high-order spatiotemporal interactions in a single layer, reducing computational cost. However, the graph convolutions may lead to a lack of important information about the state of particular countries and trade flows, and the size of the attention weight matrix still depends on the length of the time series; therefore, the sequence length is limited. In this paper, we mitigate this problem with recurrent transformers, which combine a graph transformer for spatial dependencies with a recurrent layer for temporal ones. Another paper [43] utilized a similar general idea, though they used a pre-defined function to explicitly estimate edge weights based on the previous ones; therefore, the model can handle the graph's topology dynamics. In our case, edges are assigned to trade (export/import) between countries, which makes that model useless for us. In contrast to that model, we encode hidden temporal states for the vertices and edges, which are then used to predict the wheat production and export, respectively.

Another related direction of research is the use of Transformer decoder networks to model temporal data. A previous paper [44] suggested using the Transformer decoder as a conditional quantile estimator to predict the quantiles of prediction residuals, which are used to estimate the prediction interval. They show that the decoder benefits the estimation of the prediction interval by learning temporal dependencies across past prediction residuals. In this paper, we utilize the Transformer's encoder-decoder architecture to model conditional scenarios, where export volume level between some countries is limited to a pre-defined value.

One approach is the use of log-linear gravity regressions. The test is performed by adding a dummy variable representing the presence or absence of a particular policy [45]. Another approach has been to use computable and applied general equilibrium models (CGE, AGE). These are large-scale models that can be commonly used to evaluate changes in trade policy [46]. CGE modeling is typically used when the key purpose of the analysis is to understand the impacts of policies (for instance, to understand how policy changes affect the market [47]. The most significant works in this area include the multi-commodity, partial equilibrium SWOPSIM framework [48], the Global Trade Analysis Project (GTAP) [49], the International Institute for Applied Systems Analysis (IIASA) Model [50], the GTAP-AGR model [51], the Common Agricultural Policy Regionalised Impact modelling system (CAPRI) [52], and the GARCH-MIDAS model to study the influence of geopolitical risk on the commodity markets [53]. Although those models are considered to be precise, they require a great deal of manual labor and thorough analysis to run. There are also a lot of indicators and scenarios, which are hard to analyze with such a framework, including small-scale impacts, social welfare indicators, non-market factors, etc.

In contrast, Agent-based modeling (ABM) is a quantitative framework that does not need to compute optima, meaning they are simpler to run and more tractable, allowing a higher level of realism. In this framework, agents (countries, etc.) make decisions using heuristics, myopic reasoning, and/or learning algorithms. Recent research has developed quantitative agent-based models that make time-series predictions, modeling a specific economy at a specific point in time. For example, a previous paper [54] focused on modeling trade wars considering four countries and the rest world as key participants. In [55], it was shown that ABMs can compete with VAR and DSGE models in out-of-sample trade forecasting. Unfortunately, ABMs are highly demanding of hardware resources, especially in the case of world-scale trade modeling, which is why those models are simulations by nature.

Autoregressive models can also be applied to counterfactual intervention studies. For example, in [56], they proposed a VAR model and formulated the inference of a causal model as a joint regression task, where, for inference, they used both data with and without interventions. In this paper, we stick to a similar approach, also inspired by multiple Masked language modeling frameworks in natural language processing, in which the model learns to restore masked tokens based on the existing ones.

In summary, this study addresses the limitations of existing wheat trade forecasting research, including the trade-offs between Fourier modules' frequency and time resolutions, the neglect of critical external factors (e.g., policies), and insufficient validation. This framework addresses the drawbacks of spectral features through adaptive graph attention, integrates key external variables to capture real-world market complexity, and enhances validation, thereby directly resolving the gaps in prior work.

3. Dataset

In our data collection, we integrated open information about wheat trade flows and production levels from the UN FAOSTAT and UN Comtrade databases [57,58] (Table 1). We obtained trading and production data for N countries and G products in the time period t . In this study, $G = 1$ (wheat), $t = 1993$ –2023. That period is used because the earlier indicators are not interpretable due to the large changes in the political world map from 1989 to 1993. Thus, an initial dataset was compiled on wheat exports from 245 countries over a period of 30 years. This study focuses more on how models can utilize spatial and temporal relationships to achieve better prediction accuracy; therefore, we deliberately neglected complex features such as climate or global economy indicators. We filtered the data to leave only significant wheat-exporting countries; therefore, the dataset contains information about the export volumes, with a mean value for the 1993–2023 period of at least 500,000 kilograms per year. After filtering, the number of remaining countries for analysis decreased to 211 ($N = 211$). Thus, the final dataset contains 6330 records.

Table 1. Feature set for the wheat export forecast.

Feature Set	Frequency	For	Features	Unit
Trade flows	Annual	Pair of countries	Export quantity Import quantity Re-export quantity Re-import quantity	Kilograms
Production	Annual	Country	Production quantity	Tonnes

We filtered all the unions and other non-country trade subjects from the dataset. The collected data has a small number of gaps, especially for the 2022–2023 period. We utilized several imputation techniques, including ARIMA, forward fill, linear interpolation, and

model-based imputation to fill these gaps. During the model-based imputation, we trained the same model that we tested (with the same hyperparameters) on the data that had been preprocessed with forward fill. Then, we used the trained model to fill the gaps. Then, we split the data into five-year-long non-overlapping ranges to obtain more time-series samples for training and validation. Finally, we normalized all features and targets to the interval $[0.1]$ to prevent overflows during model training and inference.

As preliminary studies with the Kolmogorov–Smirnov test showed that the target export values (except some extreme values close to zero) are expected to be Beta-distributed, we tested several losses (Figure 1). First, we applied the mean squared error (MSE) of the model output logit (presuming the logit is normally distributed) and binary cross-entropy (BCE) as the losses to train all the models. To motivate BCE loss, one may imagine a binary output Z , which has a Bernoulli distribution. The Beta-distributed model output Y would define a probability for Z . Thus, we have a special case of Beta-Binomial distribution, which collapses back to Bernoulli (see [59], page 17). We also tried to employ beta-regression loss in a straightforward manner [60], but the performance was very low. We believe this is due to optimization issues, because in our case, the Beta values and most of the outputs and targets remain close to 0.

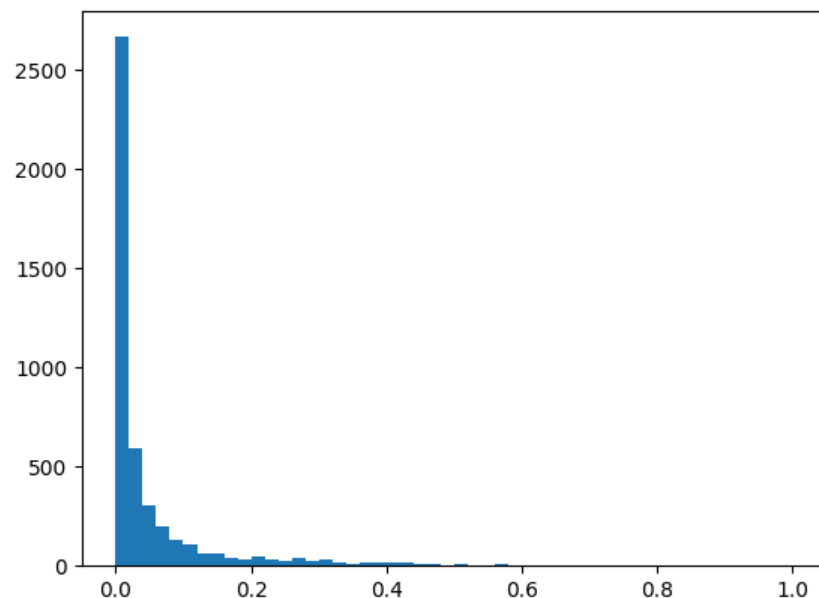


Figure 1. Distribution of the normalized export values in the dataset.

4. Forecast Models

4.1. Baseline Regression Models

In this study, we consider wheat export forecasting as a time-series prediction problem. In a food market, past-time indicators like the production level may have a great effect on future values. A common approach to solving this problem is to utilize a regression model, whereby the current output depends on the indicator values from previous time steps. The essential choice in this case is to use an autoregressive model, such as ARIMA, and a recurrent network. We utilized two well-known recurrent network architectures, namely, Long-Short Term Memory (LSTM) and the Gated Recurrent Unit (GRU), as the baselines [61,62]. These architectures apply gating mechanisms to solve the gradient-vanishing problem, which may help in catching long-term relationships between indicators and output values.

Besides LSTM and GRU, we utilized the Temporal Kolmogorov–Arnold Networks (TKANs) [28]. TKANs are recurrent extensions of the Kolmogorov–Arnold Networks, which have been empirically shown to be more accurate for many problems [27].

4.2. Recurrent Graph Transformer Model

The proposed hybrid regression model is based on a neural network with the Transformer architecture [63]. The transformer network allows one to consider the complex inter-relationships between the trade and production indicators and the predicted values for different countries. In this study, we extend the standard Graph Transformer architecture proposed in a previous paper [64]. The general idea of the Graph Transformer is to sequentially generate vector representations (embeddings) of graph vertices and edges, considering embeddings of their neighbors in the graph. As in the previous paper [43], we use temporal states for edges to create a recurrent model, i.e., to make the current vector representations independent of the previous time steps, but we encode the states in a hidden trainable embedding via a standard recurrent layer. In addition to the edge embeddings, we also make the vertex embeddings independent of the previous time steps. We use recurrent blocks (LSTM or GRU) to do so. In Figure 2b, h_{it} denotes an embedding for the vertex i at step t , and e_{ijt} denotes an embedding for the edge ij at step t . In our case, the graph transformer operates not with the observable trade indicators or output values but rather with the embeddings, which may be arbitrarily large to achieve the required complexity of the model. Another difference is that we use Fast Fourier Transformation (FFT) blocks to work with discrete spectral data instead of original values, which is preferable for Transformers. We perform FFT on the time series, with a length of 5 and a step value of 1; therefore, the frequency resolution is $1/5$.

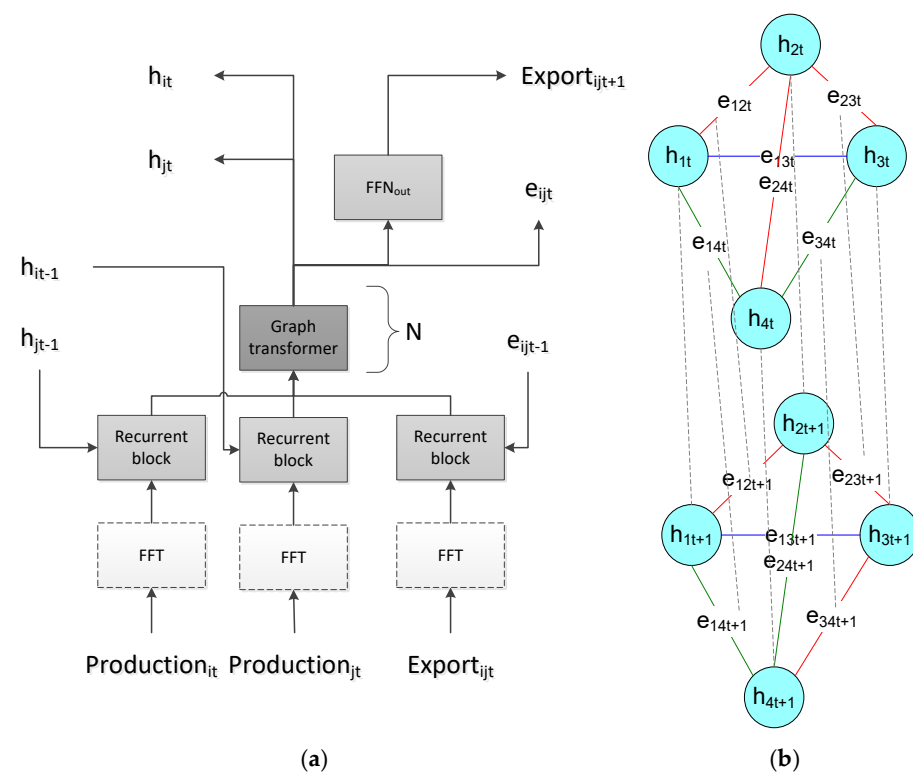


Figure 2. The hybrid wheat export forecast model. (a) The model architecture. (b) Recurrent dependencies (dashed lines) between hidden (latent) embeddings of the vertices (countries) and edges (trade flows, colored lines) at steps t and $t+1$.

Figure 2a shows the architecture of the proposed model with N graph transformer blocks. The model takes features of a pair of graph vertices i and j (production levels $Prod_i$ and $Prod_j$ in the countries i and j , and features of the edge between these vertices (export value $Export_{ij}$). First, we (optionally) apply the FFT blocks and add positional embeddings $\lambda_{1,2}$ to the obtained representations (Expressions (1), and (2)). As in the standard Graph Transformer, we use Laplacian eigenvectors to form the positional embeddings [38]. Then, we use recurrent blocks (RNN) to create embeddings of the vertex features $h_{i,t}^0, h_{j,t}^0$ of size h and edge features $e_{ij,t}^0$ of size e at each timestep t . We use output vertex and edge embeddings from the previous time step $h_{i,t-1}^N, e_{i,t-1}^N$ as the hidden state embeddings in the RNN (Expressions (3), and (4)).

$$Export_{ij}^{fft} = FFT(Export_{ij}) + \lambda_1. \quad (1)$$

$$Prod_i^{fft} = FFT(Prod_i) + \lambda_2. \quad (2)$$

$$h_{i,t}^0 = RNN(Prod_{i,t}^{fft}, h_{i,t-1}^N). \quad (3)$$

$$e_{ij,t}^0 = RNN(Export_{ij,t}^{fft} + \lambda, e_{i,t-1}^N). \quad (4)$$

Next, we iteratively use graph transformer blocks GT (Expression (5)) to form the output vertex and edge feature embeddings $h_{i,t}^N, h_{j,t}^N, e_{ij,t}^N$ [38]:

$$h_{i,t}^l, h_{j,t}^l, e_{ij,t}^l = GT(h_{i,t}^{l-1}, h_{j,t}^{l-1}, e_{ij,t}^{l-1}). \quad (5)$$

The model has N graph transformation blocks; each of them contains H heads to create linear projections of the inputs. Each transformer block works as follows: First, it applies a standard multi-head attention mechanism [63,64] to edge embeddings. Next, it uses the obtained edge attention weights to adjust vertex embeddings. Then, it processes the embeddings with a residual network. In the next step $t + 1$, these embeddings $h_{i,t}^N, h_{j,t}^N, e_{ij,t}^N$ are used to obtain $h_{i,t+1}^0, h_{j,t+1}^0, e_{ij,t+1}^0$ with RNN blocks. This way, the graph transformer helps catch spatial dependences, while the RNN learns temporal dependencies (Figure 2b).

Finally, we utilize a multilayer feed-forward network FFN_{out} with hidden *relu* activations and *sigmoid* output activation to predict next export values $Export_{ij,t+1}$ (Expression (6)) based on the edge embeddings $e_{ij,t}^N$.

$$Export_{ij,t+1} = FFN_{out}(e_{ij,t}^N). \quad (6)$$

We use a dropout mechanism to regularize the networks. In the next sections, we will refer to this model as “encoder”.

4.3. Recurrent Graph Transformer Encoder–Decoder Model

We enhance the proposed export forecast model with a decoder to provide conditional forecasting. It might be helpful when, for example, one needs to predict the next export values in a cluster of countries if the next year’s export between some of them is limited to 0. During training, the decoder takes the current production and future export values $Prod_{i,t}, Prod_{j,t}, Export_{ij,t+1}$, together with the current edge and vertex embeddings from the encoder, as inputs $e_{ij,t}^N, h_{i,t}^N, h_{j,t}^N$ (Figure 3). In contrast to the encoder, we use a simple linear feed-forward layer (Expressions (7), and (8)) instead of the RNN block to create edge and vertex embeddings $h_{i,t}^d, h_{j,t}^d$, and $e_{ij,t}^d$ (in this description, we deliberately omit FFT/inverse FFT blocks for simplicity).

$$h_{i,t}^d = FFN_{in}(Prod_{i,t} + \lambda) \quad (7)$$

$$e_{ij,t}^d = FFN_{in}(Export_{ij,t+1} + \lambda) \quad (8)$$

where $d > N$.

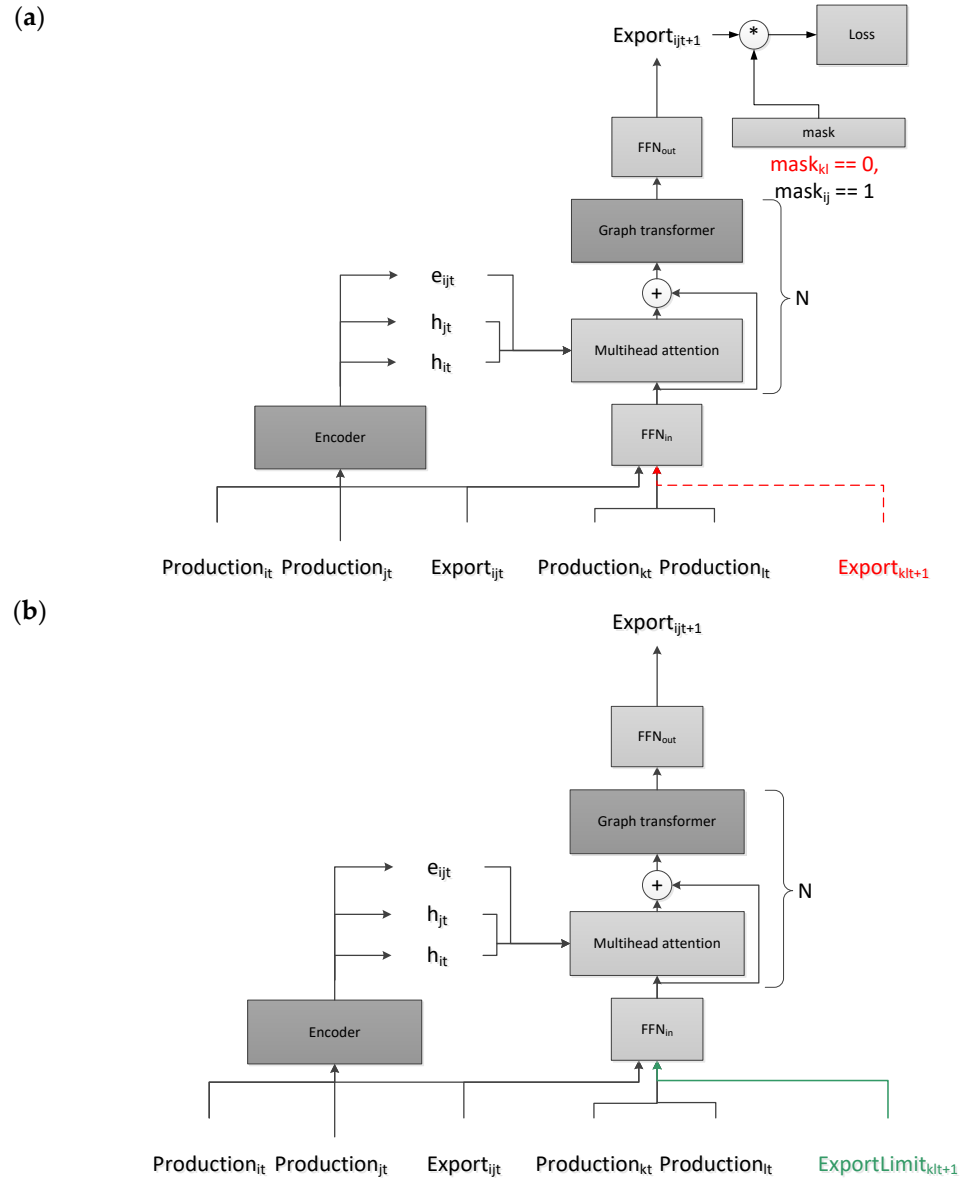


Figure 3. Encoder–decoder approach to ensure predefined constraints on the predicted export values. (a) During training, the decoder takes export and production values from the current step t as input (for example, production in countries i, j , and export between them), but for some randomly picked countries k and l , it takes export for the future step $t + 1$. The loss for these k, l pairs is masked (red dashed line). (b) During inference, the decoder takes current export and production values, but for some countries k and l , it also takes our predefined limitations (for example, it takes the limit value for the export from k to l , green line).

Then, we apply M blocks with the following structure: The block utilizes a standard multihead attention $Attention(Q, K, V)$ between the decoder vertex and edge embeddings $h_{i,t}^d, h_{j,t}^d$, and $e_{ij,t}^d$ (as a query Q), and encoder embeddings $e_{ij,t}^N, h_{i,t}^N, h_{j,t}^N$ as keys K and values V (Expressions (9), and (10)).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (9)$$

$$H_t^{d+1} = \text{Attention}(H_t^d, H_t^N) + H_t^d, E_t^{d+1} = \text{Attention}(E_t^d, E_t^N) + E_t^d, \quad (10)$$

where H_t^d is the decoder vertex embeddings of a country cluster, H_t^N is the encoder vertex embeddings of the country cluster, E_t^d is the decoder edge embeddings of the country cluster, and E_t^N is the encoder edge embeddings of the country cluster. After that, the block applies a graph attention, as in the encoder.

Finally, as in the encoder, we utilize a multilayer feed-forward network FFN with hidden *relu* activations and *sigmoid* output activation to predict the next export values.

During training, we set some randomly picked decoder's export inputs for the countries k and l to $\text{Export}_{kl,t}$ (e.g., to the values of the export from k to l for the current period t , not for the future period $t + 1$). As in BERT [65], a mask is used to hide all the decoder outputs related to the future export inputs $\text{Export}_{ij,t+1}$ during loss estimation. The best results in tests have been obtained when we masked half (50%) of the decoder export inputs.

5. Experiment Results

First, we separate the dataset into the train and validation subsets. We use all the data from 1993 to 2017 as the training dataset, and the data from 2018–2022 as the test dataset. It is worth noting that the test dataset does not have any imputed target values. We use cross-validation on the training subset to find the hyper-parameters (dropout level, number of layers, size of the hidden layers, and learning rate) using the Optuna library. All the networks have been trained on a relatively large epoch number (1000). For the ARIMA model (both for training and for data imputation), we used the AutoARIMA algorithm, which conducts the Kwiatkowski–Phillips–Schmidt–Shin differencing test to define the order of first-differencing, and then finds the model that optimizes the Akaike Information Criterion. We applied the mean squared error MSE , mean average error MAE , $MAPE$, and coefficient of determination R^2 as the validation scores (Expressions (11), (12), (13), and (14)). We used two $MAPE$ scores: $MAPE_{all}$ is evaluated on the whole test dataset, while $MAPE_{large}$ covers only the exports, which are greater than 5% of the maximum wheat export between any two countries in 1993–2023. We repeated all the tests on random subsets of the training and testing datasets to obtain standard deviations of the scores.

$$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - Y_i^*)^2 \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - Y_i^*| \quad (12)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{Y_i - Y_i^*}{Y_i} \right| \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (Y_i - Y_i^*)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} \quad (14)$$

Appendix A contains all the detailed scores and additional comments for different models and imputation parameters (ablation study). Table 2 shows the results for one-year-ahead export predictions with the baseline ARIMA model and recurrent networks on the test subset. The results confirm that recurrent neural network models significantly outperform ARIMA on that task. Namely, LSTM achieves the best R^2 score of 0.80. We believe this is because LSTM has the most complex gating mechanism to deal with gradient vanishing. The novel TKAN model also shows a competitive R^2 score of 0.79.

Table 2. Average wheat export trade forecasting errors for the baseline methods.

Model	MSE, $\times 10^{17} \text{ kg}^2$	MAE, $\times 10^8 \text{ kg}$	MAPE _{all}	MAPE _{large}	R ²
ARIMA	1.99 $\pm$ 0.13	1.82 $\pm$ 0.09	(3.03 $\pm$ 0.03) $\times 10^{21}$	0.46 $\pm$ 0.02	0.72 $\pm$ 0.03
GRU	1.57 $\pm$ 0.57	1.53 $\pm$ 0.18	13.88 $\pm$ 7.52	0.33 $\pm$ 0.02	0.77 $\pm$ 0.04
LSTM	1.37 $\pm$ 0.34	1.52 $\pm$ 0.15	5.97 $\pm$ 3.47	0.32 $\pm$ 0.02	0.80 $\pm$ 0.03
TKAN	1.47 $\pm$ 0.35	1.52 $\pm$ 0.13	5.55 $\pm$ 2.71	0.32 $\pm$ 0.02	0.79 $\pm$ 0.03

The most accurate results are highlighted with bold font.

Representing all countries from the dataset as a single graph would make attention weight matrices too large; therefore, the models would tend to overfit. Thus, we split all the data into clusters and associate each cluster with one or several large wheat exporters and a set of countries–importers. Since the initial data for the clustering is generally a relatively small, weighted graph, we applied the Affinity propagation algorithm, which can work with such data out of the box and provide stable results [66]. We have set 10 top wheat exporters as the preference parameter (“exemplars”, representatives of clusters), and as a result, obtained 10 clusters, of which the sizes vary from 4 to 15 countries. Figure 4 shows some examples of such clusters, which are related to the world’s top wheat exporters.

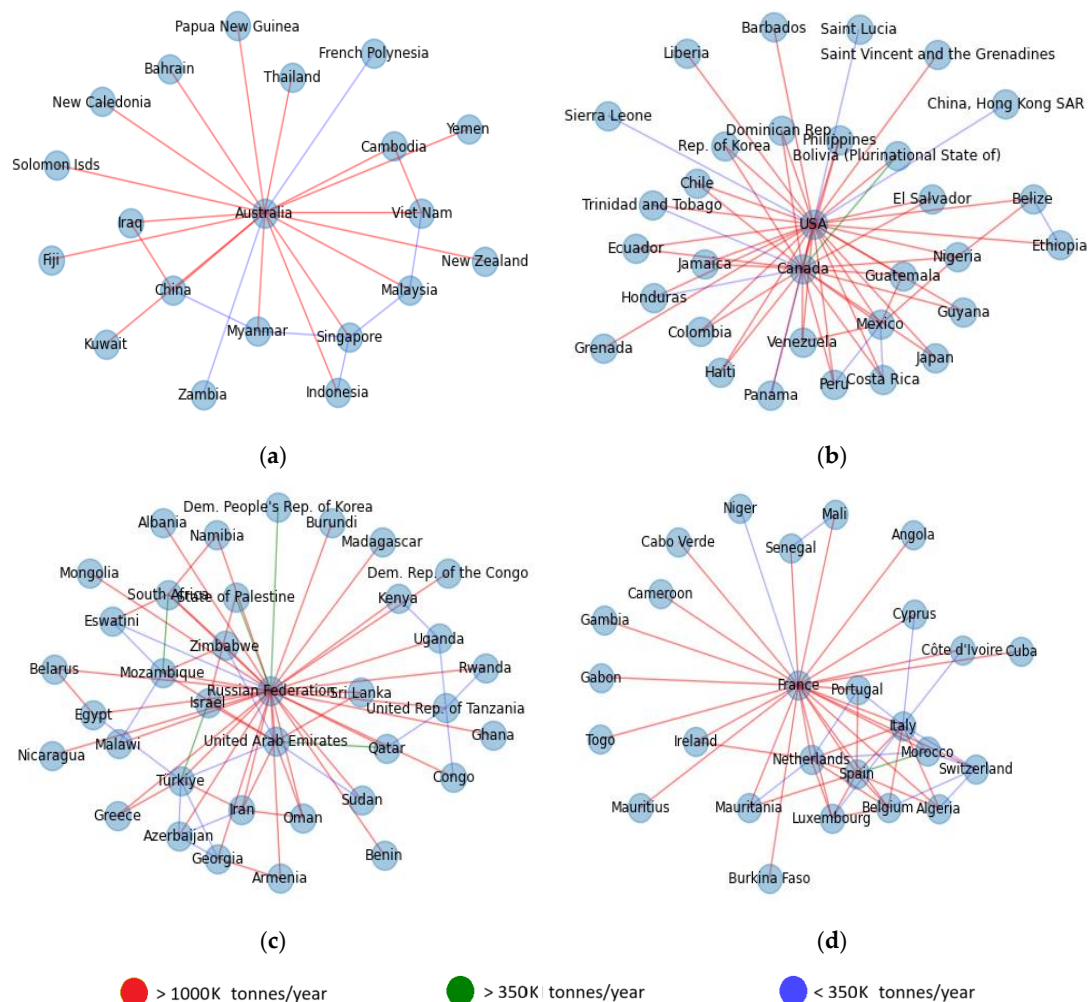


Figure 4. Wheat trade clusters for some top exporters. Vertices are countries, and edges are sustained trade flows. Edges are colorized as follows: red denotes trade flows that are above average, blue denotes trade flows that are below average, and green denotes trade flows that are close to average. (a) Australia; (b) USA; (c) Russian Federation; (d) France.

Each cluster forms a graph to be processed with the Transformer-based models. Table 3 shows the scores of the proposed models on the test dataset. The recurrent graph transformer (in the single encoder configuration) achieved the best results in terms of the R^2 score, significantly outperforming standard recurrent network-based models. The encoder–decoder model shows similar results. Introducing spectral features (with the *FFT* block) led to a very slight improvement in the modeling scores. We believe this might be related to the small block length; therefore, in the future, we will try FFT before splitting the dataset into time windows.

Table 3. Average wheat export trade forecasting errors for the proposed model.

Model	MSE, $\times 10^{17}$ kg ²	MAE, $\times 10^8$ kg	MAPE _{all}	MAPE _{large}	R ²
Recurrent graph transformer	1.01 ± 0.75	1.26 ± 0.34	1.35 ± 0.10	0.31 ± 0.04	0.85 ± 0.09
Recurrent graph transformer + spectral features	0.79 ± 0.39	1.13 ± 0.26	6.32 ± 3.95	0.25 ± 0.05	0.88 ± 0.05
Recurrent graph transformer (Encoder-Decoder)	0.71 ± 0.16	1.13 ± 0.12	17.46 ± 2.06	0.26 ± 0.05	0.88 ± 0.04

The most accurate results are highlighted with bold font.

Then, we analyzed the ability of the encoder–decoder model to deal with the assumed limitation on the sale of grain to one of the importers. Trade flows between Australia and China serve as a special case to illustrate the features of the model and demonstrate experimental results. We have selected Australia as the largest player in the global wheat grain market to perform calculations with the encoder–decoder model. According to the Comtrade database, the share of global wheat trade in Australia in 2022 was 13%, and it has the potential to expand exports at a level of +4.8% since, for the 2017–2022 period, it was classified as a fast-growing market segment. In many ways, Australia’s export potential was facilitated by the deregulation of the grain market in 2008 and the subsequent reform of the grain trading market. The key was the mandatory accreditation of exporters in accordance with the Wheat Export Marketing Act 2008, which establishes strict quality requirements, including compliance with international phytosanitary norms and product safety standards. Australia has intensified the modernization of logistics infrastructure and the implementation of innovations in logistics to optimize the operation of port terminals, introducing the Wheat Port Code of Conduct to standardize it. This has increased the transparency of capacity allocation and reduced downtime, which is critical to meeting delivery times to key Asian markets (Indonesia, Vietnam, etc.). Australia is focused on reforming its grain export supply chains and reviewing the approaches to selecting exporters to remain competitive in the global grain market.

We have performed a retrospective forecast on the 2018–2022 interval to check two scenarios with the encoder–decoder model:

- Forecast of wheat grain exports without any limitations.
- Forecast of wheat grain exports, providing the limitation on the sale of grain to one of the importers.

We selected a wheat-importing country for the forecast calculations based on a combination of two parameters: (1) in the structure of Australia’s exports, the volume of wheat grain trade should be at least 5% of the total export volume (to exclude the importers with insignificant volumes); (2) the country should have a growing demand for wheat grain. Thus, the conditions above describe China’s import of wheat grain. In 2018, China’s share of Australian wheat exports was 6.1%. At the same time, for the 2017–2022 period,

according to Comtrade, China was classified as a fast-growing market segment that has the potential to expand exports at a level of +16%. Figure 5 shows the modeling results for both scenarios.

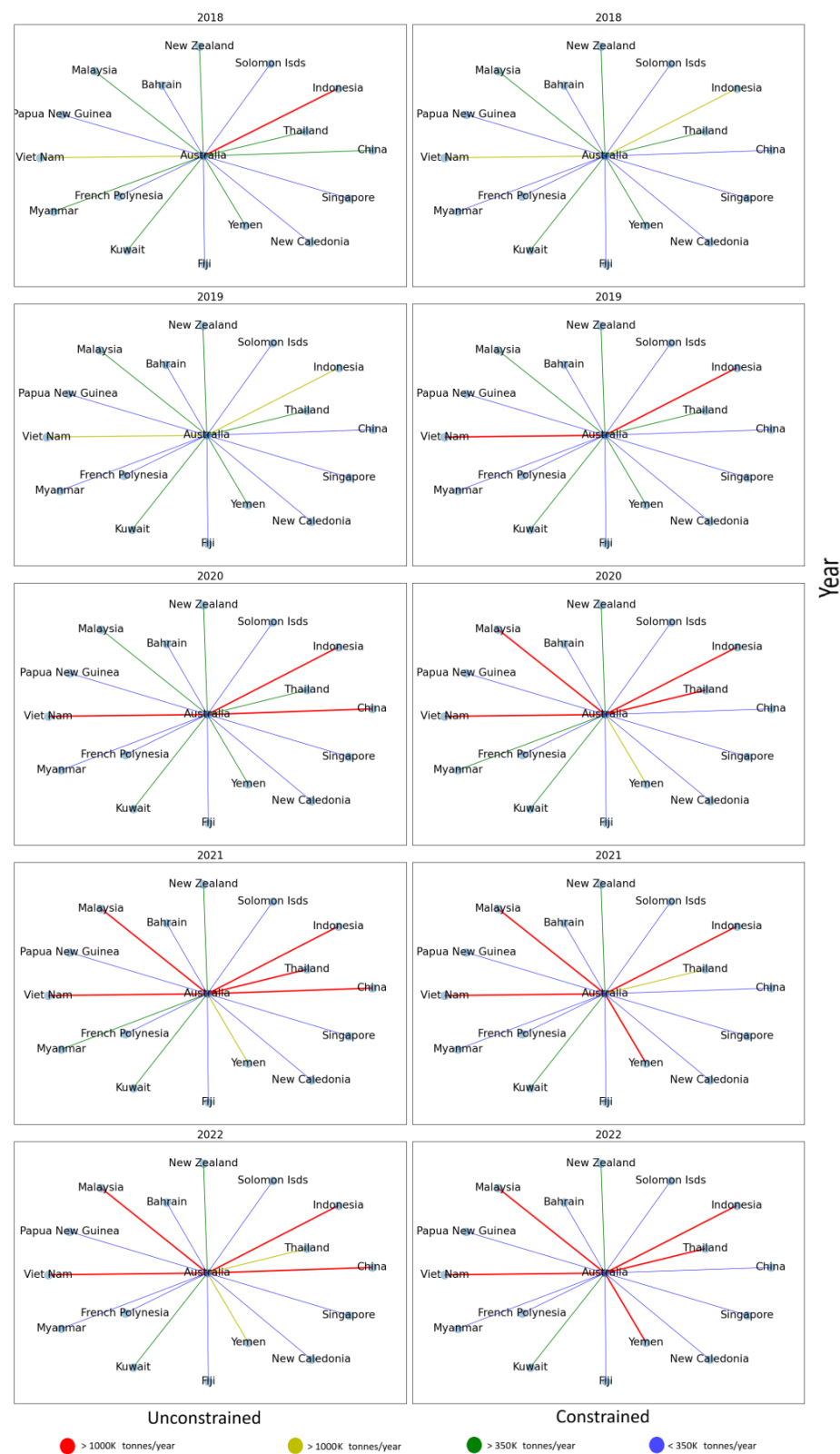


Figure 5. Modeling results for the Australian wheat export dynamics in 2018–2022. Left column shows results without any limitations; right column shows the scenario where export to China is set to zero.

The left column of the picture is the forecast of exports without restrictions, and the right column is the forecast of exports subject to restrictions on trade flows between Australia and China. For clarity of the trade flow assessment in the diagrams, the maximum volume of trade flows is colored in red (more than 20%), medium in yellow (at the level of 15–20%), low in green (5–15%), and very low in blue (less than 5%).

As part of the study, we considered a hypothetical example: the Australia–China zero-export experiment to assess possible risks and forecast export destinations when trade restrictions (sanctions) are imposed. Given that this situation does not exist in the real world, it is impossible to validate scenario forecasting using real historical trade data. However, due to the volatility of international trade, any scenario cannot be completely ruled out.

6. Discussion

The structural and dynamic analysis shows the resilience of the Australian market to the introduction of sanctions, including the restoration of the market concentration volume over the forecast period. The ratio of the export volume from the model without limitations to the export volume from the model with the trade restrictions shows the possibilities for restoring export flows. We propose a coefficient of the export recovery potential, RecoveryExport (RE):

$$RE = \frac{V_{exportTM}}{V_{exportTMs}} \quad (15)$$

where $V_{exportTM}$ is the export volume from the model without any limitations and $V_{exportTMs}$ is the export volume from the model, which introduces the limitations.

The modeling shows that by the end of the forecast period (2022), the RE coefficient was 0.74, which allows us to discuss the restoration of Australian exports after the introduction of restrictions on trade with China. In addition, the model provides feasible results, redistributing flows among importers whose import volumes are significant and ignoring ones who cannot meet the growing demand for wheat (island states, etc.). When adjusting trade flows according to the model with restrictions on trade with China, Australia's export destinations could be redirected, with an increase in the volume of deliveries; the bulk of deliveries could go to Indonesia (28.1%), Thailand (14.8%), Vietnam (20.3%), and Yemen (11.15%).

Given the roughness of the tested models, this example, of course, should be considered a demo to show the model's potential. In the future, we shall find appropriate retrospective trade shock scenarios to perform a valid counterfactual study.

7. Conclusions

This study tackles the critical challenges of modern agrifood supply chains, characterized by globally interconnected networks marked by escalating complexity and interdependence. The main actors of the agrifood market have to place greater emphasis on modern methods of forecasting supplier networks to streamline supply chain operations and enhance overall performance. The high market concentration of key commodities such as wheat amplifies vulnerabilities in trade-restricted environments, where supplier selection directly impacts global food security.

This paper introduces a novel recurrent transformer architecture, extending graph transformer frameworks to temporal data analysis in supply networks. Unlike existing models, the model includes hidden temporal states at the vertex level, explicitly encoding the dynamic production capacities of national actors of the wheat grain market. This approach enables key advancements: (1) capturing time-dependent patterns, such as seasonal production fluctuations and long-term yield trends, allows the model to enhance its predic-

tive accuracy in volatile markets; and (2) considering interconnections and dependencies between agrifood importers and exporters via the graph attention mechanism.

Another feature is that the model follows an encoder–decoder architecture, which enables, for instance, the ability to simulate alternative supplier configurations under trade restrictions. This provides insights for policymakers and agribusinesses, particularly in mitigating cascading risks from supplier concentration in grain markets. In contrast to equilibrium or agent-based approaches, the proposed model is simple to obtain (as it infers the required relationships from data) and run because it does not need hardware-demanding numerical simulations.

In the future, we will test the model in multi-agent environments integrated with behavioral economic factors (e.g., supplier risk preferences, policy response biases) to refine its predictive power under market shocks and perform counterfactual studies on retrospective data. However, critical questions remain to guide further exploration: How might quantifying dynamic trust and collaboration patterns among suppliers alter the model’s ability to forecast the redistribution of export flows after trade restrictions? Can the framework be adapted to disentangle the overlapping impacts of climate extremes and geopolitical tensions, two major drivers of wheat supply chain volatility rarely modeled together? How might incorporating sub-national production data (e.g., regional yield variations within export countries) enhance the granularity of resilience assessments? Beyond these questions, the results of this study confirm the proposed model’s accuracy in forecasting resilient wheat supply chains amid global uncertainties. Its novel hybrid architecture also contributes to the discourse on sustainable agrifood systems by aligning advanced forecasting methods with supply chain governance needs.

Author Contributions: Conceptualization and methodology, Y.O. and H.Z.; investigation and validation, D.D.; writing—review and editing, D.D., H.Z., and Y.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the Russian Science Foundation, project No. 25-28-00810. The research was also supported by the State assignment of the Ministry of Science and Higher Education of the Russian Federation, project “Software and analytical complexes for increasing the efficiency of management decisions in the Russian Federation” No. FMGF-2025-0001, No. 125031003392-3.

Data Availability Statement: The dataset is available at https://github.com/masterdoors/agro_graph_transformer (accessed on 5 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1 provides the performance scores for all the baseline models. All of the runs for all scenarios below have been repeated five times on random subsets (70% of the original sets, without repeats) of the training and testing data; *std* scores are provided. There is no significant difference between different imputation strategies, but the forward fill is slightly behind the others. We provide two *MAPE* scores: $MAPE_{all}$ is evaluated on the whole test dataset, while $MAPE_{large}$ covers only the exports, which are greater than 5% of the maximum wheat export between any two countries in 1993–2023 (i.e., it ignores all of the blue lines in Figure 5). One can note that the $MAPE_{all}$ scores are poor for all the considered combinations (the best result is 2.28), while the $MAPE_{large}$ scores are significantly better (the minimum error score is 0.31). This means the models cannot predict small exports between countries accurately, but perform better for large importers/exporters. It is worth noting

that the use of BCE loss tends to improve R^2 and $MAPE_{large}$ scores, but $MAPE_{all}$ becomes arbitrarily large.

Table A1. Performance scores for baseline forecasting models with various hyperparameters and imputation.

Model	Imputation	Loss	MSE, $\times 10^{17}$, kg ²	MAE, $\times 10^8$, kg	MAPE _{all}	MAPE _{large}	R ²
TKAN	Forward fill	BCE	1.38 $\pm$ 0.40	1.73 $\pm$ 0.58	41.80 $\pm$ 71.80	0.35 $\pm$ 0.09	0.76 $\pm$ 0.11
TKAN	Forward fill	MSE(logit)	3.41 $\pm$ 3.02	2.14 $\pm$ 1.10	6.47 $\pm$ 9.36	0.43 $\pm$ 0.23	0.58 $\pm$ 0.36
LSTM	Forward fill	BCE	1.76 $\pm$ 0.18	1.57 $\pm$ 0.06	10.28 $\pm$ 3.54	0.36 $\pm$ 0.02	0.73 $\pm$ 0.07
LSTM	Forward fill	MSE(logit)	1.97 $\pm$ 0.24	1.76 $\pm$ 0.17	2.56 $\pm$ 0.14	0.40 $\pm$ 0.03	0.74 $\pm$ 0.03
GRU	Forward fill	BCE	1.57 $\pm$ 0.57	1.53 $\pm$ 0.18	13.88 $\pm$ 7.52	0.33 $\pm$ 0.02	0.77 $\pm$ 0.04
GRU	Forward fill	MSE(logit)	2.33 $\pm$ 0.71	1.85 $\pm$ 0.24	2.64 $\pm$ 0.72	0.46 $\pm$ 0.02	0.65 $\pm$ 0.04
TKAN	Interpolation	BCE	1.61 $\pm$ 0.47	1.60 $\pm$ 0.16	7.06 $\pm$ 2.53	0.33 $\pm$ 0.02	0.78 $\pm$ 0.05
TKAN	Interpolation	MSE(logit)	1.85 $\pm$ 0.53	1.53 $\pm$ 0.19	2.28 $\pm$ 0.66	0.31 $\pm$ 0.02	0.74 $\pm$ 0.05
LSTM	Interpolation	BCE	1.37 $\pm$ 0.34	1.52 $\pm$ 0.15	5.97 $\pm$ 3.47	0.32 $\pm$ 0.02	0.80 $\pm$ 0.03
LSTM	Interpolation	MSE(logit)	2.31 $\pm$ 0.55	1.87 $\pm$ 0.19	3.02 $\pm$ 1.31	0.43 $\pm$ 0.06	0.67 $\pm$ 0.09
GRU	Interpolation	BCE	1.53 $\pm$ 0.45	1.53 $\pm$ 0.15	11.34 $\pm$ 5.31	0.33 $\pm$ 0.03	0.78 $\pm$ 0.05
GRU	Interpolation	MSE(logit)	2.54 $\pm$ 1.03	1.93 $\pm$ 0.32	3.19 $\pm$ 1.71	0.48 $\pm$ 0.08	0.65 $\pm$ 0.11
TKAN	Model-based	BCE	1.71 $\pm$ 0.29	1.60 $\pm$ 0.13	7.91 $\pm$ 2.32	0.33 $\pm$ 0.01	0.77 $\pm$ 0.04
TKAN	Model-based	MSE(logit)	3.23 $\pm$ 2.09	2.12 $\pm$ 0.93	10.59 $\pm$ 16.22	0.44 $\pm$ 0.21	0.52 $\pm$ 0.33
LSTM	Model-based	BCE	1.23 $\pm$ 0.39	1.54 $\pm$ 0.20	13.67 $\pm$ 6.05	0.34 $\pm$ 0.03	0.80 $\pm$ 0.03
LSTM	Model-based	MSE(logit)	2.48 $\pm$ 0.42	2.01 $\pm$ 0.16	2.29 $\pm$ 0.92	0.51 $\pm$ 0.03	0.64 $\pm$ 0.04
GRU	Model-based	BCE	1.94 $\pm$ 0.22	1.79 $\pm$ 0.12	13.39 $\pm$ 3.22	0.35 $\pm$ 0.02	0.76 $\pm$ 0.02
GRU	Model-based	MSE(logit)	2.14 $\pm$ 0.48	1.80 $\pm$ 0.21	3.27 $\pm$ 0.93	0.46 $\pm$ 0.10	0.70 $\pm$ 0.06
TKAN	ARIMA	BCE	1.47 $\pm$ 0.35	1.52 $\pm$ 0.13	5.55 $\pm$ 2.71	0.32 $\pm$ 0.02	0.79 $\pm$ 0.03
TKAN	ARIMA	MSE(logit)	2.32 $\pm$ 1.88	1.70 $\pm$ 0.76	3.58 $\pm$ 5.64	0.35 $\pm$ 0.12	0.68 $\pm$ 0.21
LSTM	ARIMA	BCE	1.39 $\pm$ 0.38	1.49 $\pm$ 0.13	7.61 $\pm$ 4.22	0.33 $\pm$ 0.03	0.79 $\pm$ 0.06
LSTM	ARIMA	MSE(logit)	2.67 $\pm$ 0.64	1.99 $\pm$ 0.24	2.66 $\pm$ 1.08	0.47 $\pm$ 0.05	0.63 $\pm$ 0.06
GRU	ARIMA	BCE	1.53 $\pm$ 0.41	1.58 $\pm$ 0.18	10.17 $\pm$ 3.56	0.34 $\pm$ 0.03	0.77 $\pm$ 0.05
GRU	ARIMA	MSE(logit)	2.50 $\pm$ 0.57	1.89 $\pm$ 0.20	2.63 $\pm$ 0.98	0.46 $\pm$ 0.05	0.65 $\pm$ 0.06

Table A2 shows the ablation study results for the proposed hybrid model. The primary conclusions regarding imputation strategies and losses are the same as for baselines. The best $MAPE_{all}$ score is 0.87, which is achieved for MSE loss; the best $MAPE_{large}$ is 0.25 (BCE loss). The best R^2 is 0.88 (BCE loss). Models with a large number of attention heads (2, 4) and a small number of layers (1, 2) tend to show better results. Replacing the recurrent network with a linear layer (i.e., ignoring hidden states) leads to total degradation of the model performance. Enabling the FFT block to encode inputs leads to a hardly noticeable performance improvement.

Table A2. Performance scores for the proposed recurrent graph transformer forecasting model with various hyperparameters and imputation.

Imputation	Loss	Layers	Attn. Heads	FFT	Recurrent Embeddings	MSE, $\times 10^{17}$, kg ²	MAE, $\times 10^8$, kg	MAPE _{all}	MAPE _{large}	R ²
Forward fill	MSE(logit)	1	1	+	+	1.60 $\pm$ 0.37	1.48 $\pm$ 0.18	1.03 $\pm$ 0.13	0.34 $\pm$ 0.08	0.74 $\pm$ 0.08
Forward fill	BCE	1	1	+	+	1.03 $\pm$ 0.46	1.28 $\pm$ 0.22	12.94 $\pm$ 5.73	0.30 $\pm$ 0.05	0.84 $\pm$ 0.06
Forward fill	MSE(logit)	1	1	+	-	6.68 $\pm$ 1.46	3.61 $\pm$ 0.40	3.40 $\pm$ 1.22	0.69 $\pm$ 0.08	0.15 $\pm$ 0.13
Forward fill	BCE	1	1	+	-	6.01 $\pm$ 2.07	4.41 $\pm$ 0.42	93.55 $\pm$ 97.00	0.59 $\pm$ 0.07	0.20 $\pm$ 0.15
Forward fill	MSE(logit)	1	2	+	+	2.18 $\pm$ 0.62	1.73 $\pm$ 0.24	1.04 $\pm$ 0.18	0.38 $\pm$ 0.05	0.69 $\pm$ 0.08
Forward fill	BCE	1	2	+	+	1.01 $\pm$ 0.40	1.32 $\pm$ 0.15	13.37 $\pm$ 12.45	0.28 $\pm$ 0.02	0.86 $\pm$ 0.04
Forward fill	MSE(logit)	1	2	+	-	6.22 $\pm$ 1.83	3.50 $\pm$ 0.55	3.92 $\pm$ 1.41	0.70 $\pm$ 0.07	0.25 $\pm$ 0.13
Forward fill	BCE	1	2	+	-	7.66 $\pm$ 1.28	5.22 $\pm$ 0.13	158.10 $\pm$ 65.71	0.52 $\pm$ 0.05	-0.02 $\pm$ 0.02

Table A2. Cont.

Imputation	Loss	Layers	Attn. Heads	FFT	Recurrent Embeddings	$MSE, \times 10^{17}, \text{kg}^2$	$MAE, \times 10^8, \text{kg}$	$MAPE_{all}$	$MAPE_{large}$	R^2
Forward fill	MSE(logit)	1	4	+	+	2.63 ± 1.00	1.70 ± 0.50	1.21 ± 0.26	0.44 ± 0.05	0.57 ± 0.10
Forward fill	BCE	1	4	+	+	0.79 ± 0.39	1.13 ± 0.26	6.32 ± 3.95	0.25 ± 0.05	0.88 ± 0.05
Forward fill	MSE(logit)	1	4	+	-	5.97 ± 1.61	3.31 ± 0.49	5.69 ± 2.05	0.71 ± 0.09	0.15 ± 0.16
Forward fill	BCE	1	4	+	-	5.33 ± 1.59	4.18 ± 0.48	82.81 ± 75.30	0.57 ± 0.07	0.26 ± 0.14
Forward fill	MSE(logit)	2	1	+	+	2.71 ± 1.11	1.69 ± 0.41	1.11 ± 0.18	0.37 ± 0.03	0.59 ± 0.12
Forward fill	BCE	2	1	+	+	1.25 ± 0.45	1.33 ± 0.32	8.34 ± 6.38	0.30 ± 0.03	0.82 ± 0.02
Forward fill	MSE(logit)	2	1	+	-	5.43 ± 1.59	3.20 ± 0.38	4.32 ± 1.71	0.67 ± 0.04	0.31 ± 0.15
Forward fill	BCE	2	1	+	-	6.53 ± 1.98	4.12 ± 0.54	84.29 ± 110.03	0.59 ± 0.07	0.24 ± 0.14
Forward fill	MSE(logit)	2	2	+	+	2.11 ± 1.13	1.69 ± 0.57	1.17 ± 0.32	0.37 ± 0.08	0.69 ± 0.21
Forward fill	BCE	2	2	+	+	1.83 ± 0.57	1.69 ± 0.23	8.14 ± 2.85	0.32 ± 0.04	0.76 ± 0.06
Forward fill	MSE(logit)	2	2	+	-	6.84 ± 3.34	3.54 ± 0.80	4.81 ± 2.68	0.68 ± 0.07	0.20 ± 0.14
Forward fill	BCE	2	2	+	-	5.17 ± 1.32	3.59 ± 0.68	102.70 ± 88.77	0.60 ± 0.02	0.25 ± 0.19
Forward fill	MSE(logit)	2	4	+	+	1.54 ± 0.86	1.43 ± 0.26	0.87 ± 0.06	0.32 ± 0.03	0.79 ± 0.12
Forward fill	BCE	2	4	+	+	0.84 ± 0.32	1.15 ± 0.21	6.48 ± 2.33	0.27 ± 0.05	0.86 ± 0.04
Forward fill	MSE(logit)	2	4	+	-	4.70 ± 1.50	2.81 ± 0.53	4.81 ± 1.39	0.67 ± 0.05	0.23 ± 0.12
Forward fill	BCE	2	4	+	-	6.13 ± 1.59	3.81 ± 0.56	47.83 ± 25.29	0.63 ± 0.05	0.11 ± 0.07
Forward fill	MSE(logit)	3	1	+	+	2.30 ± 0.79	1.68 ± 0.31	1.07 ± 0.08	0.34 ± 0.04	0.70 ± 0.05
Forward fill	BCE	3	1	+	+	1.90 ± 0.64	1.62 ± 0.29	5.52 ± 3.09	0.32 ± 0.02	0.78 ± 0.04
Forward fill	MSE(logit)	3	1	+	-	5.10 ± 0.47	3.09 ± 0.34	4.33 ± 1.30	0.67 ± 0.11	0.35 ± 0.02
Forward fill	BCE	3	1	+	-	5.15 ± 1.20	3.57 ± 0.51	29.56 ± 9.89	0.61 ± 0.07	0.26 ± 0.12
Forward fill	MSE(logit)	3	2	+	+	2.82 ± 0.47	1.75 ± 0.26	1.14 ± 0.31	0.38 ± 0.05	0.57 ± 0.05
Forward fill	BCE	3	2	+	+	0.95 ± 0.36	1.13 ± 0.20	6.73 ± 2.38	0.28 ± 0.05	0.85 ± 0.06
Forward fill	MSE(logit)	3	2	+	-	4.35 ± 0.63	2.98 ± 0.32	3.24 ± 0.71	0.66 ± 0.08	0.40 ± 0.05
Forward fill	BCE	3	2	+	-	4.45 ± 0.76	3.24 ± 0.19	28.82 ± 6.16	0.60 ± 0.06	0.29 ± 0.10
Forward fill	MSE(logit)	3	4	+	+	1.61 ± 0.99	1.48 ± 0.35	0.98 ± 0.15	0.31 ± 0.07	0.78 ± 0.14
Forward fill	BCE	3	4	+	+	1.55 ± 0.68	1.41 ± 0.36	12.02 ± 8.30	0.36 ± 0.12	0.77 ± 0.09
Forward fill	MSE(logit)	3	4	+	-	4.93 ± 1.37	3.19 ± 0.55	3.91 ± 1.08	0.67 ± 0.09	0.31 ± 0.09
Forward fill	BCE	3	4	+	-	6.69 ± 0.78	4.12 ± 0.28	40.27 ± 22.37	0.60 ± 0.06	0.11 ± 0.12
Forward fill	MSE(logit)	4	1	+	+	1.36 ± 0.42	1.35 ± 0.23	0.90 ± 0.05	0.33 ± 0.02	0.77 ± 0.06
Forward fill	BCE	4	1	+	+	1.15 ± 0.46	1.38 ± 0.29	5.12 ± 2.22	0.29 ± 0.03	0.85 ± 0.04
Forward fill	MSE(logit)	4	1	+	-	5.16 ± 1.71	3.24 ± 0.55	10.54 ± 11.46	0.62 ± 0.04	0.24 ± 0.06
Forward fill	BCE	4	1	+	-	3.78 ± 1.04	3.06 ± 0.46	27.06 ± 10.96	0.57 ± 0.05	0.37 ± 0.04
Forward fill	MSE(logit)	4	2	+	+	2.82 ± 0.97	1.68 ± 0.23	1.17 ± 0.10	0.35 ± 0.05	0.62 ± 0.09
Forward fill	BCE	4	2	+	+	1.03 ± 0.21	1.33 ± 0.14	7.41 ± 1.35	0.30 ± 0.02	0.84 ± 0.03
Forward fill	MSE(logit)	4	2	+	-	5.10 ± 1.73	2.94 ± 0.47	9.61 ± 6.62	0.66 ± 0.06	0.28 ± 0.18
Forward fill	BCE	4	2	+	-	5.52 ± 1.25	3.71 ± 0.45	46.39 ± 31.38	0.62 ± 0.07	0.20 ± 0.12
Forward fill	MSE(logit)	4	4	+	+	112.40 ± 3.00	32.89 ± 0.56	467.51 ± 67.11	3.38 ± 0.12	-17.69 ± 4.11
Forward fill	BCE	4	4	+	+	1.96 ± 1.54	1.95 ± 1.06	19.35 ± 32.26	0.32 ± 0.12	0.71 ± 0.29
Forward fill	MSE(logit)	4	4	+	-	4.95 ± 1.32	3.01 ± 0.44	4.65 ± 1.97	0.66 ± 0.05	0.34 ± 0.06
Forward fill	BCE	4	4	+	-	5.52 ± 1.43	3.69 ± 0.44	37.92 ± 33.77	0.65 ± 0.10	0.25 ± 0.15
Interpolation	MSE(logit)	1	1	+	+	1.71 ± 0.64	1.51 ± 0.24	0.96 ± 0.13	0.33 ± 0.01	0.75 ± 0.11
Interpolation	BCE	1	1	+	+	1.27 ± 0.41	1.46 ± 0.30	13.12 ± 11.63	0.30 ± 0.03	0.84 ± 0.03
Interpolation	MSE(logit)	1	1	+	-	5.42 ± 1.41	3.09 ± 0.53	3.56 ± 0.47	0.65 ± 0.09	0.26 ± 0.15
Interpolation	BCE	1	1	+	-	6.69 ± 2.37	4.42 ± 0.82	55.59 ± 29.87	0.53 ± 0.04	0.18 ± 0.16
Interpolation	MSE(logit)	1	2	+	+	3.91 ± 1.35	2.17 ± 0.45	1.23 ± 0.21	0.38 ± 0.05	0.52 ± 0.15
Interpolation	BCE	1	2	+	+	1.07 ± 0.40	1.34 ± 0.23	9.12 ± 7.63	0.29 ± 0.05	0.85 ± 0.05
Interpolation	MSE(logit)	1	2	+	-	4.91 ± 1.22	3.08 ± 0.40	4.63 ± 2.56	0.68 ± 0.12	0.19 ± 0.11
Interpolation	BCE	1	2	+	-	5.82 ± 1.61	4.50 ± 0.40	73.48 ± 19.07	0.52 ± 0.04	0.12 ± 0.06
Interpolation	MSE(logit)	1	4	+	+	2.46 ± 0.48	1.68 ± 0.16	1.11 ± 0.06	0.38 ± 0.04	0.63 ± 0.04
Interpolation	BCE	1	4	+	+	1.12 ± 0.39	1.21 ± 0.22	10.15 ± 4.52	0.32 ± 0.04	0.81 ± 0.04
Interpolation	MSE(logit)	1	4	+	-	5.97 ± 0.90	3.14 ± 0.22	2.89 ± 1.02	0.66 ± 0.08	0.21 ± 0.13
Interpolation	BCE	1	4	+	-	7.18 ± 1.20	5.10 ± 0.28	157.29 ± 71.37	0.51 ± 0.01	-0.00 ± 0.00
Interpolation	MSE(logit)	2	1	+	+	3.66 ± 1.14	2.11 ± 0.40	1.07 ± 0.11	0.39 ± 0.05	0.58 ± 0.08
Interpolation	BCE	2	1	+	+	1.21 ± 0.42	1.33 ± 0.29	7.55 ± 4.62	0.29 ± 0.03	0.82 ± 0.05
Interpolation	MSE(logit)	2	1	+	-	4.38 ± 1.04	2.89 ± 0.45	3.34 ± 0.71	0.62 ± 0.03	0.35 ± 0.07
Interpolation	BCE	2	1	+	-	4.22 ± 0.87	3.25 ± 0.59	24.27 ± 7.63	0.61 ± 0.08	0.31 ± 0.13
Interpolation	MSE(logit)	2	2	+	+	2.22 ± 1.08	1.59 ± 0.23	1.11 ± 0.11	0.34 ± 0.03	0.68 ± 0.16
Interpolation	BCE	2	2	+	+	1.20 ± 0.46	1.30 ± 0.25	3.81 ± 1.56	0.30 ± 0.05	0.83 ± 0.04
Interpolation	MSE(logit)	2	2	+	-	4.32 ± 1.44	2.90 ± 0.49	3.14 ± 0.73	0.60 ± 0.08	0.36 ± 0.09
Interpolation	BCE	2	2	+	-	4.37 ± 1.40	3.44 ± 0.47	31.59 ± 17.69	0.57 ± 0.03	0.35 ± 0.10
Interpolation	MSE(logit)	2	4	+	+	2.16 ± 0.90	1.56 ± 0.23	1.23 ± 0.13	0.37 ± 0.02	0.68 ± 0.13

Table A2. Cont.

Imputation	Loss	Layers	Attn. Heads	FFT	Recurrent Embeddings	MSE, $\times 10^{17}$, kg ²	MAE, $\times 10^8$, kg	MAPE _{all}	MAPE _{large}	R ²
Interpolation	BCE	2	4	+	+	1.43 $\pm$ 0.28	1.41 $\pm$ 0.06	5.24 $\pm$ 1.38	0.32 $\pm$ 0.02	0.81 $\pm$ 0.03
Interpolation	MSE(logit)	2	4	+	-	6.45 $\pm$ 1.49	3.46 $\pm$ 0.49	6.82 $\pm$ 2.65	0.81 $\pm$ 0.12	0.03 $\pm$ 0.23
Interpolation	BCE	2	4	+	-	4.20 $\pm$ 1.03	2.97 $\pm$ 0.46	25.45 $\pm$ 6.85	0.73 $\pm$ 0.09	0.32 $\pm$ 0.07
Interpolation	MSE(logit)	3	1	+	+	1.76 $\pm$ 0.89	1.53 $\pm$ 0.38	1.21 $\pm$ 0.34	0.36 $\pm$ 0.08	0.76 $\pm$ 0.07
Interpolation	BCE	3	1	+	+	0.91 $\pm$ 0.34	1.20 $\pm$ 0.16	3.97 $\pm$ 2.88	0.28 $\pm$ 0.03	0.86 $\pm$ 0.05
Interpolation	MSE(logit)	3	1	+	-	5.19 $\pm$ 0.95	3.03 $\pm$ 0.28	4.63 $\pm$ 1.72	0.68 $\pm$ 0.07	0.23 $\pm$ 0.14
Interpolation	BCE	3	1	+	-	4.80 $\pm$ 0.54	3.55 $\pm$ 0.23	50.16 $\pm$ 47.81	0.66 $\pm$ 0.05	0.25 $\pm$ 0.11
Interpolation	MSE(logit)	3	2	+	+	3.48 $\pm$ 1.64	2.02 $\pm$ 0.48	1.21 $\pm$ 0.08	0.39 $\pm$ 0.07	0.51 $\pm$ 0.20
Interpolation	BCE	3	2	+	+	0.94 $\pm$ 0.25	1.23 $\pm$ 0.14	7.29 $\pm$ 2.85	0.25 $\pm$ 0.01	0.87 $\pm$ 0.02
Interpolation	MSE(logit)	3	2	+	-	4.56 $\pm$ 1.10	2.92 $\pm$ 0.51	4.54 $\pm$ 2.57	0.65 $\pm$ 0.07	0.33 $\pm$ 0.13
Interpolation	BCE	3	2	+	-	5.80 $\pm$ 1.20	4.05 $\pm$ 0.45	35.55 $\pm$ 22.95	0.60 $\pm$ 0.05	0.25 $\pm$ 0.12
Interpolation	MSE(logit)	3	4	+	+	2.33 $\pm$ 0.79	1.68 $\pm$ 0.32	1.07 $\pm$ 0.18	0.32 $\pm$ 0.03	0.70 $\pm$ 0.04
Interpolation	BCE	3	4	+	+	0.91 $\pm$ 0.36	1.17 $\pm$ 0.23	4.46 $\pm$ 1.36	0.27 $\pm$ 0.03	0.87 $\pm$ 0.03
Interpolation	MSE(logit)	3	4	+	-	4.59 $\pm$ 1.38	2.98 $\pm$ 0.54	4.19 $\pm$ 1.45	0.65 $\pm$ 0.09	0.30 $\pm$ 0.08
Interpolation	BCE	3	4	+	-	5.80 $\pm$ 1.71	3.74 $\pm$ 0.82	55.00 $\pm$ 22.26	0.59 $\pm$ 0.07	0.19 $\pm$ 0.15
Interpolation	MSE(logit)	4	1	+	+	8.08 $\pm$ 2.56	3.98 $\pm$ 0.95	9.69 $\pm$ 2.58	0.89 $\pm$ 0.01	-0.17 $\pm$ 0.06
Interpolation	BCE	4	1	+	+	0.91 $\pm$ 0.46	1.20 $\pm$ 0.24	4.20 $\pm$ 1.19	0.26 $\pm$ 0.04	0.87 $\pm$ 0.05
Interpolation	MSE(logit)	4	1	+	-	7.08 $\pm$ 2.78	3.75 $\pm$ 0.90	6.25 $\pm$ 2.09	0.81 $\pm$ 0.15	0.02 $\pm$ 0.28
Interpolation	BCE	4	1	+	-	5.93 $\pm$ 2.15	3.58 $\pm$ 0.78	37.12 $\pm$ 21.17	0.66 $\pm$ 0.05	0.25 $\pm$ 0.13
Interpolation	MSE(logit)	4	2	+	+	2.89 $\pm$ 0.89	1.83 $\pm$ 0.24	0.97 $\pm$ 0.07	0.36 $\pm$ 0.04	0.63 $\pm$ 0.11
Interpolation	BCE	4	2	+	+	1.03 $\pm$ 0.45	1.26 $\pm$ 0.25	6.66 $\pm$ 2.96	0.31 $\pm$ 0.05	0.85 $\pm$ 0.03
Interpolation	MSE(logit)	4	2	+	-	5.65 $\pm$ 1.66	3.22 $\pm$ 0.60	5.64 $\pm$ 2.72	0.67 $\pm$ 0.08	0.27 $\pm$ 0.07
Interpolation	BCE	4	2	+	-	6.04 $\pm$ 1.73	3.79 $\pm$ 0.52	26.44 $\pm$ 12.15	0.67 $\pm$ 0.09	0.14 $\pm$ 0.14
Interpolation	MSE(logit)	4	4	+	+	19.49 $\pm$ 28.13	7.68 $\pm$ 9.84	81.31 $\pm$ 138.85	0.92 $\pm$ 0.99	-1.63 $\pm$ 3.83
Model-based	MSE(logit)	1	1	+	+	2.31 $\pm$ 0.49	1.73 $\pm$ 0.32	1.03 $\pm$ 0.18	0.40 $\pm$ 0.10	0.68 $\pm$ 0.04
Model-based	BCE	1	1	+	+	1.63 $\pm$ 0.89	1.59 $\pm$ 0.43	3.78 $\pm$ 1.43	0.31 $\pm$ 0.05	0.81 $\pm$ 0.08
Model-based	MSE(logit)	1	1	+	-	4.52 $\pm$ 1.69	2.87 $\pm$ 0.64	7.49 $\pm$ 5.39	0.74 $\pm$ 0.06	0.20 $\pm$ 0.15
Model-based	BCE	1	1	+	-	6.42 $\pm$ 0.68	4.50 $\pm$ 0.47	112.95 $\pm$ 81.72	0.53 $\pm$ 0.04	0.08 $\pm$ 0.18
Model-based	MSE(logit)	1	2	+	+	2.48 $\pm$ 0.89	1.60 $\pm$ 0.25	1.25 $\pm$ 0.33	0.37 $\pm$ 0.03	0.65 $\pm$ 0.11
Model-based	BCE	1	2	+	+	3.96 $\pm$ 2.98	3.10 $\pm$ 1.80	81.01 $\pm$ 76.31	0.37 $\pm$ 0.10	0.48 $\pm$ 0.39
Model-based	MSE(logit)	1	2	+	-	6.04 $\pm$ 1.76	3.44 $\pm$ 0.52	3.92 $\pm$ 0.83	0.68 $\pm$ 0.05	0.22 $\pm$ 0.08
Model-based	BCE	1	2	+	-	5.83 $\pm$ 0.60	4.47 $\pm$ 0.31	135.05 $\pm$ 31.64	0.55 $\pm$ 0.04	0.15 $\pm$ 0.05
Model-based	MSE(logit)	1	4	+	+	1.58 $\pm$ 0.58	1.43 $\pm$ 0.28	1.01 $\pm$ 0.20	0.38 $\pm$ 0.10	0.78 $\pm$ 0.08
Model-based	BCE	1	4	+	+	1.99 $\pm$ 1.50	1.72 $\pm$ 0.87	14.34 $\pm$ 10.81	0.33 $\pm$ 0.04	0.73 $\pm$ 0.18
Model-based	MSE(logit)	1	4	+	-	6.78 $\pm$ 1.93	3.44 $\pm$ 0.78	7.69 $\pm$ 4.01	0.74 $\pm$ 0.04	0.10 $\pm$ 0.10
Model-based	BCE	1	4	+	-	10.17 $\pm$ 8.31	5.30 $\pm$ 2.37	102.72 $\pm$ 130.70	0.65 $\pm$ 0.08	-0.36 $\pm$ 1.05
Model-based	MSE(logit)	2	1	+	+	2.90 $\pm$ 0.27	1.86 $\pm$ 0.21	0.94 $\pm$ 0.07	0.37 $\pm$ 0.02	0.63 $\pm$ 0.02
Model-based	BCE	2	1	+	+	1.61 $\pm$ 0.68	1.50 $\pm$ 0.17	11.31 $\pm$ 7.10	0.30 $\pm$ 0.04	0.81 $\pm$ 0.09
Model-based	MSE(logit)	2	1	+	-	6.68 $\pm$ 1.43	3.51 $\pm$ 0.57	3.79 $\pm$ 1.21	0.76 $\pm$ 0.11	0.14 $\pm$ 0.23
Model-based	BCE	2	1	+	-	4.15 $\pm$ 0.52	3.67 $\pm$ 0.26	121.03 $\pm$ 38.99	0.56 $\pm$ 0.03	0.27 $\pm$ 0.07
Model-based	MSE(logit)	2	2	+	+	2.12 $\pm$ 1.24	1.64 $\pm$ 0.53	0.96 $\pm$ 0.08	0.35 $\pm$ 0.06	0.69 $\pm$ 0.17
Model-based	BCE	2	2	+	+	1.27 $\pm$ 0.53	1.45 $\pm$ 0.36	5.05 $\pm$ 3.46	0.32 $\pm$ 0.07	0.84 $\pm$ 0.04
Model-based	MSE(logit)	2	2	+	-	5.86 $\pm$ 1.01	3.47 $\pm$ 0.22	4.79 $\pm$ 2.55	0.75 $\pm$ 0.04	0.14 $\pm$ 0.09
Model-based	BCE	2	2	+	-	6.55 $\pm$ 1.28	4.39 $\pm$ 0.66	73.56 $\pm$ 77.17	0.61 $\pm$ 0.06	0.08 $\pm$ 0.07
Model-based	MSE(logit)	2	4	+	+	3.57 $\pm$ 2.66	2.17 $\pm$ 0.73	3.91 $\pm$ 5.18	0.47 $\pm$ 0.21	0.56 $\pm$ 0.33
Model-based	BCE	2	4	+	+	1.29 $\pm$ 0.49	1.44 $\pm$ 0.30	5.68 $\pm$ 3.96	0.31 $\pm$ 0.04	0.83 $\pm$ 0.04
Model-based	MSE(logit)	2	4	+	-	3.99 $\pm$ 0.68	2.77 $\pm$ 0.43	3.77 $\pm$ 1.89	0.67 $\pm$ 0.09	0.33 $\pm$ 0.13
Model-based	BCE	2	4	+	-	4.70 $\pm$ 1.15	3.48 $\pm$ 0.26	29.98 $\pm$ 9.99	0.60 $\pm$ 0.06	0.30 $\pm$ 0.12
Model-based	MSE(logit)	3	1	+	+	2.49 $\pm$ 1.39	1.70 $\pm$ 0.57	1.09 $\pm$ 0.16	0.36 $\pm$ 0.05	0.63 $\pm$ 0.11
Model-based	BCE	3	1	+	+	0.81 $\pm$ 0.30	1.20 $\pm$ 0.17	4.66 $\pm$ 2.20	0.27 $\pm$ 0.02	0.87 $\pm$ 0.03
Model-based	MSE(logit)	3	1	+	-	5.46 $\pm$ 0.85	2.89 $\pm$ 0.35	4.04 $\pm$ 1.48	0.70 $\pm$ 0.04	0.29 $\pm$ 0.06
Model-based	BCE	3	1	+	-	5.23 $\pm$ 1.25	3.47 $\pm$ 0.62	19.87 $\pm$ 2.74	0.66 $\pm$ 0.08	0.27 $\pm$ 0.13
Model-based	MSE(logit)	3	2	+	+	1.69 $\pm$ 0.30	1.58 $\pm$ 0.16	1.16 $\pm$ 0.14	0.37 $\pm$ 0.04	0.73 $\pm$ 0.05
Model-based	BCE	3	2	+	+	1.42 $\pm$ 0.56	1.40 $\pm$ 0.21	12.40 $\pm$ 4.94	0.32 $\pm$ 0.04	0.78 $\pm$ 0.07
Model-based	MSE(logit)	3	2	+	-	6.67 $\pm$ 1.20	3.69 $\pm$ 0.53	10.97 $\pm$ 8.26	0.80 $\pm$ 0.13	-0.13 $\pm$ 0.44
Model-based	BCE	3	2	+	-	9.48 $\pm$ 9.17	4.23 $\pm$ 1.89	72.63 $\pm$ 106.58	0.78 $\pm$ 0.18	-0.61 $\pm$ 1.83
Model-based	MSE(logit)	3	4	+	+	2.43 $\pm$ 0.69	1.77 $\pm$ 0.32	1.22 $\pm$ 0.22	0.38 $\pm$ 0.04	0.67 $\pm$ 0.08
Model-based	BCE	3	4	+	+	2.59 $\pm$ 1.92	2.18 $\pm$ 1.44	17.61 $\pm$ 6.42	0.37 $\pm$ 0.12	0.64 $\pm$ 0.26
Model-based	MSE(logit)	3	4	+	-	5.53 $\pm$ 1.56	3.51 $\pm$ 0.46	7.47 $\pm$ 5.62	0.71 $\pm$ 0.07	0.20 $\pm$ 0.12
Model-based	BCE	3	4	+	-	5.03 $\pm$ 1.33	3.68 $\pm$ 0.59	40.47 $\pm$ 11.99	0.61 $\pm$ 0.06	0.19 $\pm$ 0.14
Model-based	MSE(logit)	4	1	+	+	46.73 $\pm$ 55.32	14.17 $\pm$ 15.61	208.27 $\pm$ 254.44	1.60 $\pm$ 1.56	-5.31 $\pm$ 7.45

Table A2. Cont.

Imputation	Loss	Layers	Attn. Heads	FFT	Recurrent Embeddings	MSE, $\times 10^{17}$, kg ²	MAE, $\times 10^8$, kg	MAPE _{all}	MAPE _{large}	R ²
Model-based	BCE	4	1	+	+	1.42 $\pm$ 0.63	1.41 $\pm$ 0.35	4.71 $\pm$ 1.91	0.31 $\pm$ 0.06	0.81 $\pm$ 0.08
Model-based	MSE(logit)	4	1	+	-	4.11 $\pm$ 1.29	3.00 $\pm$ 0.46	3.45 $\pm$ 0.91	0.65 $\pm$ 0.07	0.35 $\pm$ 0.15
Model-based	BCE	4	1	+	-	6.56 $\pm$ 1.93	4.26 $\pm$ 0.75	263.20 $\pm$ 446.90	0.67 $\pm$ 0.07	-0.05 $\pm$ 0.19
Model-based	MSE(logit)	4	2	+	+	3.05 $\pm$ 1.33	2.09 $\pm$ 0.62	1.19 $\pm$ 0.14	0.41 $\pm$ 0.04	0.57 $\pm$ 0.18
Model-based	BCE	4	2	+	+	1.10 $\pm$ 0.57	1.26 $\pm$ 0.32	7.37 $\pm$ 5.68	0.29 $\pm$ 0.07	0.84 $\pm$ 0.05
Model-based	MSE(logit)	4	2	+	-	4.74 $\pm$ 0.80	2.88 $\pm$ 0.46	5.14 $\pm$ 2.53	0.69 $\pm$ 0.09	0.22 $\pm$ 0.15
Model-based	BCE	4	2	+	-	5.12 $\pm$ 0.97	3.57 $\pm$ 0.42	51.89 $\pm$ 32.71	0.66 $\pm$ 0.11	0.25 $\pm$ 0.08
Model-based	MSE(logit)	4	4	+	+	2.13 $\pm$ 0.64	1.71 $\pm$ 0.41	1.04 $\pm$ 0.22	0.36 $\pm$ 0.03	0.69 $\pm$ 0.05
Model-based	BCE	4	4	+	+	1.74 $\pm$ 1.22	1.71 $\pm$ 0.78	10.18 $\pm$ 12.19	0.34 $\pm$ 0.15	0.78 $\pm$ 0.16
Model-based	MSE(logit)	4	4	+	-	6.02 $\pm$ 2.33	3.43 $\pm$ 1.06	6.31 $\pm$ 3.82	0.71 $\pm$ 0.04	0.22 $\pm$ 0.12
Model-based	BCE	4	4	+	-	5.37 $\pm$ 0.43	3.70 $\pm$ 0.31	38.21 $\pm$ 32.85	0.68 $\pm$ 0.05	0.30 $\pm$ 0.05
ARIMA	MSE(logit)	1	1	+	+	2.82 $\pm$ 0.68	1.72 $\pm$ 0.21	1.11 $\pm$ 0.32	0.38 $\pm$ 0.03	0.61 $\pm$ 0.08
ARIMA	BCE	1	1	+	+	1.43 $\pm$ 0.46	1.66 $\pm$ 0.36	4.93 $\pm$ 5.15	0.30 $\pm$ 0.03	0.84 $\pm$ 0.03
ARIMA	MSE(logit)	1	1	+	-	7.46 $\pm$ 1.70	3.92 $\pm$ 0.64	7.10 $\pm$ 2.79	0.79 $\pm$ 0.06	0.09 $\pm$ 0.19
ARIMA	BCE	1	1	+	-	6.87 $\pm$ 1.96	4.31 $\pm$ 0.75	37.82 $\pm$ 12.43	0.58 $\pm$ 0.03	0.19 $\pm$ 0.13
ARIMA	MSE(logit)	1	2	+	+	1.83 $\pm$ 0.35	1.67 $\pm$ 0.24	1.07 $\pm$ 0.10	0.33 $\pm$ 0.02	0.74 $\pm$ 0.04
ARIMA	BCE	1	2	+	+	1.13 $\pm$ 0.51	1.25 $\pm$ 0.22	5.81 $\pm$ 2.84	0.28 $\pm$ 0.04	0.85 $\pm$ 0.04
ARIMA	MSE(logit)	1	2	+	-	4.29 $\pm$ 0.73	2.77 $\pm$ 0.70	4.50 $\pm$ 1.30	0.65 $\pm$ 0.07	0.27 $\pm$ 0.08
ARIMA	BCE	1	2	+	-	4.48 $\pm$ 0.68	3.65 $\pm$ 0.37	61.71 $\pm$ 18.07	0.58 $\pm$ 0.10	0.30 $\pm$ 0.13
ARIMA	MSE(logit)	1	4	+	+	2.13 $\pm$ 0.64	1.55 $\pm$ 0.27	1.19 $\pm$ 0.17	0.34 $\pm$ 0.03	0.68 $\pm$ 0.02
ARIMA	BCE	1	4	+	+	1.36 $\pm$ 0.53	1.45 $\pm$ 0.30	5.02 $\pm$ 3.51	0.28 $\pm$ 0.03	0.83 $\pm$ 0.06
ARIMA	MSE(logit)	1	4	+	-	5.11 $\pm$ 2.06	3.03 $\pm$ 0.78	4.60 $\pm$ 2.12	0.72 $\pm$ 0.08	0.18 $\pm$ 0.13
ARIMA	BCE	1	4	+	-	5.18 $\pm$ 0.37	4.06 $\pm$ 0.52	103.04 $\pm$ 71.29	0.59 $\pm$ 0.10	0.23 $\pm$ 0.12
ARIMA	MSE(logit)	2	1	+	+	2.26 $\pm$ 0.55	1.64 $\pm$ 0.22	1.18 $\pm$ 0.27	0.36 $\pm$ 0.05	0.66 $\pm$ 0.06
ARIMA	BCE	2	1	+	+	1.31 $\pm$ 0.35	1.54 $\pm$ 0.33	4.90 $\pm$ 2.21	0.31 $\pm$ 0.05	0.84 $\pm$ 0.02
ARIMA	MSE(logit)	2	1	+	-	4.80 $\pm$ 0.65	3.08 $\pm$ 0.29	2.98 $\pm$ 0.56	0.71 $\pm$ 0.10	0.29 $\pm$ 0.11
ARIMA	BCE	2	1	+	-	4.37 $\pm$ 0.61	3.33 $\pm$ 0.23	20.57 $\pm$ 12.20	0.62 $\pm$ 0.02	0.34 $\pm$ 0.08
ARIMA	MSE(logit)	2	2	+	+	1.60 $\pm$ 0.70	1.37 $\pm$ 0.30	1.09 $\pm$ 0.21	0.31 $\pm$ 0.05	0.74 $\pm$ 0.08
ARIMA	BCE	2	2	+	+	1.18 $\pm$ 0.31	1.35 $\pm$ 0.28	3.71 $\pm$ 1.08	0.28 $\pm$ 0.03	0.83 $\pm$ 0.05
ARIMA	MSE(logit)	2	2	+	-	5.69 $\pm$ 0.84	3.28 $\pm$ 0.20	4.15 $\pm$ 2.24	0.65 $\pm$ 0.04	0.22 $\pm$ 0.09
ARIMA	BCE	2	2	+	-	5.58 $\pm$ 3.10	3.85 $\pm$ 0.96	33.48 $\pm$ 13.82	0.59 $\pm$ 0.04	0.32 $\pm$ 0.13
ARIMA	MSE(logit)	2	4	+	+	2.22 $\pm$ 0.65	1.75 $\pm$ 0.39	0.89 $\pm$ 0.13	0.36 $\pm$ 0.07	0.71 $\pm$ 0.07
ARIMA	BCE	2	4	+	+	1.14 $\pm$ 0.47	1.33 $\pm$ 0.27	7.47 $\pm$ 4.63	0.29 $\pm$ 0.05	0.82 $\pm$ 0.07
ARIMA	MSE(logit)	2	4	+	-	5.63 $\pm$ 1.24	3.26 $\pm$ 0.62	4.74 $\pm$ 1.95	0.66 $\pm$ 0.07	0.23 $\pm$ 0.15
ARIMA	BCE	2	4	+	-	6.03 $\pm$ 0.21	4.97 $\pm$ 0.12	165.61 $\pm$ 19.61	0.52 $\pm$ 0.02	-0.00 $\pm$ 0.00
ARIMA	MSE(logit)	3	1	+	+	1.88 $\pm$ 1.10	1.53 $\pm$ 0.40	0.91 $\pm$ 0.07	0.32 $\pm$ 0.04	0.75 $\pm$ 0.10
ARIMA	BCE	3	1	+	+	1.10 $\pm$ 0.28	1.29 $\pm$ 0.12	2.83 $\pm$ 1.15	0.28 $\pm$ 0.02	0.85 $\pm$ 0.02
ARIMA	MSE(logit)	3	1	+	-	5.09 $\pm$ 1.56	3.19 $\pm$ 0.57	6.22 $\pm$ 2.54	0.74 $\pm$ 0.07	0.27 $\pm$ 0.04
ARIMA	BCE	3	1	+	-	5.57 $\pm$ 1.87	4.11 $\pm$ 0.80	45.67 $\pm$ 13.26	0.62 $\pm$ 0.10	0.04 $\pm$ 0.26
ARIMA	MSE(logit)	3	2	+	+	1.52 $\pm$ 0.40	1.34 $\pm$ 0.31	0.95 $\pm$ 0.07	0.33 $\pm$ 0.06	0.76 $\pm$ 0.05
ARIMA	BCE	3	2	+	+	1.28 $\pm$ 0.34	1.40 $\pm$ 0.16	4.33 $\pm$ 3.43	0.29 $\pm$ 0.01	0.83 $\pm$ 0.02
ARIMA	MSE(logit)	3	2	+	-	6.15 $\pm$ 1.71	3.38 $\pm$ 0.65	6.42 $\pm$ 4.26	0.74 $\pm$ 0.09	0.20 $\pm$ 0.17
ARIMA	BCE	3	2	+	-	5.71 $\pm$ 1.51	3.88 $\pm$ 0.41	59.30 $\pm$ 27.34	0.64 $\pm$ 0.06	0.17 $\pm$ 0.12
ARIMA	MSE(logit)	3	4	+	+	2.47 $\pm$ 0.94	1.76 $\pm$ 0.37	1.06 $\pm$ 0.21	0.35 $\pm$ 0.05	0.68 $\pm$ 0.10
ARIMA	BCE	3	4	+	+	1.18 $\pm$ 0.35	1.31 $\pm$ 0.20	8.88 $\pm$ 1.92	0.31 $\pm$ 0.02	0.83 $\pm$ 0.02
ARIMA	MSE(logit)	3	4	+	-	6.67 $\pm$ 1.12	3.66 $\pm$ 0.45	5.23 $\pm$ 1.56	0.67 $\pm$ 0.08	0.17 $\pm$ 0.14
ARIMA	BCE	3	4	+	-	5.56 $\pm$ 1.26	3.96 $\pm$ 0.46	42.09 $\pm$ 22.79	0.61 $\pm$ 0.08	0.17 $\pm$ 0.06
ARIMA	MSE(logit)	4	1	+	+	2.46 $\pm$ 1.10	1.75 $\pm$ 0.47	1.19 $\pm$ 0.22	0.34 $\pm$ 0.02	0.68 $\pm$ 0.06
ARIMA	BCE	4	1	+	+	1.14 $\pm$ 0.35	1.36 $\pm$ 0.24	5.56 $\pm$ 2.55	0.31 $\pm$ 0.04	0.83 $\pm$ 0.04
ARIMA	MSE(logit)	4	1	+	-	4.19 $\pm$ 1.12	2.77 $\pm$ 0.45	3.37 $\pm$ 0.64	0.68 $\pm$ 0.10	0.26 $\pm$ 0.14
ARIMA	BCE	4	1	+	-	5.69 $\pm$ 1.94	3.48 $\pm$ 0.60	28.93 $\pm$ 8.54	0.71 $\pm$ 0.13	0.18 $\pm$ 0.17
ARIMA	MSE(logit)	4	2	+	+	1.96 $\pm$ 0.80	1.46 $\pm$ 0.32	1.08 $\pm$ 0.23	0.34 $\pm$ 0.07	0.68 $\pm$ 0.10
ARIMA	BCE	4	2	+	+	1.34 $\pm$ 0.50	1.33 $\pm$ 0.23	9.42 $\pm$ 13.70	0.28 $\pm$ 0.03	0.82 $\pm$ 0.04
ARIMA	MSE(logit)	4	2	+	-	4.32 $\pm$ 0.67	2.70 $\pm$ 0.21	4.91 $\pm$ 1.34	0.67 $\pm$ 0.08	0.30 $\pm$ 0.07
ARIMA	BCE	4	2	+	-	5.71 $\pm$ 1.48	3.65 $\pm$ 0.57	39.23 $\pm$ 24.22	0.59 $\pm$ 0.06	0.22 $\pm$ 0.11
ARIMA	MSE(logit)	4	4	+	+	2.93 $\pm$ 1.47	2.02 $\pm$ 0.49	1.26 $\pm$ 0.28	0.39 $\pm$ 0.05	0.63 $\pm$ 0.12
ARIMA	BCE	4	4	+	+	1.43 $\pm$ 0.65	1.45 $\pm$ 0.33	7.75 $\pm$ 4.61	0.31 $\pm$ 0.04	0.80 $\pm$ 0.05
ARIMA	MSE(logit)	4	4	+	-	5.30 $\pm$ 0.72	2.96 $\pm$ 0.39	7.09 $\pm$ 4.87	0.65 $\pm$ 0.04	0.30 $\pm$ 0.09
ARIMA	BCE	4	4	+	-	5.02 $\pm$ 1.93	3.49 $\pm$ 0.67	28.37 $\pm$ 15.94	0.61 $\pm$ 0.08	0.23 $\pm$ 0.21
ARIMA	MSE(logit)	1	1	-	+	1.94 $\pm$ 0.50	1.59 $\pm$ 0.22	1.26 $\pm$ 0.19	0.36 $\pm$ 0.06	0.74 $\pm$ 0.09
ARIMA	BCE	1	1	-	+	1.15 $\pm$ 0.28	1.28 $\pm$ 0.19	7.44 $\pm$ 0.99	0.32 $\pm$ 0.02	0.82 $\pm$ 0.02

Table A2. Cont.

Imputation	Loss	Layers	Attn. Heads	FFT	Recurrent Embeddings	$MSE, \times 10^{17}, \text{kg}^2$	$MAE, \times 10^8, \text{kg}$	$MAPE_{all}$	$MAPE_{large}$	R^2
ARIMA	MSE(logit)	1	1	-	-	1.01 ± 0.75	1.26 ± 0.34	1.35 ± 0.10	0.31 ± 0.04	0.85 ± 0.09
ARIMA	BCE	1	1	-	-	1.08 ± 0.42	1.35 ± 0.26	10.39 ± 2.99	0.30 ± 0.02	0.83 ± 0.05
ARIMA	MSE(logit)	1	2	-	+	2.23 ± 0.53	1.66 ± 0.24	1.19 ± 0.18	0.34 ± 0.02	0.69 ± 0.06
ARIMA	BCE	1	2	-	+	1.01 ± 0.59	1.20 ± 0.18	9.77 ± 2.42	0.29 ± 0.03	0.85 ± 0.06
ARIMA	MSE(logit)	1	2	-	-	1.17 ± 0.55	1.25 ± 0.19	1.37 ± 0.16	0.32 ± 0.04	0.81 ± 0.05
ARIMA	BCE	1	2	-	-	1.34 ± 0.37	1.45 ± 0.11	7.49 ± 4.18	0.29 ± 0.02	0.84 ± 0.03
ARIMA	MSE(logit)	1	4	-	+	2.08 ± 0.73	1.53 ± 0.24	1.14 ± 0.08	0.31 ± 0.04	0.74 ± 0.07
ARIMA	BCE	1	4	-	+	1.36 ± 0.41	1.44 ± 0.24	8.82 ± 5.09	0.30 ± 0.01	0.81 ± 0.03
ARIMA	MSE(logit)	1	4	-	-	1.59 ± 0.41	1.41 ± 0.16	1.21 ± 0.10	0.31 ± 0.03	0.79 ± 0.04
ARIMA	BCE	1	4	-	-	1.35 ± 0.46	1.33 ± 0.12	10.27 ± 2.82	0.31 ± 0.02	0.82 ± 0.05
ARIMA	MSE(logit)	2	1	-	+	1.70 ± 0.53	1.54 ± 0.23	1.48 ± 0.38	0.33 ± 0.02	0.77 ± 0.05
ARIMA	BCE	2	1	-	+	1.33 ± 0.57	1.44 ± 0.21	5.52 ± 2.16	0.30 ± 0.03	0.83 ± 0.05
ARIMA	MSE(logit)	2	1	-	-	1.72 ± 0.99	1.39 ± 0.24	1.36 ± 0.16	0.33 ± 0.04	0.77 ± 0.07
ARIMA	BCE	2	1	-	-	1.69 ± 0.67	1.42 ± 0.11	7.03 ± 2.51	0.31 ± 0.02	0.77 ± 0.08
ARIMA	MSE(logit)	2	2	-	+	2.17 ± 0.42	1.59 ± 0.10	1.16 ± 0.07	0.35 ± 0.03	0.72 ± 0.05
ARIMA	BCE	2	2	-	+	1.17 ± 0.30	1.33 ± 0.15	6.51 ± 3.28	0.28 ± 0.02	0.83 ± 0.03
ARIMA	MSE(logit)	2	2	-	-	1.93 ± 1.17	1.61 ± 0.42	1.45 ± 0.19	0.31 ± 0.02	0.76 ± 0.07
ARIMA	BCE	2	2	-	-	1.90 ± 0.90	1.50 ± 0.36	7.16 ± 2.52	0.32 ± 0.04	0.76 ± 0.10
ARIMA	MSE(logit)	2	4	-	+	2.62 ± 0.69	1.65 ± 0.18	1.34 ± 0.15	0.37 ± 0.03	0.60 ± 0.07
ARIMA	BCE	2	4	-	+	1.91 ± 0.43	1.64 ± 0.28	8.38 ± 5.63	0.31 ± 0.03	0.75 ± 0.05
ARIMA	MSE(logit)	2	4	-	-	1.65 ± 0.42	1.56 ± 0.22	1.26 ± 0.23	0.35 ± 0.04	0.79 ± 0.04
ARIMA	BCE	2	4	-	-	1.26 ± 0.33	1.43 ± 0.14	7.09 ± 3.06	0.28 ± 0.03	0.82 ± 0.03
ARIMA	MSE(logit)	3	1	-	+	1.16 ± 0.36	1.25 ± 0.10	1.21 ± 0.13	0.32 ± 0.02	0.81 ± 0.04
ARIMA	BCE	3	1	-	+	1.81 ± 0.47	1.47 ± 0.24	6.81 ± 2.72	0.32 ± 0.03	0.76 ± 0.05
ARIMA	MSE(logit)	3	1	-	-	1.72 ± 0.81	1.62 ± 0.52	1.53 ± 0.48	0.37 ± 0.11	0.78 ± 0.09
ARIMA	BCE	3	1	-	-	1.52 ± 0.69	1.53 ± 0.37	7.38 ± 2.11	0.29 ± 0.02	0.80 ± 0.05
ARIMA	MSE(logit)	3	2	-	+	1.45 ± 0.40	1.39 ± 0.16	1.67 ± 0.33	0.32 ± 0.04	0.79 ± 0.04
ARIMA	BCE	3	2	-	+	1.56 ± 0.75	1.58 ± 0.41	9.16 ± 8.66	0.30 ± 0.03	0.80 ± 0.06
ARIMA	MSE(logit)	3	2	-	-	1.30 ± 0.00	1.14 ± 0.00	1.47 ± 0.00	0.30 ± 0.00	0.82 ± 0.00

Table A3 shows the performance score for the encoder–decoder model with different hyperparameters. Most of the outcomes here are the same as for the encoder-only model. The “Masked ratio” shows the ratio of exports from the current year t that we use as the input in the decoder (i.e., this is the ratio of exports to be restored).

Table A3. Performance scores for the proposed encoder–decoder model with various hyperparameters and imputation.

Imputation	Loss	FFT	Recurrent Embeddings	Masked Ratio	$MSE, \times 10^{17}, \text{kg}^2$	$MAE, \times 10^8, \text{kg}$	$MAPE_{all}$	$MAPE_{large}$	R^2
Forward fill	MSE(logit)	+	-	0.1	8.82 ± 2.53	4.22 ± 0.82	10.25 ± 0.99	0.88 ± 0.02	-0.17 ± 0.04
Forward fill	BCE	+	-	0.1	1.35 ± 0.67	1.76 ± 0.44	12.53 ± 9.51	0.35 ± 0.03	0.82 ± 0.05
Forward fill	MSE(logit)	+	+	0.1	2.53 ± 0.63	1.92 ± 0.33	1.92 ± 0.41	0.37 ± 0.05	0.70 ± 0.04
Forward fill	BCE	+	+	0.1	1.69 ± 0.22	1.89 ± 0.09	7.73 ± 1.56	0.38 ± 0.06	0.79 ± 0.04
Forward fill	MSE(logit)	+	+	0.5	1.58 ± 0.18	1.46 ± 0.11	1.10 ± 0.15	0.32 ± 0.03	0.73 ± 0.04
Forward fill	BCE	+	+	0.5	0.71 ± 0.16	1.13 ± 0.12	17.46 ± 2.06	0.26 ± 0.05	0.88 ± 0.04
Model-based	MSE(logit)	+	-	0.1	8.78 ± 0.00	4.56 ± 0.00	6.88 ± 0.00	0.80 ± 0.00	-0.13 ± 0.00
Model-based	MSE(logit)	+	+	0.1	2.91 ± 1.39	1.92 ± 0.37	1.45 ± 0.16	0.37 ± 0.03	0.63 ± 0.06

Table A3. Cont.

Imputation	Loss	FFT	Recurrent Embeddings	Masked Ratio	MSE, $\times 10^{17}$, kg ²	MAE, $\times 10^8$, kg	MAPE _{all}	MAPE _{large}	R ²
ARIMA	MSE(logit)	+	+	0.1	1.86 $\pm$ 1.04	1.77 $\pm$ 0.51	2.79 $\pm$ 0.75	0.35 $\pm$ 0.05	0.73 $\pm$ 0.13
ARIMA	BCE	+	+	0.1	1.35 $\pm$ 0.38	1.64 $\pm$ 0.26	10.90 $\pm$ 4.18	0.35 $\pm$ 0.02	0.81 $\pm$ 0.03
Interpolation	MSE(logit)	-	+	0.1	2.02 $\pm$ 0.96	1.71 $\pm$ 0.31	1.98 $\pm$ 0.50	0.37 $\pm$ 0.07	0.68 $\pm$ 0.16
Interpolation	BCE	-	+	0.1	2.40 $\pm$ 0.41	2.21 $\pm$ 0.16	5.44 $\pm$ 2.36	0.40 $\pm$ 0.05	0.74 $\pm$ 0.03
Interpolation	MSE(logit)	-	-	0.1	2.28 $\pm$ 1.28	2.12 $\pm$ 0.64	3.27 $\pm$ 1.20	0.62 $\pm$ 0.18	0.66 $\pm$ 0.13
Interpolation	BCE	-	-	0.1	1.97 $\pm$ 0.39	1.64 $\pm$ 0.19	6.67 $\pm$ 2.65	0.34 $\pm$ 0.03	0.76 $\pm$ 0.03
Interpolation	MSE(logit)	-	+	0.3	1.68 $\pm$ 0.43	1.50 $\pm$ 0.17	1.46 $\pm$ 0.38	0.34 $\pm$ 0.04	0.73 $\pm$ 0.08
Interpolation	BCE	-	+	0.3	3.33 $\pm$ 1.82	2.82 $\pm$ 1.47	76.73 $\pm$ 90.06	0.38 $\pm$ 0.10	0.50 $\pm$ 0.36

References

- Food and Agriculture Organisation. *World Food and Agriculture*; FAO: Rome, Italy, 2015.
- Caparas, M.; Zobel, Z.; Castanho, A.D.; Schwalm, C.R. Increasing risks of crop failure and water scarcity in global breadbaskets by 2030. *Environ. Res. Lett.* **2021**, *16*, 104013. [\[CrossRef\]](#)
- Zhang, Y.T.; Zhou, W.-X. Structural evolution of international crop trade networks. *Front. Phys.* **2022**, *10*, 926764. [\[CrossRef\]](#)
- Duan, J.; Nie, C.; Wang, Y.; Yan, D.; Xiong, W. Research on Global Grain Trade Network Pattern and Its Driving Factors. *Sustainability* **2022**, *14*, 245. [\[CrossRef\]](#)
- Food and Agriculture Organisation. *The State of Agricultural Commodity Markets*; FAO: Rome, Italy, 2022.
- Burkholz, R.; Schweitzer, F. International crop trade networks: The impact of shocks and cascades. *Environ. Res. Lett.* **2019**, *14*, 114013. [\[CrossRef\]](#)
- Seekell, D.; Carr, J.; Dell'Angelo, J.; D'Odorico, P.; Fader, M.; Gephart, J.; Kumm, M.; Magliocca, N.; Porkka, M.; Puma, M. Resilience in the global food system. *Environ. Res. Lett.* **2017**, *12*, 025010. [\[CrossRef\]](#) [\[PubMed\]](#)
- Burkholz, R.; Garas, A.; Schweitzer, F. How damage diversification can reduce systemic risk. *Phys. Rev. E* **2016**, *93*, 042313. [\[CrossRef\]](#)
- Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food security: The challenge of feeding 9 billion people. *Science* **2010**, *327*, 812–818. [\[CrossRef\]](#)
- Cabell, J.F.; Myles, O. An indicator framework for assessing agroecosystem resilience. *Ecol. Soc.* **2012**, *17*, 1–13. [\[CrossRef\]](#)
- Schipanski, M.E.; MacDonald, G.K.; Rosenzweig, S.; Chappell, M.J.; Bennett, E.M.; Kerr, R.B.; Blesh, J.; Crews, T.; Drinkwater, L.; Lundgren, J.G.; et al. Realizing resilient food systems. *BioScience* **2016**, *66*, 600–610. [\[CrossRef\]](#)
- Robu, R.G.; Alexoaei, A.P.; Cojanu, V.; Miron, D. The cereal network: A baseline approach to current configurations of trade communities. *Agric. Food Econ.* **2024**, *12*, 24. [\[CrossRef\]](#)
- Schoenherr, T.; Kanak, G.; Montalbano, A.; Patel, S.; Bourlakis, M.; Sawyerr, E.; Cong, W.F. *Frontiers in Agri-Food Supply Chains: Frameworks and Case Studies*; Burleigh Dodds Science Publishing: Cambridge, UK, 2024.
- Davide, M. S-MARL: An Algorithm for Single-To-Multi-Agent Reinforcement Learning: Case Study: Formula 1 Race Strategies. Available online: <https://www.diva-portal.org/smash/get/diva2:1763095/FULLTEXT01.pdf> (accessed on 2 August 2025).
- Ahmar, A.S.; Singh, P.K.; Ruliana, R.; Pandey, A.K.; Gupta, S. Comparison of ARIMA, SutteARIMA, and Holt-Winters, and NNAR Models to Predict Food Grain in India. *Forecasting* **2023**, *5*, 138–152. [\[CrossRef\]](#)
- Qader, S.H.; Utazi, C.E.; Priyatikanto, R.; Najmaddin, P.; Hama-Ali, E.O.; Khwarahm, N.R.; Dash, J. Exploring the use of Sentinel-2 datasets and environmental variables to model wheat crop yield in smallholder arid and semi-arid farming systems. *Sci. Total Environ.* **2023**, *869*, 161716. [\[CrossRef\]](#)
- Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sens.* **2020**, *12*, 2291. [\[CrossRef\]](#)
- Leucci, A.C.; Ghinoi, S.; Sgargi, D.; Wesz Junior, V.J. VAR models for dynamic analysis of prices in the agri-food system. In *Agricultural Cooperative Management and Policy: New Robust, Reliable and Coherent Modelling Tools*; Springer International Publishing: Cham, Switzerland, 2014; pp. 3–21.
- Lutkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: Berlin/Heidelberg, Germany, 2005.
- Li, X.; Yuan, J. DeepVARwT: Deep learning for a VAR model with trend. *J. Appl. Stat.* **2025**, *52*, 1–27. [\[CrossRef\]](#)

21. Rumánková, L. Evaluation of market relations in soft milling wheat agri-food chain. *AGRIS on-line. Pap. Econ. Inform.* **2016**, *8*, 133–141.
22. Zhao, Y.; Ye, L.; Pinson, P.; Tang, Y.; Lu, P. Correlation-constrained and sparsity-controlled vector autoregressive model for spatio-temporal wind power forecasting. *IEEE Trans. Power Syst.* **2018**, *33*, 5029–5040. [\[CrossRef\]](#)
23. Han, K.; Leem, K.; Choi, Y.R.; Chung, K. What drives a country's fish consumption? Market growth phase and the causal relations among fish consumption, production and income growth. *Fish. Res.* **2022**, *254*, 106435. [\[CrossRef\]](#)
24. Xiong, T.; Li, C.; Bao, Y.; Hu, Z.; Zhang, L. A combination method for interval forecasting of agricultural commodity futures prices. *Knowl. Based Syst.* **2015**, *77*, 92–102. [\[CrossRef\]](#)
25. Iniyar, S.; Varma, V.A.; Naidu, C.T. Crop yield prediction using machine learning techniques. *Adv. Eng. Softw.* **2023**, *175*, 103326. [\[CrossRef\]](#)
26. Panda, S.K.; Mohanty, S.N. Time series forecasting and modeling of food demand supply chain based on regressors analysis. *IEEE Access* **2023**, *11*, 42679–42700. [\[CrossRef\]](#)
27. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T.Y.; Tegmark, M. Kan: Kolmogorov-Arnold networks. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
28. Xu, K.; Chen, L.; Wang, S. Kolmogorov-Arnold networks for time series: Bridging predictive power and interpretability. *arXiv* **2024**, arXiv:2406.02496. [\[CrossRef\]](#)
29. Genet, R.; Inzirillo, H. A Temporal Kolmogorov-Arnold Transformer for Time Series Forecasting. *arXiv* **2024**, arXiv:2406.02486. [\[CrossRef\]](#)
30. Bhattacharya, P.; Mukherjee, T. EleKAN: Temporal Kolmogorov-Arnold Networks for Price and Demand Forecasting Framework in Smart Cities. In Proceedings of the International Conference on Frontiers of Electronics, Information and Computation Technologies, Singapore, 22–24 June 2024; pp. 182–192.
31. Le, D.; Rajasegarar, S.; Luo, W.; Nguyen, T.T.; Angelova, M. Navigating Uncertainty: Gold Price Forecasting with Kolmogorov-Arnold Networks in Volatile Markets. In Proceedings of the 2024 IEEE Conference on Engineering Informatics (ICEI), Melbourne, Australia, 20–21 November 2024; pp. 1–9.
32. Ibañez, S.C.; Monterola, C.P. A Global Forecasting Approach to Large-Scale Crop Production Prediction with Time Series Transformers. *Agriculture* **2023**, *13*, 1855. [\[CrossRef\]](#)
33. Ma, B.; Xue, Y.; Chen, J.; Sun, F. Meta-Learning Enhanced Trade Forecasting: A Neural Framework Leveraging Efficient Multicommodity STL Decomposition. *Int. J. Intell. Syst.* **2024**, *2024*, 6176898. [\[CrossRef\]](#)
34. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier neural operator for parametric partial differential equations. *arXiv* **2020**, arXiv:2010.08895.
35. Buchholz, T.O.; Jug, F. Fourier image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1846–1854.
36. He, S.; Lin, G.; Li, T.; Chen, Y. Frequency-Domain Fusion Transformer for Image Inpainting. *arXiv* **2025**, arXiv:2506.18437. [\[CrossRef\]](#)
37. Li, Z.; Liu, T.; Peng, W.; Yuan, Z.; Wang, J. A transformer-based neural operator for large-eddy simulation of turbulence. *Phys. Fluids* **2024**, *36*, 065167. [\[CrossRef\]](#)
38. Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; Hoi, S. Learning deep time-index models for time series forecasting. In Proceedings of the 14th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
39. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Li, H. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3848–3858. [\[CrossRef\]](#)
40. Li, Y.; Moura, J.M.F. Forecaster: A Graph Transformer for Forecasting Spatial and Time-Dependent Data. In Proceedings of the European Conference on Artificial Intelligence, Santiago de Compostela, Spain, 31 August–8 September 2020.
41. Banerjee, S.; Dong, M.; Shi, W. Spatial-temporal synchronous graph transformer network (stsgt) for COVID-19 forecasting. *Smart Health* **2022**, *26*, 100348. [\[CrossRef\]](#)
42. Liu, H.; Dong, Z.; Jiang, R.; Deng, J.; Chen, Q.; Song, X. Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023.
43. Wang, H.; Chen, J.; Pan, T.; Dong, Z. STGformer: Efficient Spatiotemporal Graph Transformer for Traffic Forecasting. *arXiv* **2024**, arXiv:2410.00385. [\[CrossRef\]](#)
44. Hu, S.; Zou, G.; Lin, S.; Wu, L.; Zhou, C.; Zhang, B.; Chen, Y. Recurrent transformer for dynamic graph representation learning with edge temporal states. *arXiv* **2023**, arXiv:2304.10079.
45. Lee, J.; Xu, C.; Xie, Y. Transformer Conformal Prediction for Time Series. In Proceedings of the ICML 2024 Workshop on Structured Probabilistic Inference and Generative Modeling, Vienna, Austria, 21–27 July 2024.
46. Caliendo, L.; Parro, F. *Gains from Trade: A Model for Counterfactual Trade Policy Analysis*; The University of Chicago Working Papers; The University of Chicago: Chicago, IL, USA, 2009.

47. Yu, Z.; Han, J.; Shi, X.; Yang, Y. Estimates of the trade and global value chain (GVC) effects of China's Pilot Free Trade Zones: A research based on the quantitative trade model. *J. Asia Pac. Econ.* **2025**, *30*, 1255–1302. [\[CrossRef\]](#)
48. Roningen, V.O.; Dixit, P.M. *Economic Implications of Agricultural Policy Reform in Industrial Market Economics*; Rapport Techniques de l'USDA, Ages 80–36; United States Department of Agriculture, Economic Research Service: Washington, DC, USA, 1989.
49. Hertel, T.W. *Global Trade Analysis: Modeling and Applications*; Cambridge University Press: Cambridge, MA, USA, 1997.
50. Parikh, K.S.; Fisher, G.; Frohberg, K.; Gulbrandsen, O. *Towards Free Trade in Agriculture*; Martinus Mijhoff Publishers: Laxemburg, Austria, 1988.
51. Keeney, R.M.; Hertel, T. *GTAP-AGR: A Framework for Assessing the Implications of Multilateral Changes in Agricultural Policies: GTAP Technical Paper*; Purdue University: West Lafayette, IN, USA, 2005.
52. Britz, W.; Leip, A. Development of marginal emission factors for N losses from agricultural soils with the DNDC-CAPRI meta-model. *Agric. Ecosyst. Environ.* **2009**, *133*, 267–279. [\[CrossRef\]](#)
53. Gong, X.; Xu, J. Geopolitical risk and dynamic connectedness between commodity markets. *Energy Econ.* **2022**, *110*, 106028. [\[CrossRef\]](#)
54. Mashkova, A.; Bakhtizin, A. Algorithm and data structures of the agent-based model of trade wars. In Proceedings of the 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 13–15 October 2021; pp. 1–6.
55. Poledna, S.; Miess, M.G.; Hommes, C.; Rabitsch, K. Economic forecasting with an agent-based model. *Eur. Econ. Rev.* **2023**, *151*, 104306. [\[CrossRef\]](#)
56. Butler, K.; Iloska, M.; Djurić, P.M. On counterfactual interventions in vector autoregressive models. In Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26–30 August 2024; IEEE: Piscataway, NJ, USA; pp. 1987–1991.
57. Food and Agriculture Organization of the United Nations. Available online: <http://www.fao.org/faostat/en/> (accessed on 2 August 2025).
58. UN Comtrade: International Trade Statistics. Available online: <https://comtradeplus.un.org/TradeFlow> (accessed on 2 August 2025).
59. Hodel, F.; Booth, J. *The Beta-Binomial Distribution*; Cornell University: Ithaca, NY, USA, 2023.
60. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [\[CrossRef\]](#)
61. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#)
62. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555. [\[CrossRef\]](#)
63. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. In Proceedings of the 30th Advances in Neural Information Processing Systems, Long Beach, CA, USA, 7–9 December 2017.
64. Dwivedi, V.P.; Bresson, X. A generalization of transformer networks to graphs. *arXiv* **2020**, arXiv:2012.09699.
65. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
66. Dueck, D. Affinity Propagation: Clustering Data by Passing Messages. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.