# FOODTRADE: A BENCHMARK FOR GRAPH-BASED LEARNING ON TEMPORAL FOOD TRADE NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

This work introduces FoodTrade, a dynamic temporal network dataset covering global trade in four staple grains from 1992 to 2023, comprehensively curated from multiple public sources, preprocessed, and ready-to-use for graph-based machine learning. The dataset presents a combination of challenges often encountered in machine learning, such as a noisy, low-sample, high-dimensional, and heavy-tailed structure with temporal distribution shifts and the presence of hidden confounders. With these characteristics, FoodTrade offers a new and difficult benchmark for graph-based learning at a time when the community is increasingly concerned about stagnated progress due to limitations of current benchmarks. We analyze the evolution of the link structure and the presence of network shocks, formulate predictive tasks at the node and edge levels, build an evaluation framework, and benchmark several baselines across different model classes. While temporal GNNs may seem like a natural choice for modeling the food trade network, they are not currently the best-performing model. Heuristics, econometric, and tabular ML methods often perform as well as or better than GNNs. Therefore, this dataset contributes to the call for more meaningful benchmarks that facilitate effective evaluation of graph-learning methods beyond simple solutions. We outline potential directions for better modeling of the food trade network and improving predictions of critical variables. Therefore, the dataset paves the way for developing specialized predictive tools that invite collaboration with domain experts to advance impactful data-driven AI solutions to assess global food security.

## 1 INTRODUCTION

Despite enough calories being produced globally, it is estimated that 318 million people will face crisis levels of hunger in 2027 (World Food Programme (WFP), 2025). Therefore, the urgency and significance of taking steps towards firstly assessing, and secondly, improving the state of global food security are paramount.

Food security is greatly affected by shocks in the international food trade network, which are sudden and unexpected exogenous events that disrupt trade flows. For example, shipping restrictions in the Hormuz Strait during the ongoing Iran conflict have led to a sharp increase in energy and fertilizer prices, potentially leading to lower crop yields due to decreased fertilizer use (International Food Policy Research Institute, 2026). This adds to the concern about global food price inflation and rising staple food prices over recent years. In addition to geopolitical conflicts such as Russia-Ukraine war causing trade barriers, pandemics like COVID-19 and natural disasters such as droughts and floods (now also worsened by climate change) all impact agricultural production and trade flows (FAO, 2025). Cascading effects of such shocks can significantly aggravate disruptions in the trade network and restrict access to food.

Therefore, addressing food insecurity requires a thorough study of the systemic risk in the international food trade network (Burkholz & Schweitzer, 2019). Even today, models of the international food trade network are still heavily dependent on classical economic approaches such as the Gravity model (Santos Silva & Tenreyro, 2006), or on network science descriptions (Ji et al., 2024). While graph-based machine learning methods such as temporal graph neural networks (TGNNs) have been evaluated for temporal networks across various applications such as weather (Yu et al., 2023), traffic flow (Yu et al., 2018; Guo; Zheng et al., 2020), road safety (Nippani et al., 2023),



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pandemics (Panagopoulos et al., 2021), neuroimaging (Said et al., 2023), power grids (Varbella et al., 2024), and urban knowledge graphs (Ning et al., 2023), their application in the economics domain is limited. However, current research is exploring how artificial intelligence could reshape the design of international trade policy in the coming years (Antràs, 2025). In line with this vision, we expect the development of predictive tools that encourage collaboration with domain experts to advance impactful data-driven solutions to assess systemic risk in the global trade network.

As an essential step in this direction, we introduce FoodTrade, a dynamic temporal network dataset covering global trade in four staple grains from 1992 to 2023, sourced from the Food and Agriculture Organization of the United Nations (FAO) statistical databases (FAO, 2026). While FAO's compiled food and agricultural data has been used extensively for analysis (Zhang & Zhou, 2023; Li et al., 2026; Xu et al., 2024; Zhao et al., 2025; Monken et al., 2021), its use in the machine learning community remains limited. A major obstacle to this is the several processing steps required to prepare data for predictive tasks, which FoodTrade provides.

FoodTrade is a comprehensive dataset curated from multiple public sources, pre-processed, cleaned, and presented as a dynamic temporal network dataset, ready to use for machine learning. Along with the dataset, we design and build an extensive evaluation framework applicable to several model classes, including heuristic, tabular, sequence, MPGNN, and TGNN models. We formulate tasks at the node- and edge-level, i.e., predicting the annual supply of a crop per country and the caloric trade flow between countries, respectively. These are critical variables in the network that have a direct impact on food security (Wang et al., 2025; Silvestrini et al., 2023).

Beyond enabling advanced ML methods for economic modeling, FoodTrade introduces a challenging real-world benchmark for the graph machine learning community, which currently relies heavily on benchmarks that are either synthetic or have limited complexity and structural variation (Bechler-Speicher et al., 2025; Karn & Jensen, 2025; Rubio-Madrigal & Burkholz, 2026).

Trade networks form naturally complex and dynamic graphs, influenced by various factors. As a result, this dataset poses a diverse set of challenges that distinguish it from existing datasets, including low sample size, a high noise ratio, attributed edges, heavy-tailed distributions in both features and target variables, and flow conservation. While these characteristics are often encountered in machine learning, their combination in this dataset is unique and non-trivial. Consequently, it can serve as an important benchmark for the graph learning community to evaluate design principles for sample efficiency, robustness to noise, temporal distribution shifts, variations in density, performance under skewed data, and conservation of flow. Additionally, FoodTrade is a relatively small yet challenging dataset. This lowers computational cost barriers, making AI-driven food-security research accessible to a broader, more diverse global community.

Therefore, the FoodTrade dataset serves a dual purpose: it provides, firstly, opportunities for the development of specialized GNN architectures tailored to the economic tasks, and secondly, a realistic and demanding benchmark with complex real-world tasks where the structure, attributes of nodes and edges, and temporal component all have to be leveraged.

Our main contributions are summarized as follows:

- We introduce FoodTrade[1], a comprehensive cleaned dataset modeling the complex, real-world multi-crop trade network dataset as dynamic, directed, and attributed graphs, along with detailed exploratory data analysis.
- We posit FoodTrade as a demanding benchmark for temporal graph-based learning that presents several co-existing challenges and serves as a valuable testbed for techniques to tackle them, thereby addressing the concern about limitations of current benchmarks.
- We build an evaluation framework and compare models across several classes on multiple tasks, showing that heuristic baselines often perform the best, which highlights the need for better models that go beyond simple solutions.
- We discuss intended uses of the dataset, with promising directions to improve current solutions and to define additional tasks that reflect meaningful real-world problems related to global food security.

[1]Data and code available at: `https://anonymous.4open.science/r/foodtrade-8fb3/`

Table 1: Dataset dimensions.

| # nodes | # edges | # time steps | # node attributes | # edge attributes | # tasks evaluated |
|---|---|---|---|---|---|
| 204 | 172409 | 32 | 110 | 40 | 3 |

## 2 Related work

**Temporal graph datasets.** Several recent works have contributed datasets for temporal graph learning and explored the usage of GNNs for predicting international trade flows. The DGB benchmark (Poursafaei et al., 2022) introduces six dynamic graph datasets, including `UN-Trade`, a weighted, directed, food and agriculture trading graph between 181 nations and spanning over 30 years, where edge weights represent normalized import and export values and the task is link prediction. The `tgbn-trade` (Huang et al., 2023) dataset is the UN trading network from 1986 to 2016, where edges correspond to the trade value of all agricultural products from one country to another, and the task is predicting the proportion of agricultural trade values from one nation to other nations during the next year. These datasets serve as current benchmarks for evaluating temporal models on trade-related tasks (Tian et al., 2024; Gravina et al., 2024; Li et al., 2025; Feldman et al., 2026a). However, they include only a single weight (the total trade monetary value) along each edge, which is not sufficient to encapsulate the diverse determinants of trade flow. FoodTrade includes a comprehensive set of edge attributes, including both the traded volume, as well as the value at a crop-specific level, that allow better modeling of the trade network. Finally, unlike datasets in TGB, FoodTrade's difficulty is not largely driven by scale as FoodTrade is fairly small yet challenging.

While not concerning food trade or food security, other datasets with similar characteristics (noisy, low sample, high-dimensional, heavy-tailed structure) have been proposed, such as CSN (Sapiezynski et al., 2019) or the Enron dataset (Klimt & Yang, 2004). Although the raw Enron corpus contains textual features, its constructions as temporal benchmarks commonly used in (Poursafaei et al., 2022; Huang et al., 2024; Karmim et al., 2024; Wang et al., 2021) omit these attributes. More recently, DTGB (Zhang et al., 2024) incorporated Enron's textual attributes for destination-node retrieval, edge-category classification, and textual-relation generation. CSN (also named 'Contact') has a maximum of one node and edge attribute each (Poursafaei et al., 2022). As FOodTrade differs from these datasets widely in terms of domain, temporal granularity, and node and edge attribution, it is a complementary benchmark.

**Multi-relational temporal datasets.** TGB 2.0 (Gastinger et al., 2024) introduces multi-relational temporal graph benchmarks from domains such as knowledge graphs, software, social forums, politics, and app markets. FoodTrade, adds a distinct economic-domain benchmark by modeling temporal food trade across multiple crops. Beyond single-crop prediction, the dataset naturally supports a multiplex view of the trade network, where each layer represents a crop. While multiplex modeling is outside the scope of this work, this dataset enables future work to model dependencies across crop-specific trade, such as substitution between crops, complementarity among jointly traded or consumed crops, and shared exposure to common exporters, importers, or trade corridors.

## 3 Dataset description

**Sources.** FoodTrade is sourced from the CEPII's Gravity Database (Conte et al., 2022) (OPEN LICENCE 2.0) and the United Nations' Food and Agriculture Organization FAOSTAT databases (FAO, 2026): Crop Production (QCL), Trade Summary (TCL), Trade Matrix (TM), Macro-Economic Indicators (MK), Producer Prices (PP), Population (OA), Food Balance Sheets (FBS). While the UN Comtrade repository is also a popular choice for extracting trade data, it is not always possible to release a processed dataset due to licensing constraints imposed by the UNSD, as noted by (Sellami et al., 2024). Alternatively, we source our data from the FAO statistical databases, which are distributed under CC BY 4.0 terms, and thus do not restrict further usage by the ML community.

**Graph model.** We represent the global trade of each staple crop as a *temporal, directed, attributed graph* observed at discrete yearly snapshots. For each year $t$, the network is a graph $G_t = (V_t, E_t)$,

where nodes $V_t$ correspond to countries and directed edges $E_t \subseteq V_t \times V_t$ represent trade flows from an exporter to an importer, as illustrated in Fig. 1. Each node $v \in V_t$ has an attribute vector $\mathbf{x}_v^{(t)}$ composed of country-level features at time $t$ (e.g., crop production, macroeconomic indicators, demographics, etc.). Each edge $(u, v) \in E_t$ has an attribute vector $\mathbf{z}_{uv}^{(t)}$ describing bilateral factors (e.g., distance, shared-language, etc.) and the observed trade flow representing the quantity and value of exports from $u$ to $v$.

The sequence $\{G_t\}_{t=1}^{T}$ forms a collection of temporal snapshots that captures both structural evolution (changes in the set of active trade links) and dynamic variation in node and edge attributes over time, enabling prediction tasks such as forecasting available supply of a crop per country (node regression), future edge existence (link prediction), or trade volumes across existing edges (edge regression). The data can be accessed as a sequence of snapshots or a sequence of sliding windows for a supervised machine learning task, as discussed in §4. Tables 4, 5, 6 provide the list of node and edge attributes with descriptions.

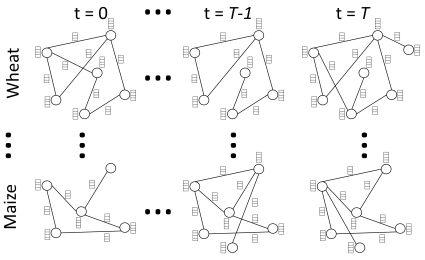


Figure 1: Global food trade network.

### 3.1 Data pre-processing

We pre-process the data in multiple steps. Firstly, the required raw data for the four staple grains in the specified time period is filtered from the downloaded bulk data and extracted into long format (where each row represents a single observation). The resulting data is further filtered in the second step to select the country (node) subset where each country has at least one crop trade link in the data.

Next, the data is cleaned to prepare the final dataset. This includes handling countries that split during the time period and/or cease to exist using an *Active* flag, removing self-loops, integer encoding categorical features, creating flags for missing features in the sourced data, imputing missing values using various strategies depending on the variable, harmonizing inconsistencies between trade quantities reported by importers and exporters, and conversion of all mass units to convert kilocalories (FAO, 2001) for analysis across different crops. We detail the strategy for imputation and reporting inconsistencies in Appendix B.

Finally, the data from FAOSTAT and the gravity database are merged. While we analyze outliers, we decide against treating them differently due to the inherent heterogeneity in the data and the heavy-tailed distributions, as discussed in §3.2.

We deliberately do not limit the dataset to only the cleaner or more robust parts of the data, but rather provide a useable dataset with maximal information, which can then be harnessed in different task-dependent ways such as to impute based on predictions, select only important features, or to engineer further node-level or global features. Robustness to intrinsic data noise is an algorithmic objective, not just a data cleaning problem.

### 3.2 Dataset characteristics

We present a non-exhaustive list of characteristics that make FoodTrade challenging and that can encourage the development of new graph-based learning methods to tackle them. Fig. 2 shows the heavy-tailed structure and temporal distribution shift in the data, while figures supporting other characteristics are deferred to Appendix C.

**Distributional and temporal shift.** Early training windows can differ markedly from later training windows and from the final test years. Consequently, models are evaluated on both their ability to predict within a stable distribution and their capacity to generalize under temporal shift.

**Long-tailed structure.** Trade volumes are unevenly distributed across country pairs, with a small number of large flows coexisting with many small or intermittent flows. Similarly, countries vary substantially in their number of trade partners, their participation across commodities, and their overall import or export volume. Performance on dominant trade relationships may obscure poor generalization on smaller, rarer, or more vulnerable flows.

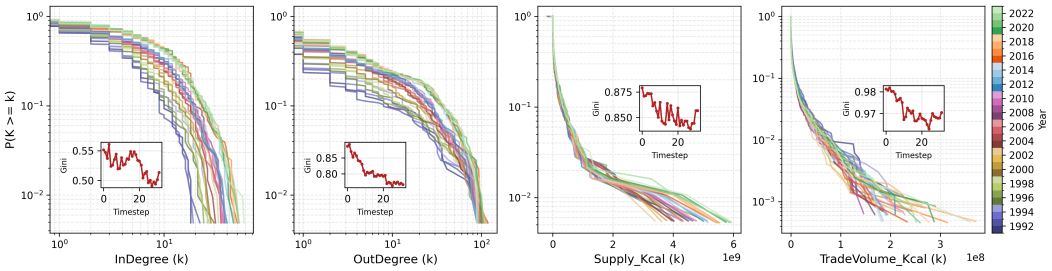



Figure 2: Temporal distribution shift in the wheat trade network structure and critical variables. The heavy-tailed structure of the degree and variable distributions remains across years.

**Network shocks.** Geopolitical events, climate anomalies, policy changes, conflicts, and macroeconomic disruptions can substantially alter patterns of food supply and demand. These events act as hidden confounders that affect both the topology and the weights of trade relationships. We provide a shock analysis in Appendix C.2

**Flow conservation.** Incoming food flows plus domestic production should, in theory, be balanced by domestic use, outgoing flows, stock changes, and losses. This is central to modeling cascades triggered by shocks, where countries may adjust trade flows in response to deficits, and therefore is particularly relevant for food security.

**Connectedness, centrality, and structural dependence.** A relatively small set of countries can occupy highly central positions in the network. These countries may act as major exporters, importers, intermediaries, or regional hubs. Correctly predicting future trade, therefore, requires more than modeling pairwise interactions independently: the graph structure itself contains information about dependency, substitution, exposure, and systemic influence.

**Fairness and heterogeneous impact.** Less economically powerful or food-insecure countries may be more vulnerable to disruptions, even when their trade volumes are relatively small. Conversely, countries with large trade volumes or high centrality may generate larger cascading effects when their trade patterns change. Therefore, fairness cannot be reduced to treating all countries identically.

**Noisy and missing data.** The dataset aggregates reports from many countries, each with different reporting standards. As a result, the data contains missing values, delays in annual reporting, and variation in the level of detail provided across countries and commodities. These issues introduce measurement noise that distinguishes this benchmark from cleaner, more controlled graph datasets.

**Limited training data.** Unlike many dynamic networks, such as financial market networks or online interaction networks, food trade data is observed at an annual resolution. Consequently, the number of temporal snapshots is small: even multi-decade datasets contain only a few dozen yearly observations. This increases the importance of inductive bias, structural modeling, and robust evaluation under data scarcity.

**High-dimensional data.** Modeling the complex food trade network requires representing not only trade volumes but also the wide range of factors that impact supply, demand, and trade constraints and decisions. We therefore construct an extensive set of unilateral (country-level) and bilateral (dyad-level) features spanning economic, political, geographic, demographic, and environmental domains. For selected variables, we also retain multiple sources.

## 4 BENCHMARKING

While many worthwhile tasks could potentially be defined on FoodTrade, we focus, in this work, on two critical variables: caloric net supply per country and bilateral trade volume for a crop.

**Node prediction: net supply.** The node regression task predicts the total supply of a crop available to a country for a given year. We define net supply as per the basic identity from the FAOSTAT Food

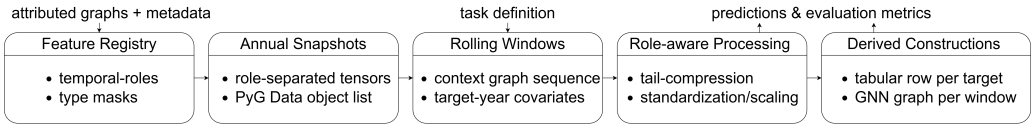



Figure 3: FoodTrade dataset model and evaluation framework overview.

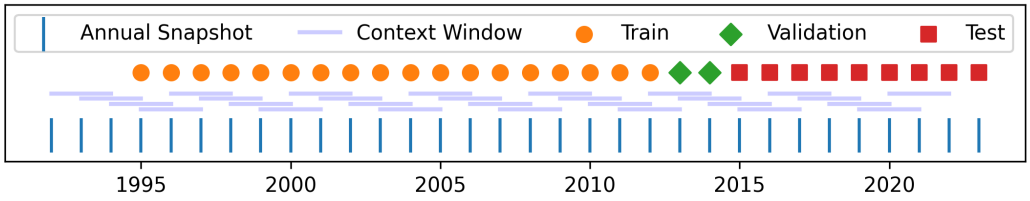



Figure 4: Rolling windows with stride and horizon one. Each window of length $l$ is the historical context $\{t-1, ..., t-l\}$ used to predict for year $t$. The last 30% of the series is reserved for testing.

Balance Sheet (FAO, 2020). Concretely, for each crop, year, and country, we compute:

$$supply = production + imports - exports - stockVariation$$

where *production* is as reported by the country, *imports* and *exports* are the total trade volumes aggregated along incoming and outgoing edges, respectively, and *stockVariation* represents internal changes in the country's reserves. *stockVariation* is negative when an item is removed from stock, i.e., added to the available supply, and positive if an item is added to stock, i.e., removed from the available supply. We also consider another version of this task, i.e. *supply per capita (supply/cap) = supply/population.* Normalizing the country size effect, this reduces the extreme heavy-tailedness of the total supply task and is a more relevant metric for food security.

**Edge prediction: trade volume.** The edge regression predicts the volume of trade of a crop per pair of countries. We assume the network structure of the target year is known, i.e., predictions are made only for edges that exist in the target year. This assumption is grounded in the standard practice of a two-step hurdle model, which is a well-established statistical and machine learning approach used for data with an excess of zero values by breaking the prediction into two sequential decisions: a participation decision (whether the value is zero or positive) and a magnitude decision (how large the value is, given that it is positive). This is suitable given the low density of the network (Fig.8, Appendix C). We isolate the two tasks, deferring link prediction to future work, as it merits independent research into sampling strategies and evaluation metrics suitable for FoodTrade.

**Other possible tasks.** Currently, we focus on the above node- and edge-level tasks. However, this dataset allows for the formulation of more meaningful tasks, also at the graph-level. We discuss some promising directions as follows: 1) Characterizing thr state of the trade network by predicting future trade links between countries and estimating global indicators like global GDP or trade volume. 2) Anomaly detection to identifying outlier years with network shocks caused by events such as pandemics and wars. 3) Shock response simulation by predicting cascade sizes in the network due to shock propagation, trade reroutes, and devising metrics for resilience to shocks. 4) classifying whether a country has sufficient nutrition available to sustain its population, given consumption patterns derived from the usage of the crop for food, population data, and an average minimum dietary requirement of 2000 Kcal/day/capita (varied based on demographics such as location, age, and gender).

A limitation for graph-level tasks is the scarce training data due to the annual granularity of the network snapshots. Given the complex nature of the network dynamics, a generative model to predict the full network in the next time step is substantially more difficult and potentially requires more data. However, such higher-order tasks can also be composed of the currently proposed standard tasks at the node and edge level. Therefore, the first challenge is to reliably model them before moving on to further tasks. Otherwise, accumulated errors can be more difficult to diagnose and address.

Table 2: Mean test sMAPE by task over 3 seeds for Wheat. **Bold**/underlined/*italic* indicates best/second-/third-best models within error. Gravity model variants are in Table 7 (§B.2).

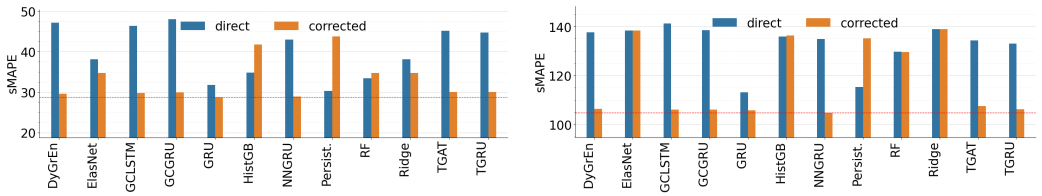
| Model | Supply/cap | Supply | TradeVolume |
|---|---|---|---|
| Persistent-forecast | 30.30 ± 0.00 | 30.37 ± 0.00 | *115.22 ± 0.00* |
| Moving-average | **29.37 ± 0.00** | **29.61 ± 0.00** | **108.81 ± 0.00** |
| ElasticNet | 38.10 ± 0.00 | 41.34 ± 0.00 | 138.36 ± 0.00 |
| HistGB | 34.86 ± 0.00 | 36.05 ± 0.00 | 135.96 ± 0.29 |
| Random-forest | 33.49 ± 0.05 | 33.53 ± 0.01 | 129.74 ± 0.05 |
| Ridge-regresion | 38.10 ± 0.00 | 41.32 ± 0.00 | 138.99 ± 0.00 |
| GRU | *32.81 ± 2.04* | *32.10 ± 0.30* | 112.85 ± 0.28 |
| GAT | 90.29 ± 5.68 | 85.55 ± 3.25 | 158.32 ± 1.35 |
| GeneralConv | 51.58 ± 1.42 | 62.82 ± 3.32 | 162.66 ± 0.32 |
| GINE | 52.52 ± 2.24 | 61.79 ± 1.27 | 120.42 ± 1.34 |
| TranformerConv | 39.38 ± 2.11 | 55.79 ± 1.30 | 120.35 ± 1.88 |
| DyGrEncoder | 47.18 ± 1.43 | 53.82 ± 3.02 | 137.88 ± 0.22 |
| GCLSTM | 45.79 ± 0.93 | 54.07 ± 4.63 | 139.48 ± 2.08 |
| GConvGRU | 48.90 ± 1.15 | 56.73 ± 3.95 | 138.43 ± 0.50 |
| GConvLSTM | 49.49 ± 1.28 | 58.25 ± 13.47 | 140.33 ± 1.50 |
| TGAT | 51.07 ± 5.14 | 50.32 ± 3.12 | 133.65 ± 0.61 |
| TGRU | 45.44 ± 0.87 | 50.45 ± 2.42 | 133.23 ± 0.67 |
| NNConvGRU | 44.02 ± 1.80 | 49.05 ± 5.05 | 133.66 ± 2.00 |

**Evaluation framework.** Fig.3 provides a high-level overview of the dataset modeling and evaluation pipeline. Given the cleaned data (i.e., annual edge list and node and edge features), we define feature-level metadata such as data types, temporal roles that distinguish static, only historically known dynamic, and future-known dynamic features, and heavy-tailed variable flags. This constitutes a central feature registry. Then, annual directed trade-graph snapshots are built as standard PyTorch Geometric Data objects. These snapshots are combined into rolling temporal windows in chronological order according to the task definition, which includes chosen crop, predictive task, context length $l$, and train/test split ratios. We illustrate the rolling window mechanism in Fig. 4.

With the training splits created, the data is can optionally be processed to apply tail-compression, such as $\log$ and standard or custom scaling, which is fit only to the train data. At this stage, the data is ready for supervised learning by temporal GNNs, which take as input a sequence of graphs. Subsequently, each window can be represented in a model-specific form. Tabular models use one row per target node or edge with lagged targets and role-aware covariates, sequence models preserve temporal ordering across context years, and graph models retain the full network structure for message passing. These constructions can be derived at varying levels of temporal feature information from the data.

**Models.** We evaluate several models ranging in variety in terms of their temporal, structural, and attribute awareness that we classify as follows. Naive baselines persistent-forecast and moving-average predict the last value and the average of the last $l$ values in the time series as the target. Tabular models include ElasticNet, HistGB, Random Forest, Ridge regression, and the sequence model GRU. Non-temporal message passing GNNs (MPGNNs) include GAT, GINE, GeneralConv, and TransformerConv. Most temporal GNNs (TGNNs) such as DyGrEncoder, GCLSTM, GConvGRU, and GConvLSTM only support a scalar edge weight, while few edge-aware TGNNs such as TGAT, TGRU, and NNGRU support multiple edge attributes. A traditional econometric baseline to forecast bilateral trade flows is the Gravity baseline and thus is only directly applicable to the edge task. It is often estimated using Poisson Pseudo-Maximum Likelihood (PPML)(Santos Silva & Tenreyro, 2006). We detail model descriptions and hyperparameter tuning in Appendix B.

**Metrics.** We use symmetric Mean Absolute Percentage Error (sMAPE) as our evaluation metric for regression. Given actual values $A$ and forecasted values $F$, sMAPE is calculated as




(a) Supply/cap (b) TradeVolume

Figure 5: Residual corrected models with persistence as the base. Direct model results from Table 2.

$\text{sMAPE}(F, A) = (2/n)\sum_{i=1}^{n} |F_i - A_i|/|A_i| + |F_i|$ and has range $[0, 200]\%$. We discuss why sMAPE is suitable for our tasks in Appendix B.1.

**Results.** Using the evaluation framework described above, Table 2 shows test performance for the three tasks on the wheat trade network. Tables 8, 9, and 10 in Appendix B.2 show test performance for soy, rice, and maize networks, which yield similar quantitative results and qualitative insights.

Naive baselines remain the best-performing across tasks. This is in line with the results by Huang et al. (2023). Ensembles such as Random Forest and HistGB outperform the GNNs on the node-level supply prediction task, whereas the sequence model GRU improves upon them, likely by using the temporal structure better. Among GNNs, edge-aware temporal GNNs outperform others as they can potentially utilize all aspects of the data better than their non-temporal counterparts. Yet, they are currently insufficient for modeling the complexities of the trade network.

We also perform residual correction, a common approach in time series forecasting, particularly for trade networks. We fix persistence-forecast as the base model and use different models to learn residuals of the baseline. This is akin to learning the change in supply and trade volumes between consecutive years. We compare the result with the same model used to learn the task, which we call the direct model. As shown in Fig.5, the corrected model has a substantially lower error than the corresponding direct model for GNNs. The improvement over persistent-forecast is smaller for the supply task (30.3% to 28.9%) than on the trade volume task (115.2% to 104.7%), both by NNGRU.

An important caveat is that for target edges not present in the previous year, a persistence-forecast predicts 0, resulting in the maximum error (sMAPE = 200%). We denote edges that persist in consecutive years (roughly 74% and 67% in the train and test set, respectively) as $E_p$ and which do not as $E_p^{'}$. Note that the models are still trained on the complete edge set, and this distinction is only to analyze how much residual correction benefits each subset, if at all.

Table 3: Test error by target edge subsets.

| target edges | | $E_p^{'}$ | $E_p$ | $E$(all) |
|---|---|---|---|---|
| persistent | | 200.00 | 84.94 | 115.22 |
| NNGRU | dir. | 152.31 | 128.36 | 134.66 |
| | cor. | 155.22 | 84.97 | 103.43 |
| | persist.($E_p$) + dir.($E_p^{'}$) | | | 102.66 |
| GGRU | dir. | 152.61 | 131.48 | 137.03 |
| | cor. | 164.5 | 85.10 | 106.00 |
| | persist.($E_p$) + dir.($E_p^{'}$) | | | 102.74 |
| TGRU | dir. | 150.61 | 125.57 | 132.16 |
| | cor. | 165.73 | 85.04 | 107.26 |
| | persist.($E_p$) + dir.($E_p^{'}$) | | | **102.22** |
| GRU | dir. | 158.42 | 96.90 | 113.09 |
| | cor. | 158.53 | 86.99 | 105.81 |
| | persist.($E_p$) + dir.($E_p^{'}$) | | | 104.29 |

Table 3 shows that for $E_p$, the corrected model performs similar to persistent-forecast whereas the direct model is much worse. For $E_p^{'}$, the corrected model is worse than the direct model for GNNs. Combining the better model for each subset, i.e. persistence for $E_p$ and the TGNN for $E_p^{'}$ improves the overall error. This simple combination model demonstrates that the data contains a heterogeneous signal and highlights how, with domain- and task-aligned strategies, heuristic baselines can be surpassed.

A similar analysis to Table 3 can be carried out with the moving-average baseline and dividing the target edge set by the number of years they appear in the context window. Furthermore, whether the direct model trained only on $E_p$ is better than the model trained on all edges, and how the efficacy of

link prediction, which removes the assumption that the target edge-set is known, affects overall results remains to be seen. Instead of using a simple binary of persistence between consecutive years, more nuanced edge subsets can be constructed, informed by our link analysis in Appendix C.1. For example, errors can be investigated based on their volume percentile. As Fig.13 shows, the percentage of $E_p$ reduces with trade volume. Given the heterogeneity in the volatility and capacity of trade links, different methods could potentially specialize across different subgroups of the distribution, leading to approaches inspired by Mixture-of-Experts that leverage model diversity and comprehensive data to model the food trade network and improve predictions of critical variables. However, this does not imply that a domain-specialized modeling approach is strictly necessary. Rather, insights obtained via FoodTrade can inform learning problems with similar challenges beyond the trade domain.

## 5 Discussion

Our results align with a broader challenge in the graph-learning community, i.e., current methods are not yet able to reliably solve complex tasks on real-world networks. Seemingly advanced methods struggle to capture features, structure, and temporal aspects simultaneously. This highlights the need, as identified by (Bechler-Speicher et al., 2025), for more complex, structured datasets, to provide benchmarks that go beyond tasks for which simple solutions are sufficient (Rubio-Madrigal & Burkholz, 2026) and where graph-learning methods can add more value.

The weak performance of current temporal GNNs may arise from several interacting factors, including architectural limitations, training strategies, overfitting to noise, and data heterogeneity. Disentangling these factors is the next step in improving upon these models. The results suggest limitations in current architectures or training strategies that obstruct them from exploiting long-term temporal patterns, covariates, and network dependencies. This also raises an interesting question about the sources of the observed error: to what extent does it reflect limitations of current models, and to what extent does it arise from uncertainty inherent in the available data, especially when the underlying ground truth is difficult to determine. From this perspective, FoodTrade may provide a useful setting not only for developing improved predictive architectures, but also for studying the intrinsic predictability of evolving trade networks, and the range of performance that can reasonably be expected. This also makes the dataset potentially interesting for the uncertainty-quantification community.

**Limitations** While FoodTrade and the associated evaluation framework provide a rich and more realistic benchmark for graph learning, several limitations exist. Firstly, the current choice of topological measures may not be able to fully capture higher-order dependencies or latent economic factors, and more advanced feature selection and engineering may be required. Moreover, currently, each crop is treated separately, so the evaluation only partially exploits the multiplex perspective of the interdependent trade networks that the data offers. Secondly, data quality and completeness also pose challenges. Reporting standards between the developing and developed world may differ significantly, leading to gaps in the data. Furthermore, predictive models may be unfairly biased towards certain countries due to their high centrality in the network, while less economically powerful or food-insecure countries may be more vulnerable to disruptions. Conversely, countries with large trade volumes or high centrality may generate larger cascading effects when their trade patterns change. Therefore, we emphasize accounting for the asymmetry in the data when designing and evaluating predictive models.

**Broader impact** The dataset and associated tasks, whether evaluated or discussed, lay the groundwork necessary to encourage the adoption of ML methods to model the international food trade network. The data can be extended to include indicators for shocks to help benchmark models for cascade prediction and systemic risk assessment (Jafari et al., 2024; T et al., 2018; Burkholz & Schweitzer, 2019). An in-depth analysis by country classification, such as Low Income Food Deficit Countries (LIFDCs) and Net Food Importing Developing Countries (NFIDCs), may help gain more insights and highlight vulnerabilities of specific groups. Incorporating causal modeling and counterfactual questions can enhance the dataset's relevance for real-world AI-aided decision making. Such models could inspire policies to help achieve food security, which is listed as a goal and 'matter of priority' by the UN in its Agenda for Sustainable Development (United Nations General Assembly, 2015). They could also drive progress in graph based machine learning to address challenging problems like the presented benchmark.

### AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, help create figures and LaTeX tables, and help code the evaluation framework, particularly its structure. We thoroughly read, understood, ran, and tested all code. We take responsibility for the final content of this work.

## REFERENCES

Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. 33.

Pol Antràs. The uncharted waters of international trade. *Journal of the European Economic Association*, 2025. Forthcoming/Working Paper.

Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M. Bronstein, Mathias Niepert, Bryan Perozzi, Mikhail Galkin, and Christopher Morris. Position: Graph learning will lose relevance due to poor benchmarks. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.14546.

Rebekka Burkholz and Frank Schweitzer. International crop trade networks: the impact of shocks and cascades. *Environmental Research Letters*, 14(11):114013, October 2019. ISSN 1748-9326. doi: 10.1088/1748-9326/ab4864. URL http://dx.doi.org/10.1088/1748-9326/ab4864.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Maddalena Conte, Pierre Cotterlaz, and Thierry Mayer. The cepii gravity database. Working Papers 2022-05, CEPII, 2022. URL https://www.cepii.fr/CEPII/en/publications/wp/abstract.asp?NoDoc=13432.

FAO. *Food Balance Sheets: A Handbook*. FAO, Rome, 2001. URL https://www.fao.org/4/x9892e/x9892e00.pdf. Accessed April 2026.

FAO. *Food Balance Sheets: A handbook*. FAO, Rome, 2020. URL https://fao.org. Accessed: 2025-04-01.

FAO. *The Impact of Disasters on Agriculture and Food Security 2025 – Digital solutions for reducing risks and impacts.* Rome. FAO, 2025. doi: 10.4060/cd7185en.

FAO. Faostat statistics database, 2026. URL https://www.fao.org/faostat/en/. Accessed: 21 April 2026.

Or Feldman, Krishna Sri Ipsit Mantri, Moshe Eliasof, and Chaim Baskin. Revisiting node affinity prediction in temporal graphs, 2026a. URL https://arxiv.org/abs/2510.06940.

Or Feldman, Krishna Sri Ipsit Mantri, Moshe Eliasof, and Chaim Baskin. Revisting node affinity prediction in temporal graphs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=6UvkemEgK3.

Benito E Flores. A pragmatic view of accuracy measurement in forecasting. *Omega*, 14(2):93–98, 1986. ISSN 0305-0483. doi: https://doi.org/10.1016/0305-0483(86)90013-7. URL https://www.sciencedirect.com/science/article/pii/0305048386900137.

Julia Gastinger, Shenyang Huang, Mikhail Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany, and Guillaume Rabusseau. Tgb 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2024. URL https://arxiv.org/abs/2406.09639.

Alessio Gravina, Giulio Lovisotto, Claudio Gallicchio, Davide Bacciu, and Claas Grohnfeldt. Long range propagation on continuous-time dynamic graphs. 2024. URL https://api.semanticscholar.org/CorpusID:270258341.

Qiang Huang, Xin Wang, Susie Xi Rao, Zhichao Han, Zitao Zhang, Yongjun He, Quanqing Xu, Yang Zhao, Zhigao Zheng, and Jiawei Jiang. Benchtemp: A general benchmark for evaluating temporal graph neural networks. In *ICDE*, pp. 4044–4057, 2024.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2023. URL https://arxiv.org/abs/2307.01026.

International Food Policy Research Institute. Conflict shocks and food systems, 2026. URL https://www.ifpri.org/landing/conflict-shocks-food-systems-blog/.

Yaghoob Jafari, Helena Engemann, and Andrea Zimmermann. The multiple dimensions of resilience in agricultural trade networks. *Q Open*, 4, 09 2024. doi: 10.1093/qopen/qoae024.

Gaojian Ji, Honglin Zhong, Harold L. Feukam Nzudie, Peng Wang, and Peipei Tian. The structure, dynamics, and vulnerability of the global food trade network. *Journal of Cleaner Production*, 434: 140439, 2024. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2023.140439. URL https://www.sciencedirect.com/science/article/pii/S0959652623045973.

Yannis Karmim, Marc Lafon, Raphael Fournier-S'niehotta, and Nicolas THOME. Supra-laplacian encoding for transformer on dynamic graphs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=vP9qAzr2Gw.

Isha Karn and David Jensen. The impact of data characteristics on gnn evaluation for detecting fake news, 2025. URL https://arxiv.org/abs/2512.06638.

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi (eds.), *Machine Learning: ECML 2004, 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004, Proceedings*, volume 3201 of *Lecture Notes in Computer Science*, pp. 217–226. Springer, 2004. doi: 10.1007/978-3-540-30115-8_22. URL https://doi.org/10.1007/978-3-540-30115-8_22.

Dongyuan Li, Shiyin Tan, Ying Zhang, Ming Jin, Shirui Pan, Manabu Okumura, and Renhe Jiang. Dyg-mamba: Continuous state space modeling on dynamic graphs, 2025. URL https://arxiv.org/abs/2408.06966.

Zebiao Li, Xueying Wu, and Chengyi Tu. The global food trade network as a complex adaptive system: A review of structure, evolution, and resilience, 2026. URL https://arxiv.org/abs/2601.12710.

Anderson Monken, Flora Haberkorn, Munisamy Gopinath, Laura Freeman, and Feras Batarseh. Graph neural networks for modeling causality in international trade. *The International FLAIRS Conference Proceedings*, 34, 04 2021. doi: 10.32473/flairs.v34i1.128485.

Yansong Ning, Hao Liu, Hao Wang, Zhenyu Zeng, and Hui Xiong. Uukg: Unified urban knowledge graph dataset for urban spatiotemporal prediction. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2023. URL https://arxiv.org/abs/2306.11443.

Abhinav Nippani, Dongyue Li, Haotian Ju, Haris N. Koutsopoulos, and Hongyang R. Zhang. Graph neural networks for road safety modeling: Datasets and evaluations for accident analysis. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2023. URL https://arxiv.org/abs/2311.00164.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer graph neural networks for pandemic forecasting. In *AAAI*, 2021. URL https://arxiv.org/abs/2009.08388.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2022. URL https://arxiv.org/abs/2207.10128.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. *CoRR*, abs/2006.10637, 2020.

Celia Rubio-Madrigal and Rebekka Burkholz. Fixed aggregation features can rival GNNs. In *Forty-third International Conference on Machine Learning*, 2026.

Anwar Said, Roza G. Bayrak, Tyler Derr, Mudassir Shabbir, Daniel Moyer, Catie Chang, and Xenofon Koutsoukos. Neurograph: Benchmarks for graph machine learning in brain connectomics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2023. URL https://arxiv.org/abs/2306.06202.

S. Salcedo-Sanz, D. Casillas-Pérez, J. Del Ser, C. Casanova-Mateo, L. Cuadra, M. Piles, and G. Camps-Valls. Persistence in complex systems. *Physics Reports*, 957:1–73, 2022. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2022.02.002. URL https://www.sciencedirect.com/science/article/pii/S0370157322000497. Persistence in complex systems.

João MC Santos Silva and Silvana Tenreyro. The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658, 2006. doi: 10.1162/rest.88.4.641.

Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):315, 2019.

Bassem Sellami, Chahinez Ounoughi, Tarmo Kalvet, Marek Tiitis, and Diego Rincon-Yanez. Harnessing graph neural networks to predict international trade flows. *Big Data and Cognitive Computing*, 8(6), 2024. ISSN 2504-2289. doi: 10.3390/bdcc8060065. URL https://www.mdpi.com/2504-2289/8/6/65.

Yunsheng Shi, Huang Zhengjie, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. pp. 1548–1554, 08 2021. doi: 10.24963/ijcai.2021/214.

Murilo Mazzotti Silvestrini, Nick W. Smith, and Flavia Mori Sarti. Evolution of global food trade network and its effects on population nutritional status. *Current Research in Food Science*, 6: 100517, 2023. ISSN 2665-9271. doi: https://doi.org/10.1016/j.crfs.2023.100517. URL https://www.sciencedirect.com/science/article/pii/S2665927123000850.

Distefano T, Laio F, Ridolfi L, and Schiavo S. Shock transmission in the international food trade network, 2018. URL https://doi.org/10.1371/journal.pone.0200639.

Yuxing Tian, Yiyan Qi, and Fan Guo. Freedyg: Frequency enhanced continuous-time dynamic graph model for link prediction. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:269530723.

United Nations General Assembly. Transforming our world: the 2030 agenda for sustainable development. Resolution A/RES/70/1, United Nations, New York, October 2015. URL https://undocs.org/A/RES/70/1.

Anna Varbella, Kenza Amara, Blazhe Gjorgiev, Mennatallah El-Assady, and Giovanni Sansavini. Powergraph: A power grid benchmark dataset for graph neural networks. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2024. URL https://arxiv.org/abs/2402.02827.

Xiaopeng Wang, Chengyi Tu, Shuhao Chen, Sicheng Wang, Ying Fan, Samir Suweis, and Paolo D'Odorico. Quantifying global food trade: A net caloric content approach to food trade network analysis. *Agricultural Systems*, 230:104475, 2025. ISSN 0308-521X. doi: https://doi.org/10.1016/j.agsy.2025.104475. URL https://www.sciencedirect.com/science/article/pii/S0308521X2500215X.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KYPz4YsCPj.

World Food Programme (WFP). WFP 2026 global outlook. hunger and hope: Innovative solutions to address food insecurity. Technical report, World Food Programme, Rome, 2025. URL https://wfp.tind.io/record/131186.

Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rJeW1yHYwH.

Hao Xu, Niu Niu, Dongmei Li, and Chengjie Wang. A dynamic evolutionary analysis of the vulnerability of global food trade networks. *Sustainability*, 16(10), 2024. ISSN 2071-1050. doi: 10.3390/su16103998. URL https://www.mdpi.com/2071-1050/16/10/3998.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI-2018, pp. 3634–3640. International Joint Conferences on Artificial Intelligence Organization, July 2018. doi: 10.24963/ijcai.2018/505. URL http://dx.doi.org/10.24963/ijcai.2018/505.

Sungduk Yu, Zeyuan Hu, Akshay Subramaniam, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C. Will, Gunnar Behrens, Julius J. M. Busecke, Nora Loose, Charles I. Stern, Tom Beucler, Bryce Harrop, Helge Heuer, Benjamin R. Hillman, Andrea Jenney, Nana Liu, Alistair White, Tian Zheng, Zhiming Kuang, Fiaz Ahmed, Elizabeth Barnes, Noah D. Brenowitz, Christopher Bretherton, Veronika Eyring, Savannah Ferretti, Nicholas Lutsko, Pierre Gentine, Stephan Mandt, J. David Neelin, Rose Yu, Laure Zanna, Nathan Urban, Janni Yuval, Ryan Abernathey, Pierre Baldi, Wayne Chuang, Yu Huang, Fernando Iglesias-Suarez, Sanket Jantre, Po-Lun Ma, Sara Shamekh, Guang Zhang, and Michael Pritchard. Climsim-online: A large multi-scale dataset and framework for hybrid ml-physics climate emulation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track.*, 2023. URL https://arxiv.org/abs/2306.08754.

Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang, Jie Shao, and Rex Ying. Dtgb: A comprehensive benchmark for dynamic text-attributed graphs. *arXiv preprint arXiv:2406.12072*, 2024.

Yin-Ting Zhang and Wei-Xing Zhou. Quantifying the status of economies in international crop trade networks: A correlation structure analysis of various node-ranking metrics. *Chaos, Solitons & Fractals*, 172:113567, 2023. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2023.113567. URL https://www.sciencedirect.com/science/article/pii/S096007792300468X.

Zhimeng Zhao, Lili Xu, Haoyan Ma, Xuesong Zhang, and Liping Tang. The evolution of global food trade systems and their resilience in response to covid-19: Performance across nations. *Agriculture*, 15(16), 2025. ISSN 2077-0472. doi: 10.3390/agriculture15161761. URL https://www.mdpi.com/2077-0472/15/16/1761.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction, 2020. URL https://arxiv.org/abs/1911.08415.

# A Dataset details

## A.1 Data Cleaning

**Imputation** We use linear interpolation for time-varying country-level macroeconomic indicators and demographic features and forward filling (or Last Observation Carried Forward (LOCF)) for features from the gravity database, as they are mostly static, unlikely to change commonly, or their last known value is a simple, reasonable forecast. The remaining features, mostly crop-level regarding production and trade, are conservatively imputed with 0 to enable numerical computation. We report the missing rate in Fig. 6 in Appendix C.

**Reporting inconsistencies.** To address inconsistencies between trade quantities and values reported by importers and exporters, we construct harmonized trade volume and value variables. When both parties report non-zero values, the harmonized measure is defined as the mean of the two reports, whereas when only one party reports, we use the reporting party's value.

## A.2 Feature metadata

We also list and describe the node and edge features in Tables 4, 5, and 6. Additional data is in CSV metadata files with the following columns are provided:

- Feature - name of feature
- Temporal Role - categorize the temporality of the feature into static, dynamic historical context, and dynamic known future
- Data Type - ID (identifier), NUM (numeric), BIN (binary), CAT (categorical), YEAR
- Heavy Tail Flag - An umbrella flag for whether the distribution of the feature is (usually) right-skewed with an apparent heavy tail. This allows preprocessing to handle the feature by optionally compressing the tail using one of log, asinh, box-cox, yeo-johnson transformations.
- Source - FAOSTAT (Database name: QCL, TCL, FBS, FBS, PP, MK, TM) , CEPII-Gravity, or Custom (computed by authors)
- Description - brief summary of what the feature represents

Table 4: Node features. YEAR means that the feature is also used in target year.

| Feature | Type | Source | Description | YEAR |
|---|---|---|---|---|
| Country | ID | FAOSTAT | Country | |
| CountryCode_M49 | ID | FAOSTAT / CEPII-Gravity | Country's M_49 Identifier Code | |
| CountryCode_ISO3 | ID | FAOSTAT / CEPII-Gravity | Country's ISO_3 Identifier Code | |
| Crop | ID | FAOSTAT | Crop | |
| Year | NUM | FAOSTAT | Year | |
| DomesticSupplyFB_Kcal | NUM | FAOSTAT FBS | Domestic supply quantity from FBS | |
| ExpFB_Kcal | NUM | FAOSTAT FBS | Export Quantity (from FBS for crop+products) | |
| ExpQuantity_Kcal | NUM | FAOSTAT TCL | Export Quantity (reported by country) | |
| ExpValue_K$ | NUM | FAOSTAT TCL | Export Value (reported by country) | |
| FatSupply/cap/d_Kcal | NUM | FAOSTAT FBS | Fat supply quantity per capita per day from FBS | |
| FatSupply_Kcal | NUM | FAOSTAT FBS | Fat supply quantity | |
| HarvestArea_ha | NUM | FAOSTAT QCL | Area harvested | |
| ImpFB_Kcal | NUM | FAOSTAT FBS | Import Quantity (from FBS for crop+products) | |
| ImpQuantity_Kcal | NUM | FAOSTAT TCL | Import Quantity (reported by country) | |
| ImpValue_K$ | NUM | FAOSTAT TCL | Import Value (reported by country) | |
| Losses_Kcal | NUM | FAOSTAT FBS | Losses | |
| Processed_Kcal | NUM | FAOSTAT FBS | Processed | |
| ProdFB_Kcal | NUM | FAOSTAT FBS | Production (from FBS for crop+products) | x |
| ProdPrice/t_$ | NUM | FAOSTAT PP | Producer Price (USD/tonne) | |
| ProdPriceIdx_ | NUM | FAOSTAT PP | Producer Price Index (2014-2016 = 100) | |
| Production_Kcal | NUM | FAOSTAT QCL | Production | x |
| ProteinSupply/cap/d_Kcal | NUM | FAOSTAT FBS | Protein supply quantity per capita per day from FBS | |
| ProteinSupply_Kcal | NUM | FAOSTAT FBS | Protein supply quantity | |
| Residuals_Kcal | NUM | FAOSTAT FBS | Residuals | |
| StockVar_Kcal | NUM | FAOSTAT FBS | Stock Variation | |
| SupplyFB/cap_Kcal | NUM | FAOSTAT FBS | Domestic supply quantity per capita from FBS | |
| UseFeed_Kcal | NUM | FAOSTAT FBS | Used for Feed | |
| UseFood_Kcal | NUM | FAOSTAT FBS | Used for Food | |
| UseOther_Kcal | NUM | FAOSTAT FBS | Other uses (non-food) | |
| UseSeed_Kcal | NUM | FAOSTAT FBS | Used for Seed | |
| UseTourist_Kcal | NUM | FAOSTAT FBS | Tourist consumption | |
| Yeild_kg/ha | NUM | FAOSTAT QCL | Yield | |
| AggrExpVol_Kcal | NUM | FAOSTAT TM | Summed Export Quantity (over outgoing edges in trade matrix) | |
| AggrExpVal_K$ | NUM | FAOSTAT TM | Summed Export Value (over outgoing edges in trade matrix) | |
| AggrImpVol_Kcal | NUM | FAOSTAT TM | Summed Import Quantity (over incoming edges in trade matrix) | |
| AggrImpVal_K$ | NUM | FAOSTAT TM | Summed Import Value (over incoming edges in trade matrix) | |
| GDP_AnnGrowth/cap_% | NUM | FAOSTAT MK | Annual growth US$ per capita | |
| GDP_AnnGrowth2015/cap_% | NUM | FAOSTAT MK | Annual growth US$ per capita, 2015 prices | |
| GDP_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| GDP_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| GDP_Val/cap_$ | NUM | FAOSTAT MK | Value US$ per capita | |
| GDP_Val2015/cap_$ | NUM | FAOSTAT MK | Value US$ per capita, 2015 prices | |
| GDP_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| GDP_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| GFCF_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| GFCF_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| GFCF_ShareGDP2015_% | NUM | FAOSTAT MK | Share of GDP US$, 2015 prices | |
| GFCF_ShareGDP_% | NUM | FAOSTAT MK | Share of GDP US$ | |
| GFCF_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| GFCF_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| GNI_AnnGrowth/cap_% | NUM | FAOSTAT MK | Annual growth US$ per capita | |
| GNI_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| GNI_Val/cap_$ | NUM | FAOSTAT MK | Value US$ per capita | |
| GNI_Val_M$ | NUM | FAOSTAT MK | Value US$ | |

Table 5: Node features. Continued.

| Feature | Type | Source | Description | YEAR |
|---|---|---|---|---|
| GO(AFF)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$ | |
| GO(AFF)_Val2015_M$ | NUM | FAOSTAT MK | Value US$ | |
| GO(Agr)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$ | |
| GO(Agr)_Val2015_M$ | NUM | FAOSTAT MK | Value US$ | |
| Pop_Female_K | NUM | FAOSTAT OA | Total Population - Both Sexes | x |
| Pop_Male_K | NUM | FAOSTAT OA | Total Population - Male | x |
| Pop_Rural_K | NUM | FAOSTAT OA | Total Population - Female | x |
| Pop_Total_K | NUM | FAOSTAT OA | Rural population | x |
| Pop_Urban_K | NUM | FAOSTAT OA | Urban population | x |
| VA(AFF)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(AFF)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(AFF)_ShareGDP2015_% | NUM | FAOSTAT MK | Share of GDP US$, 2015 prices | |
| VA(AFF)_ShareGDP_% | NUM | FAOSTAT MK | Share of GDP US$ | |
| VA(AFF)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(AFF)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| VA(Agr)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(Agr)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(Agr)_ShareGDP_% | NUM | FAOSTAT MK | Share of GDP US$ | |
| VA(Agr)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(Agr)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| VA(ManFB)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(ManFB)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(ManFB)_RatioVA(AFF)_ | NUM | FAOSTAT MK | Ratio of Value Added (Agriculture, Forestry and Fishing) US$ | |
| VA(ManFB)_ShareVA(TotMan)_% | NUM | FAOSTAT MK | Share of Value Added (Total Manufacturing) US$ | |
| VA(ManFB)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(ManFB)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| VA(ManFBTP)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(ManFBTP)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(ManFBTP)_RatioVA(AFF)_ | NUM | FAOSTAT MK | Ratio of Value Added (Agriculture, Forestry and Fishing) US$ | |
| VA(ManFBTP)_ShareGDP2015_% | NUM | FAOSTAT MK | Share of GDP US$, 2015 prices | |
| VA(ManFBTP)_ShareGDP_% | NUM | FAOSTAT MK | Share of GDP US$ | |
| VA(ManFBTP)_ShareVA(TotMan)_% | NUM | FAOSTAT MK | Share of Value Added (Total Manufacturing) US$ | |
| VA(ManFBTP)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(ManFBTP)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| VA(ManTP)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(ManTP)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(ManTP)_RatioVA(AFF)_ | NUM | FAOSTAT MK | Ratio of Value Added (Agriculture, Forestry and Fishing) US$ | |
| VA(ManTP)_ShareVA(TotMan)_% | NUM | FAOSTAT MK | Share of Value Added (Total Manufacturing) US$ | |
| VA(ManTP)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(ManTP)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| VA(TotMan)_AnnGrowth2015_% | NUM | FAOSTAT MK | Annual growth US$, 2015 prices | |
| VA(TotMan)_AnnGrowth_% | NUM | FAOSTAT MK | Annual growth US$ | |
| VA(TotMan)_RatioVA(AFF)_ | NUM | FAOSTAT MK | Ratio of Value Added (Agriculture, Forestry and Fishing) US$ | |
| VA(TotMan)_ShareGDP2015_% | NUM | FAOSTAT MK | Share of GDP US$, 2015 prices | |
| VA(TotMan)_ShareGDP_% | NUM | FAOSTAT MK | Share of GDP US$ | |
| VA(TotMan)_Val2015_M$ | NUM | FAOSTAT MK | Value US$, 2015 prices | |
| VA(TotMan)_Val_M$ | NUM | FAOSTAT MK | Value US$ | |
| Active | CAT | FAOSTAT / CEPII-Gravity | 1 if country exists currently | x |
| Region | CAT | UNSD | geographical region | x |
| gmt_offset_2020 | NUM | CEPII-Gravity | GMT offset in 2020 of the country (hours) | x |
| legal_old | CAT | CEPII-Gravity | Origin of the country's legal system. Possible values are british, french, german, scandinavian, or socialist | x |
| legal_new | CAT | CEPII-Gravity | Origin of the country's legal system, after the fall of the USSR. Possible values are british, french, german, scandinavian, or socialist | x |
| gatt | CAT | CEPII-Gravity | 1 if the country is a GATT member | x |
| wto | CAT | CEPII-Gravity | 1 if the country is a WTO member | x |
| eu | CAT | CEPII-Gravity | 1 if country is a EU member | x |
| entry_cost | NUM | CEPII-Gravity | Cost of business start-up procedures (% of GNI per capita) | x |
| entry_proc | NUM | CEPII-Gravity | Number of start-up procedures to register a business | x |
| entry_time | NUM | CEPII-Gravity | Days required to start a business. | x |
| entry_tp | NUM | CEPII-Gravity | Days required to start a business + number of procedures to start a business | x |

Table 6: Edge features. YEAR means that the feature is also used in target year.

| Feature | Type | Source | Description | YEAR |
|---|---|---|---|---|
| CountryExp | ID | FAOSTAT | Exporter Country | |
| CountryImp | ID | FAOSTAT | Importer Country | |
| CountryCodeM49Exp | ID | FAOSTAT / CEPII-Gravity | Exporter Country's M49 Code | |
| CountryCodeM49Imp | ID | FAOSTAT / CEPII-Gravity | Importer Country's M49 Code | |
| CountryCodeISO3Exp | ID | FAOSTAT / CEPII-Gravity | Exporter Country's ISO3 Code | |
| CountryCodeISO3Imp | ID | FAOSTAT / CEPII-Gravity | Importer Country's ISO3 Code | |
| Crop | ID | FAOSTAT | Crop | |
| Year | NUM | FAOSTAT | Year | |
| TradeVolByExp_Kcal | NUM | FAOSTAT TM | Trade Volume reported by exporter | |
| TradeValByExp_K$ | NUM | FAOSTAT TM | Trade Value reported by exporter | |
| TradeVolByImp_Kcal | NUM | FAOSTAT TM | Trade Volume reported by importer | |
| TradeValByImp_K$ | NUM | FAOSTAT TM | Trade Value reported by importer | |
| MeanTradeVol_Kcal | NUM | FAOSTAT TM | Mean of Trade Volume reported by exporter and importer if both parties report, otherwise equal to only the reporting party's value | |
| MeanTradeVal_K$ | NUM | FAOSTAT TM | Mean of Trade Value reported by exporter and importer if both parties report, otherwise equal to only the reporting party's value | |
| heg_o | BIN | CEPII-Gravity | 1 if origin is current or former hegemon of destination | x |
| heg_d | BIN | CEPII-Gravity | 1 if destination is current or former hegemon of origin | x |
| distw_harmonic | NUM | CEPII-Gravity | Population-weighted distance between most populated cities (harmonic mean) | x |
| distw_arithmetic | NUM | CEPII-Gravity | Population-weighted distance between most populated cities (arithmetic mean) | x |
| distw_harmonic_jh | NUM | CEPII-Gravity | Weighted distance by J. Hinz (harmonic mean) | x |
| distw_arithmetic_jh | NUM | CEPII-Gravity | Weighted distance by J. Hinz (arithmetic mean) | x |
| dist | NUM | CEPII-Gravity | Geodesic distance between most populated cities (km) | x |
| distcap | NUM | CEPII-Gravity | Geodesic distance between capital cities (km) | x |
| contig | BIN | CEPII-Gravity | Dummy equal to 1 if countries are contiguous | x |
| diplo_disagreement | NUM | CEPII-Gravity | UN diplomatic disagreement score | x |
| scaled_sci_2021 | NUM | CEPII-Gravity | Social connectedness index in 2021 | x |
| comlang_off | BIN | CEPII-Gravity | 1 if countries share common official or primary language | x |
| comlang_ethno | BIN | CEPII-Gravity | 1 if countries share a common language spoken by at least 9% of the population | x |
| comcol | BIN | CEPII-Gravity | 1 if countries share a common colonizer post 1945 | x |
| col45 | BIN | CEPII-Gravity | 1 if countries are or were in colonial relationship post 1945 | x |
| comleg_pretrans | BIN | CEPII-Gravity | 1 if countries share common legal origins before 1991 | x |
| comleg_posttrans | BIN | CEPII-Gravity | 1 if countries share common legal origins after 1991 | x |
| transition_legalchange | BIN | CEPII-Gravity | 1 if common legal origin changed in 1991 | x |
| comrelig | NUM | CEPII-Gravity | Religious proximity index | x |
| col_dep_ever | BIN | CEPII-Gravity | 1 if pair ever was in colonial or dependency relationship (including before 1948) | x |
| col_dep | BIN | CEPII-Gravity | 1 if pair currently in colonial or dependency relationship | x |
| col_dep_end_year | BIN | CEPII-Gravity | Independence year from concerned hegemon (includes colonial ties before 1948) | x |
| col_dep_end_conflict | BIN | CEPII-Gravity | 1 if independence from the concerned hegemon involved a conflict | x |
| sibling_ever | BIN | CEPII-Gravity | 1 if pair ever had the same colonizer (including before 1948) | x |
| sibling | BIN | CEPII-Gravity | 1 if pair currently has the same colonizer | x |
| sever_year | NUM | CEPII-Gravity | Severance year for pairs that ever had the same colonizer (includes colonial ties before 1948): corresponds to the independence year of the first independent sibling | x |
| sib_conflict | BIN | CEPII-Gravity | 1 if pair ever had the same colonizer and independence involved a conflict with the hegemon (includes colonial ties before 1948) | x |
| fta_wto | BIN | CEPII-Gravity | 1 if the country pair is engaged in a regional trade agreement, source WTO supplemented by Thierry Mayer | x |
| fta_wto_raw | BIN | CEPII-Gravity | 1 if the country pair is engaged in a regional trade agreement, source WTO | x |
| rta_coverage | CAT | CEPII-Gravity | Coverage of the trade agreement. 0 = "no trade agreement". 1 = "goods only", 2 = "services only", 3 = "goods and services" | |
| rta_type | CAT | CEPII-Gravity | Type of trade agreement. PSA = "Partial Scope Agreement", FTA = "Free Trade Agreement", CU = "Customs Union", EIA = "Economic Integration Agreement". | |
| manuf_tradeflow_baci | NUM | CEPII-Gravity | Trade flow of manufactured goods (in thousands current US$) (source: BACI) | |
| MeanTradeVolAsFracOfAggrExp | NUM | FAOSTAT TM | The volume along a trade link as a fraction of the exporting country's total exports | |
| MeanTradeVolAsFracOfAggrImp | NUM | FAOSTAT TM | The volume along a trade link as a fraction of the importing country's total imports | |

# B ADDITIONAL EVALUATION DETAILS

## B.1 EXPERIMENTAL SETUP

We run the benchmarking on a 48GB NVIDIA RTX A6000 GPU in an internal cluster. We describe the evaluated models and their chosen hyperparameters as follows:

**Naive baselines** The two most simple baselines for temporal series are `persistent forecast`(Salcedo-Sanz et al., 2022; Huang et al., 2023) and `moving_average`. `persistent forecast` assigns the value of the previous timestep to the future predicted value; that is, to assume that the series does not change, whereas `moving_average` averages the last $k$ values of the series. In TGB (Huang et al., 2023), the two naive baselines have been the best performing baseline, until a recent node affinity-focused model upranked them (Feldman et al., 2026b).

**Gravity variants** This is a traditional econometric baseline to forecast bilateral trade flows, and thus is only directly applicable to the edge tasks. It is often estimated using Poisson Pseudo-Maximum Likelihood (PPML)(Santos Silva & Tenreyro, 2006). In its simplest form, trade between two countries is modeled as proportional to the GDP of the two countries and inversely proportional to the distance between them. We denote this model as `Gravity-simple`. It is common to include additional unilateral (e.g. population, crop production, etc.) and bilateral (e.g. shared language, borders, trade agreements, etc.) covariates. We denote the model using such features from the CEPII Gravity data as `Gravity-cepii`. Finally, we include all additional edge features and node features of both countries as covariates in the `Gravity-all-feats` model.

**Tabular models** We evaluate several classical tabular regression baselines, including Ridge, ElasticNet, Random Forest, and histogram gradient boosting (HistGB). The tabular constructions input to these models are derived from the data at different levels, which serves as a hyperparameter. Specifically, i) 'target' uses lagged target values only, ii) 'node-only' uses lagged node features (for the edge task, node features of both exporter and importer are used), iii) 'node' uses both lagged target and node feature values. For edge tasks, additional options are iv) 'edge-only' uses only the lagged edge features and finally v) 'edge' uses lagged targets, edge features and node features of both endpoints. These baselines do not perform graph message passing and serve as non-graph, non-temporal tabular comparators.

**GRU** This baseline is a non-graph GRU sequence model over the tabular temporal features (Cho et al., 2014). It builds a fixed-length feature sequence for each node or edge, then applies a GRU across the temporal dimension. The final hidden state is passed through an MLP readout to predict the target.

**Non-temporal GNNs** We use four non-temporal edge-feature-aware message-passing GNN layers to evaluate four baselines: GAT, GINE, GeneralConv, TranformerConv. The model operates on a single graph construction, with lagged values optionally added as node and edge features, of each temporal window.

**Temporal GNNs** Standard temporal GNN architectures that only support edge-weight scalars include GConvGRU, GConvLSTM, GCLSTM, and DyGrEncoder. For edge-aware models, we use the edge-feature-aware graph transformer layers (Shi et al., 2021) inspired by (Xu et al., 2020; Rossi et al., 2020). TGRU and NNConvGRU firstly use transformer-style graph attention at each timestep and edge-conditioned message passing, respectively, followed by a GRU recurrence along the temporal dimension. TGAT also encodes each graph snapshot with graph attention, but instead of a GRU, it aggregates the resulting node representations across the temporal window using multi-head attention. We choose general-purpose temporal GNNs and not ones specialized for other domains such as traffic or climate modeling.

**Hyperparameter tuning** Temporal GNNs were thoroughly hyperparameter-tuned. We tune the learning rate and dropout, with learning rate in $\{10^{-2}, 10^{-3}, 10^{-4}\}$, and dropout in $\{0, 0.2, 0.5\}$. The best hyperparameters for the two main tasks are $(10^{-3}, 0.2)$. Other hyperparameters include hidden dimension $(512)$, layers $(3)$, heads $(4)$, and epochs $(1000)$. The other models are baselines that

Table 7: Test performance for gravity model. Note that all standard deviations are 0.

| Feature Set | Wheat | Maize | Rice | Soy |
|---|---|---|---|---|
| additional_features | **138.66** | **157.33** | **158.06** | **148.88** |
| cepii_covariates | 153.34 | 171.07 | 168.74 | 164.01 |
| simple | 159.03 | 178.10 | 182.07 | 177.51 |

Table 8: Test performance by task over 3 seeds for Maize.

| Model | AggrSupply/cap_Kcal | AggrSupply_Kcal | MeanTradeVol_Kcal |
|---|---|---|---|
| persistent-forecast | $25.45 \pm 0.00$ | $25.57 \pm 0.00$ | $113.86 \pm 0.00$ |
| moving-average | $25.02 \pm 0.00$ | $25.45 \pm 0.00$ | $110.49 \pm 0.00$ |
| elasticnet | $35.78 \pm 0.00$ | $46.74 \pm 0.00$ | $162.69 \pm 0.00$ |
| histgb | $36.17 \pm 0.00$ | $38.80 \pm 0.00$ | $160.28 \pm 2.84$ |
| random-forest | $30.22 \pm 0.02$ | $31.38 \pm 0.09$ | $141.01 \pm 0.06$ |
| ridge | $35.78 \pm 0.00$ | $46.71 \pm 0.00$ | $162.69 \pm 0.00$ |
| gru | $25.03 \pm 0.67$ | $25.50 \pm 0.14$ | $110.75 \pm 1.85$ |
| mpgnn-gat | $88.74 \pm 3.56$ | $86.94 \pm 2.99$ | $178.52 \pm 1.02$ |
| mpgnn-general | $43.92 \pm 0.74$ | $71.53 \pm 3.35$ | $182.49 \pm 0.92$ |
| mpgnn-gine | $57.09 \pm 1.71$ | $63.45 \pm 2.61$ | $117.40 \pm 0.38$ |
| mpgnn-transformer | $31.24 \pm 2.96$ | $54.54 \pm 3.97$ | $119.54 \pm 0.98$ |
| dygrencoder | $51.99 \pm 3.25$ | $51.15 \pm 4.80$ | $138.64 \pm 0.59$ |
| gclstm | $52.13 \pm 1.25$ | $50.36 \pm 1.54$ | $136.97 \pm 0.86$ |
| gconvgru | $50.91 \pm 1.96$ | $51.52 \pm 3.83$ | $138.17 \pm 0.94$ |
| gconvlstm | $52.48 \pm 2.13$ | $48.41 \pm 2.67$ | $138.86 \pm 0.93$ |
| tgat | $46.24 \pm 1.80$ | $43.39 \pm 2.51$ | $136.77 \pm 0.92$ |
| tgru | $67.20 \pm 17.24$ | $135.63 \pm 2.09$ | $55.35 \pm 8.91$ |

either have relatively few hyperparameters or are expected to generally perform worse than TGNNs, such as a plain GNN without a temporal component. Given the limited performance of the stronger model relative to the persistence baseline, we decided against extensively tuning models that received less information. We therefore tuned only basic parameters, such as model width and learning rate, sufficiently to ensure stable training. Further tuning of the remaining hyperparameters did not yield meaningful improvements on the validation set.

**Metric** sMAPE is suitable for our tasks because the target values vary substantially in scale, and a symmetric normalization evaluates the errors relative to the magnitudes of both observed and predicted flows. This avoids a strong dependence on the observed value, which is inherent in MAPE, and, unlike MAPE, assigns equal error to over- and under-estimations that are equivalent by a multiplicative factor (Flores, 1986). However, true target values of zero lead to maximum error, which could disproportionately distort aggregated results. Since non-positive net crop supply in the data ($11.2\%$ of values in the wheat trade) typically reflects reporting inconsistencies rather than physical realities, we compute sMAPE only over positive targets. We avoid trying to correct these anomalies by adjusting trade volumes or stock variations, as this could likely introduce cascading violations of flow conservation.

### B.2 Additional results

We report results for the wheat trade network in the main text in 2, and for the soy, maize, and rice networks as follows in Tables 10, 8, and 9, respectively.

Table 9: Test performance by task over 3 seeds for Rice.

| Model | AggrSupply/cap_Kcal | AggrSupply_Kcal | MeanTradeVol_Kcal |
|---|---|---|---|
| persistent-forecast | 56.47 ± 0.00 | 56.59 ± 0.00 | 129.82 ± 0.00 |
| moving-average | 53.31 ± 0.00 | 53.50 ± 0.00 | 122.98 ± 0.00 |
| elasticnet | 60.98 ± 0.00 | 79.24 ± 0.00 | 173.47 ± 0.00 |
| histgb | 57.24 ± 0.00 | 64.05 ± 0.00 | 166.50 ± 0.00 |
| random-forest | 54.66 ± 0.02 | 55.37 ± 0.09 | 147.31 ± 0.02 |
| ridge | 60.98 ± 0.00 | 79.47 ± 0.00 | 173.47 ± 0.00 |
| gru | 52.57 ± 1.12 | 53.81 ± 0.14 | 122.47 ± 1.10 |
| mpgnn-gat | 117.74 ± 1.63 | 117.77 ± 2.99 | 179.00 ± 3.13 |
| mpgnn-general | 78.45 ± 3.44 | 108.07 ± 3.35 | 185.37 ± 0.22 |
| mpgnn-gine | 82.56 ± 0.77 | 86.16 ± 2.61 | 123.09 ± 0.79 |
| mpgnn-transformer | 65.12 ± 2.59 | 80.74 ± 3.97 | 123.08 ± 0.23 |
| dygrencoder | 72.04 ± 1.37 | 72.04 ± 4.80 | 138.80 ± 1.68 |
| gclstm | 71.52 ± 0.47 | 70.41 ± 1.54 | 138.31 ± 0.82 |
| gconvgru | 73.67 ± 1.54 | 76.78 ± 3.83 | 136.67 ± 0.46 |
| gconvlstm | 75.95 ± 0.51 | 71.89 ± 2.67 | 138.86 ± 0.56 |
| tgat | 68.46 ± 0.97 | 73.37 ± 2.51 | 137.06 ± 0.81 |
| tgru | 70.50 ± 0.77 | 72.01 ± 2.09 | 134.24 ± 0.53 |

Table 10: Test performance by task over 3 seeds for Soy.

| Model | AggrSupply/cap_Kcal | AggrSupply_Kcal | MeanTradeVol_Kcal |
|---|---|---|---|
| persistent-forecast | 57.60 ± 0.00 | 57.74 ± 0.00 | 116.73 ± 0.00 |
| moving-average | 57.06 ± 0.00 | 57.30 ± 0.00 | 112.55 ± 0.00 |
| elasticnet | 74.09 ± 0.00 | 82.96 ± 0.00 | 173.52 ± 0.00 |
| histgb | 72.85 ± 0.00 | 78.27 ± 0.00 | 158.09 ± 7.55 |
| random-forest | 66.48 ± 0.10 | 70.30 ± 0.04 | 139.77 ± 0.04 |
| ridge | 74.09 ± 0.00 | 82.96 ± 0.00 | 173.51 ± 0.00 |
| gru | 58.07 ± 0.59 | 61.77 ± 1.87 | 112.98 ± 1.38 |
| mpgnn-gat | 120.35 ± 0.84 | 121.42 ± 1.74 | 172.25 ± 3.70 |
| mpgnn-general | 102.89 ± 4.17 | 113.36 ± 3.75 | 187.56 ± 0.36 |
| mpgnn-gine | 92.12 ± 2.16 | 99.17 ± 2.23 | 119.18 ± 1.06 |
| mpgnn-transformer | 79.12 ± 4.94 | 85.95 ± 4.70 | 119.63 ± 1.41 |
| dygrencoder | 94.58 ± 0.84 | 92.38 ± 5.52 | 140.50 ± 1.00 |
| gclstm | 95.31 ± 0.29 | 83.29 ± 0.92 | 140.02 ± 1.27 |
| gconvgru | 92.79 ± 3.32 | 89.66 ± 3.34 | 140.66 ± 1.20 |
| gconvlstm | 100.31 ± 9.07 | 87.81 ± 1.88 | 140.21 ± 1.47 |
| tgat | 82.47 ± 0.54 | 84.73 ± 4.48 | 115.45 ± 2.21 |
| tgru | 92.14 ± 0.41 | 87.91 ± 4.87 | 134.00 ± 0.19 |

## C ADDITIONAL DATA ANALYSIS

**Missing Rate.** Figure 6 shows the missing rate in the data for all crop trade networks. The observed increase in missing values after 2021 arises due to the CEPII Gravity dataset being available only until 2021. While most variables are static and we use LOCF imputation to fill data for 2022 and 2023, we indicate the data as missing for transparency.

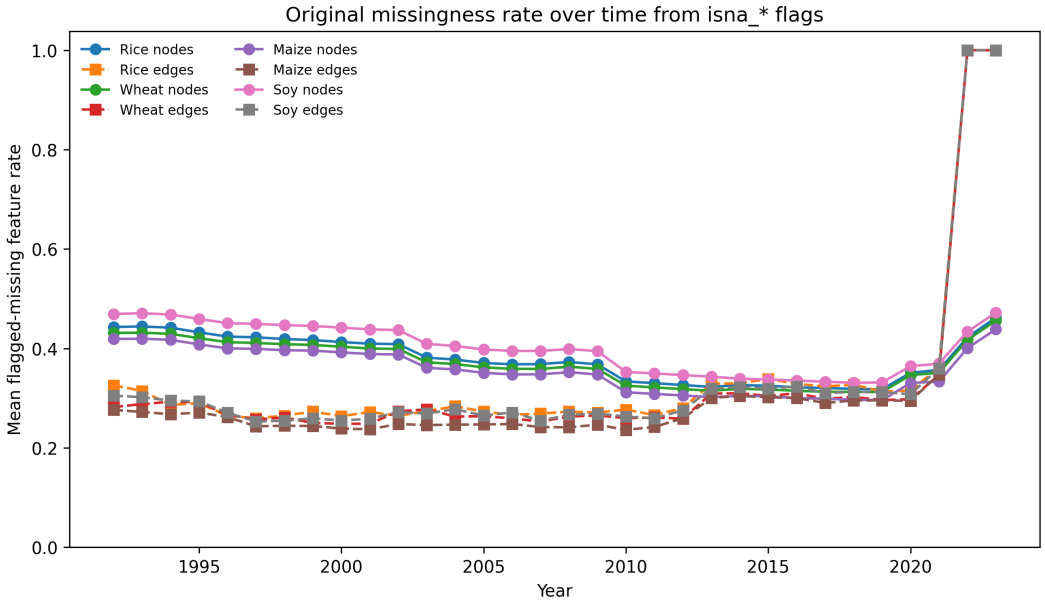



Figure 6: Missing data rate.

**Centrality.** Figure 7 shows how much of the total trade volume is concentrated amongst the countries top-ranked by trade volume, as listed in Table 11.

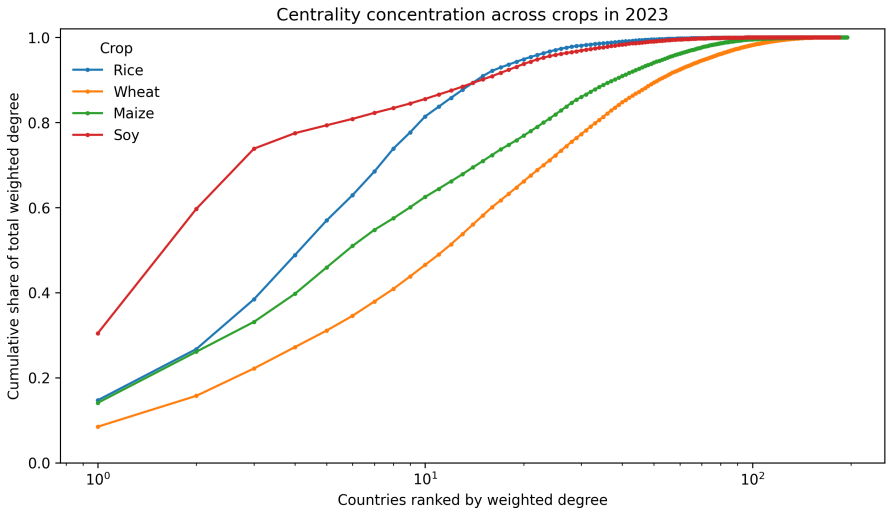



Figure 7: The centrality concentration across different crops in 2023: cumulative share of total weighted degree of countries against their rank weighted degree.

**Connectedness** The increase in density of the crop trade networks over the years (see Fig.8) suggests that the networks have become more interconnected. However, at the same time, the average

Table 11: The top five countries ranked by weighted degree (total imports + total exports) in 2023.

| Crop | Rank | Country | Weighted Degree | Unweighted Degree |
|---|---|---|---|---|
| Rice | 1 | United States of America | 4.90E+07 | 73 |
| Rice | 2 | Viet Nam | 4.01E+07 | 29 |
| Rice | 3 | Cambodia | 3.90E+07 | 9 |
| Rice | 4 | Brazil | 3.46E+07 | 44 |
| Rice | 5 | Mexico | 2.71E+07 | 5 |
| Wheat | 1 | Russian Federation | 1.36E+09 | 73 |
| Wheat | 2 | Australia | 1.17E+09 | 69 |
| Wheat | 3 | Canada | 1.03E+09 | 112 |
| Wheat | 4 | United States of America | 8.06E+08 | 123 |
| Wheat | 5 | Ukraine | 6.24E+08 | 90 |
| Maize | 1 | Brazil | 2.35E+09 | 130 |
| Maize | 2 | United States of America | 2.00E+09 | 184 |
| Maize | 3 | China, mainland | 1.17E+09 | 85 |
| Maize | 4 | Ukraine | 1.10E+09 | 88 |
| Maize | 5 | Argentina | 1.04E+09 | 127 |
| Soy | 1 | China, mainland | 4.21E+09 | 111 |
| Soy | 2 | Brazil | 4.04E+09 | 71 |
| Soy | 3 | United States of America | 1.96E+09 | 134 |
| Soy | 4 | Argentina | 5.02E+08 | 31 |
| Soy | 5 | Paraguay | 2.57E+08 | 19 |

path length over the years (see Fig.9) remains relatively stable. This could possibly indicate that an increase in density is driven by large exporters forming more edges with more and smaller importers, thereby increasingly participate in trade (see Fig.10), and there is a tendency to form a stronger core-periphery structure, i.e. the large exporters (hubs) get more links and the other countries connect to those. This could, in turn, increase dependency on the hubs.

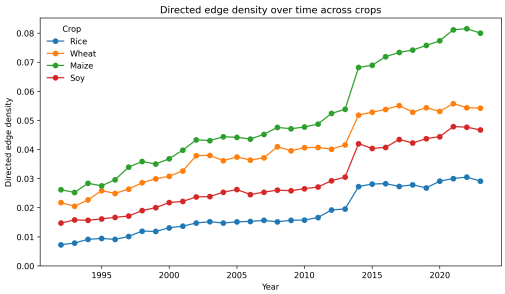



Figure 8: Directed edge density.

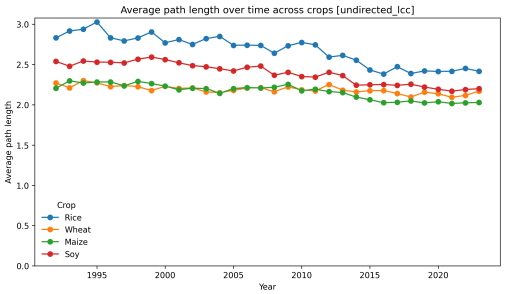



Figure 9: Average path length.

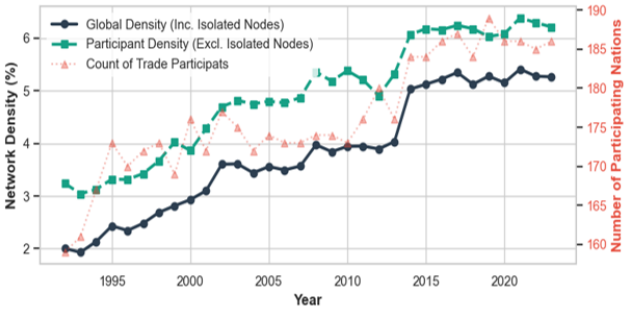



Figure 10: Global participation in wheat trade.

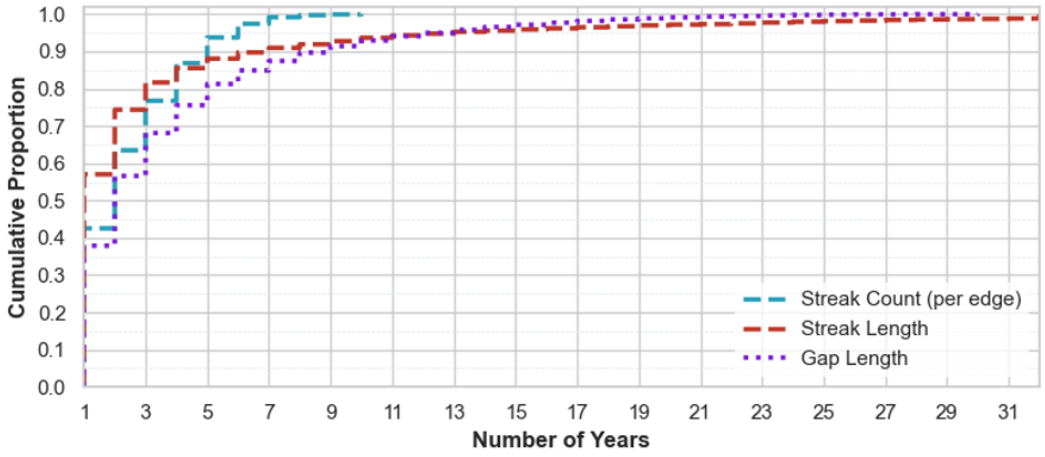



Figure 11: Distributions of the wheat trade links with respect to average streak length, (presence in consecutive years), average gap length between consecutive streaks, and the number of streaks (which shares the same x-axis but doesn't represent number of years.)

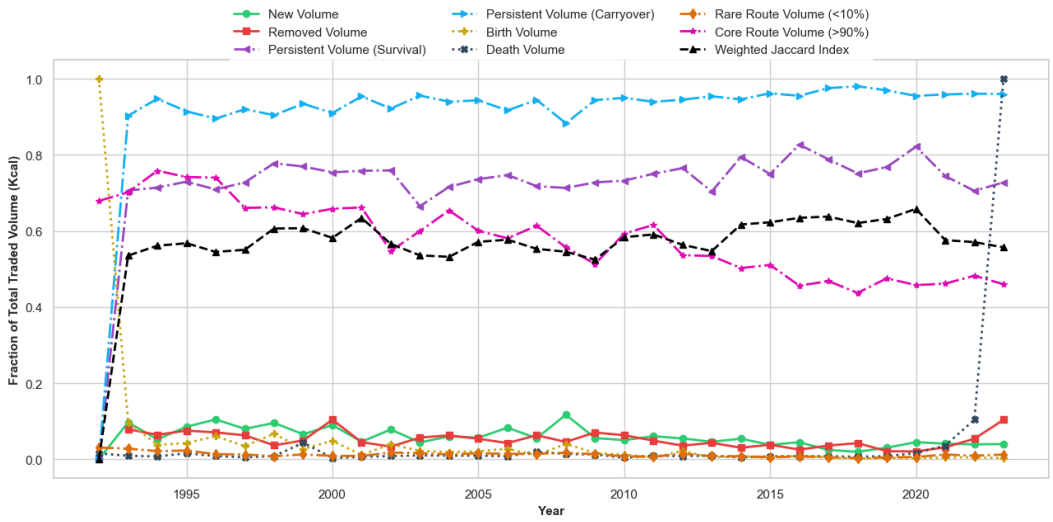



Figure 12: Illustration of the annual fraction of trade volume carried by different categories of edges such as edges that are 'born' (observed for the first time), 'die' (not observed past that year), core (appear >90% of the timeline), rare (appear < 10% of the timeline), year-on-year presence changes ('new' edges added or removed in comparison to previous year). Trade volume across edges persistent in consecutive years is reported from two perspectives: as a fraction of the previous year's edge volume (survival), and as a fraction of the latter year's edge volume (carryover). The weighted Jaccard index measures how similar the weighted trade-network structure is between two consecutive years, accounting for both the presence of trade links and their trade volumes. Note that since the range of Jaccard Index is also [0,1], it shares the same axis but doesn't represent trade volume fraction. Also note that the extreme values of 0 and 1 at years 1992 and 2023 are only due to these years being the endpoints of the data timeline.

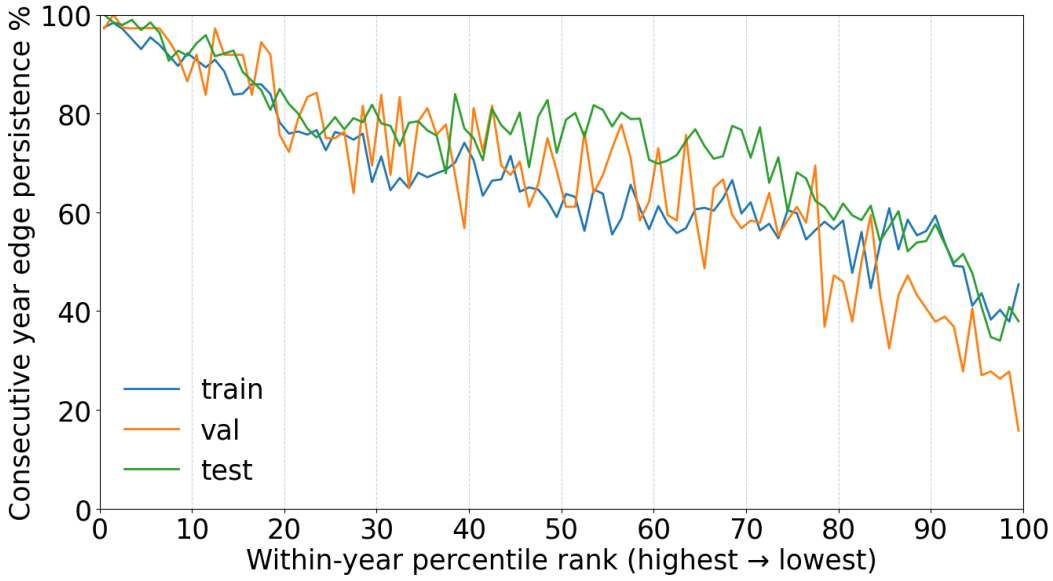



Figure 13: The percentage of edges in wheat trade that are persistent in consecutive years in each 1-percentage-point percentile bin. Edge volumes are ranked separately within each year, with lower percentile ranks corresponding to higher-volume edges, and the percentage is computed for edges in each split to observe the distribution difference in train and test sets.

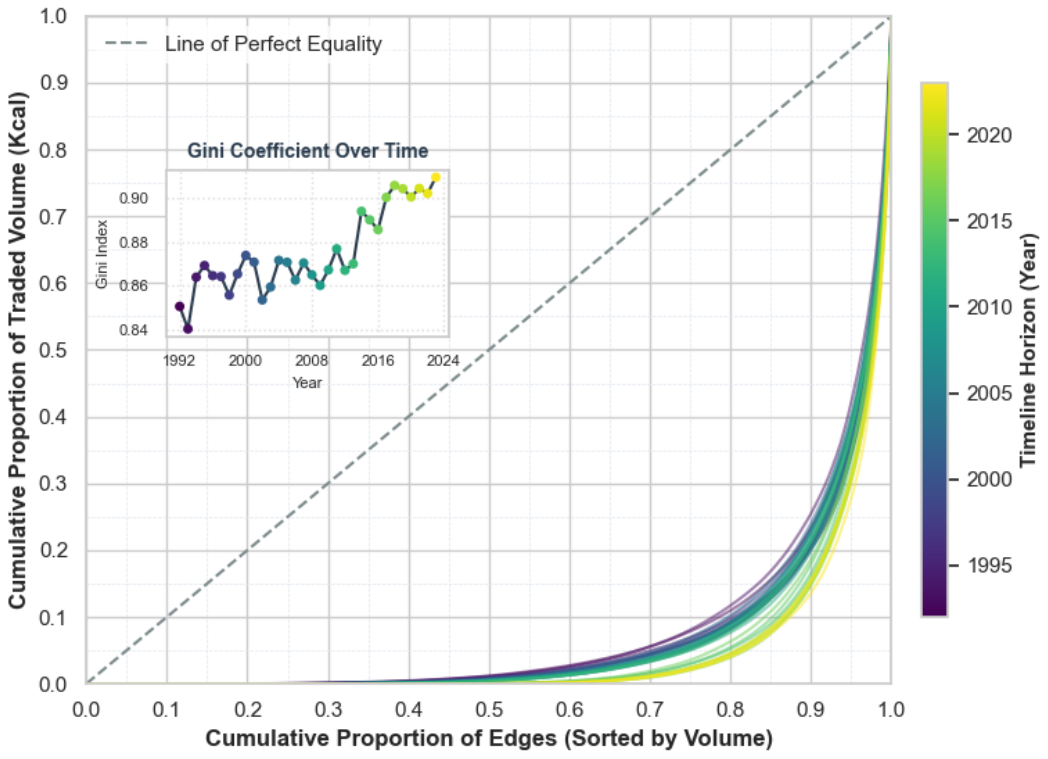



Figure 14: Wheat trade has become more concentrated over the years with the fraction of global wheat trade volume carried by the top 10% of the largest trade links increasing approximately from 75% in 1992 to 85% in 2023.

### C.1 LINK ANALYSIS

We study the analysis of links in the food trade network through the following lenses, accompanied by Figures 11 , 12, 13, and 14:

**Presence and Capacity** Of all observed edges, 80% have an average streak length of 3 years or less, and only 0.0284% appear in all 32 years. Although the capacity of the persistent trade links (appearing > 90% of the timeline) shows a declining trend, where the percentage of global trade such edges constitute reduces from 75% in 1994 to 46% in 2023, they continue to be the core of the network.

**Concentration and Volatility** : We measure the Hellinger distance between pairs of years for country-wise export and import share distributions to quantify how much trade patterns change between them, in terms of the trading partners involved and/or their respective shares. Based on the mean Hellinger distance between consecutive years, the top 5 most volatile importers and exporters of wheat are:

- Importers: Argentina, Niger, Democratic People's Republic of Korea, Somalia, India
- Exporters: Egypt, Thailand, Republic of Korea, Norway, China, Hong Kong

We also measure the Herfindahl-Hirschman Index (HHI) per country and year to quantify the concentration of trade among partners. Low HHI values indicate diversified trade, whereas high HHI values indicate dependence on a few partners, which can potentially imply resilience or vulnerability, respectively. Among wheat importers active for all 32 years, and based on the mean annual HHI:

- Most volatile: Saint Vincent and the Grenadines (in the Caribbean), with a mean annual HHI of 0.987, and a historical maximum of 1.0.
- Most stable: Italy, with a mean annual HHI of 0.166 and a historical maximum of 0.407.

Such measures can guide feature engineering and selection, which remain major machine learning challenges. Moreover, such analyses highlight the inherent heterogeneity in the data and the dataset's potential to serve as a benchmark for unsupervised learning tasks, such as clustering countries and/or trade relationships to identify archetypes and network roles.

### C.2 SHOCK ANALYSIS

Table 12: High-risk shocks identified using rolling five-year percentage changes

| Crop | Variable | Rank using rolling (5-year window) percentage change (negative) $z$-score | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Wheat | Supply/cap | Benin (2017) | Bahamas (2007) | Antigua and Barbuda (2021) | Maldives (2010) | South Sudan (2020) |
| | Imports | United Republic of Tanzania (2008) | Rwanda (2023) | Haiti (2010) | Libya (2022) | Sri Lanka (2022) |
| | Exports | Slovakia (2017) | Oman (1998) | Denmark (2004) | Ukraine (2022) | Argentina (2023) |
| Rice | Supply/cap | Cabo Verde (2008) | Botswana (2013) | Papua New Guinea (2010) | Lebanon (2019) | Tunisia (2020) |
| | Imports | Sierra Leone (2014) | Mexico (2012) | Ireland (2017) | Germany (2008) | Canada (2016) |
| | Exports | United Kingdom of Great Britain and Northern Ireland (2010) | United States of America (2022) | Pakistan (2019) | France (2013) | Netherlands (Kingdom of the) (2017) |

Table 13: Selected global grain and food-market shocks

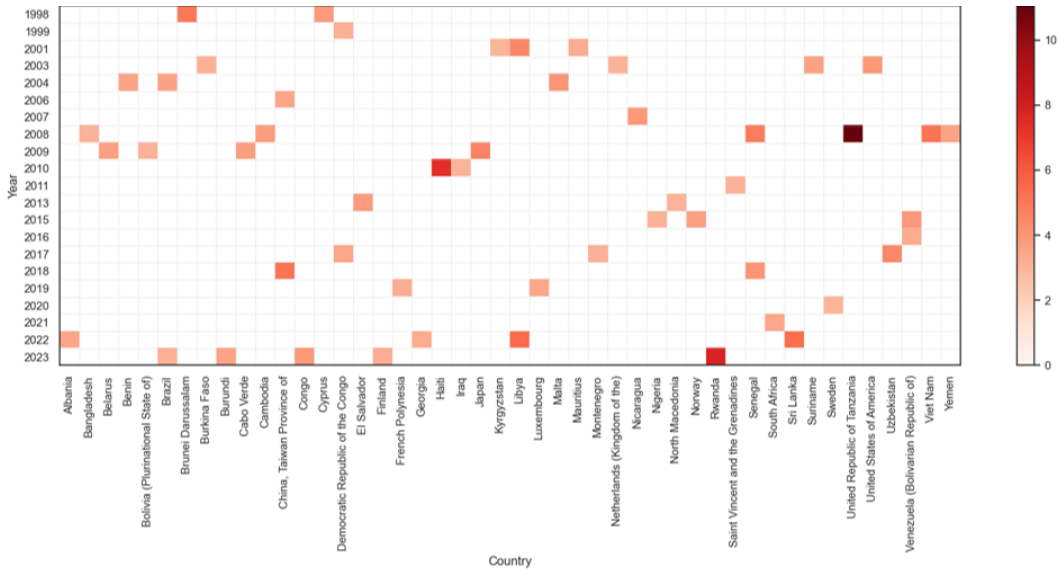
| Year(s) | Event |
|---|---|
| 1995–1996 | Global grain price spike |
| 2007–2008 | Global food-price crisis |
| 2010–2012 | Russian drought and wheat export ban<br>India–Thailand rice export reallocation |
| 2012 | Weather-driven supply shock |
| 2020 | COVID-related export restrictions |
| 2022 | Russia–Ukraine war shock |

We now provide shock detection analysis. We measure percentage change, i.e. $\frac{x_t - x_{t-1}}{x_t}$ in supply/cap, total imports, and exporters for all countries. Instead of using raw year-on-year values, we opt for a normalized z-score using a rolling window over the past 5 years. This ensures that each country's abnormality is assessed against its recent historical baseline. The complete history is not used to account for the gradual temporal shifts over the years.

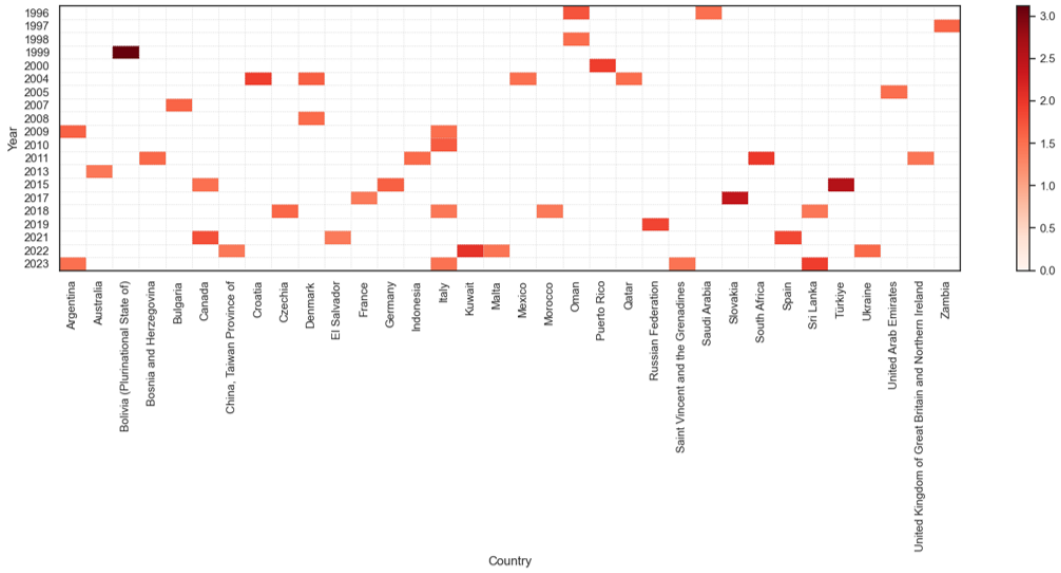
In Table 12, we report the top 5 shocks as (Country, Year) pairs for multiple variables and crops. A caveat in this analysis is that, because of annual data, shocks may appear a year later than the triggering global event due to calendar-year versus marketing-year reporting. We observe that these top shocks often align with well-known global periods of disruption in the international trade network, particularly of wheat, as identified in Table 13. We also show top trade shifts in particular historically relevant year transitions in Figure 17.

We note that the most impacted countries are import-dependent Caribbean countries and least developed or developing nations, particularly in Africa, where food insecurity levels are critical. This is also corroborated by Fig 15a, which shows the countries in each year that face a significant reduction in wheat imports, accompanied, in 15b, by countries that show a substantial reduction in wheat exports.

The shocks discussed so far essentially negative (an unusual decrease), but their relationship with 'positive' shocks (an unusual increase) is interesting to analyze. For this, instead of percentage change, we use log growth, i.e. $\log \frac{x_t}{x_{t-1}}$, for the 5 year rolling window z-score. This is so because,




(a) Imports




(b) Exports

Figure 15: Shock intensity, computed as a rolling percentage change z-score, is indicated by the colorbar. Only changes with a z-score > |3| are shown as shocks.

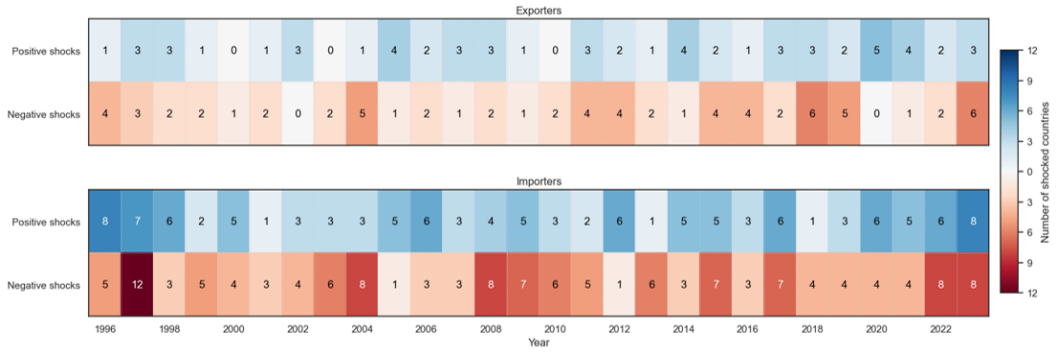



Figure 16: Number of countries experiencing positive and negative wheat trade shocks where positive and negative shocks refer to an unusual increase or decrease, respectively, in exports or imports.

unlike percentage change, log growth is symmetric for postive and negative changes, making them comparable. Some insights from Fig.16 are as follows:

- Negative exporter shocks are usually followed by positive exporter shocks as compensation.
- Negative exporter shocks accompanied by negative importer shocks indicate that the countries were highly dependent on the shocked exporters and unable to secure imports from other countries.
- Negative exporter shocks accompanied by positive importer shocks could indicate either exporter countries meeting their demand through imports, or dependent importers rerouting and stockpiling.
- The following positive import shocks could indicate recovery for both the shocked exporters and importers. Again, shocks impacting the most number of countries can be mapped to global events in corresponding years.

Finally, we emphasize that this general shock analysis is intended as a starting point. The dataset could support a much broader range of questions, including how countries respond to shocks, which exporters tend to fill gaps left by disrupted suppliers, and how substitution occurs across trading partners or crops. Methodologically, more robust detection methods can be developed that better account for volatility and trends. However, such analyses are not straightforward to generalize, as the appropriate approach depends on the research question, the intended application, and the desired level of interpretability. These extensions are therefore beyond the scope of this paper, which focuses on demonstrating the dataset's potential and outlining future research directions it enables, particularly in the context of food security.

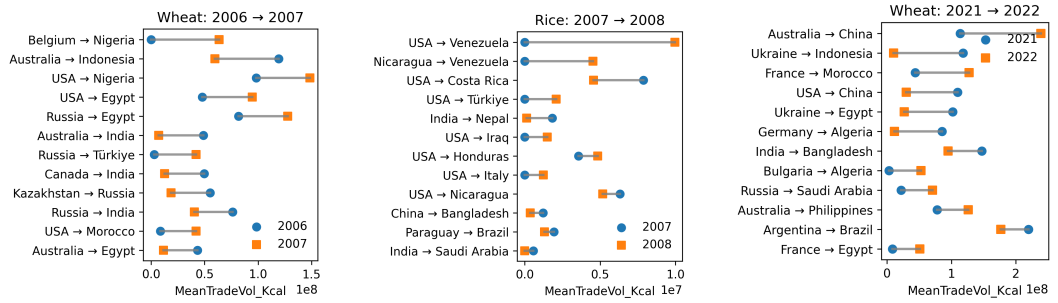
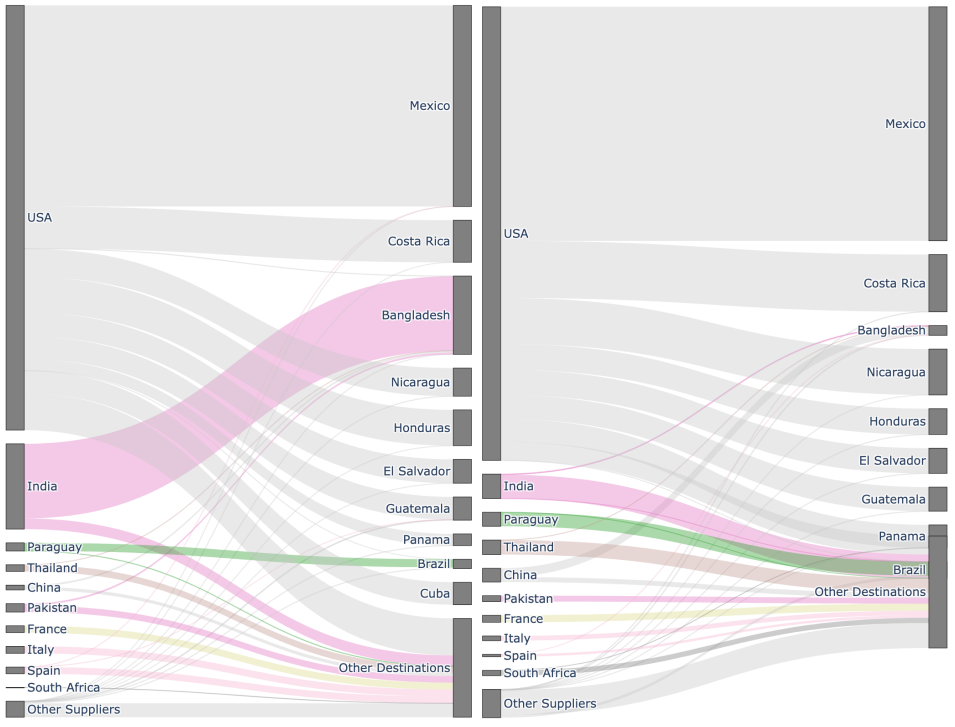
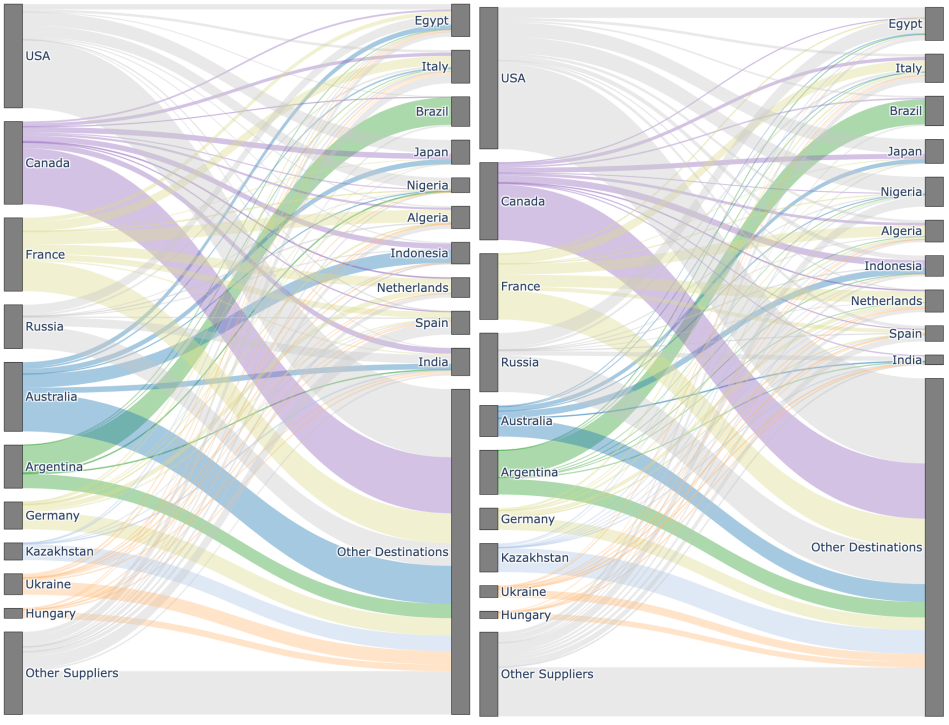



Figure 17: Top trade shifts in response to network shocks: In 2006–07, a supply-side shock in wheat trade was driven primarily by drought. In 2007–08, the rice turmoil was largely policy-driven due to export restrictions. In 2021–22, the Russia-Ukraine conflict resulted in wheat trade disruptions.




(a) Rice trade (2006)

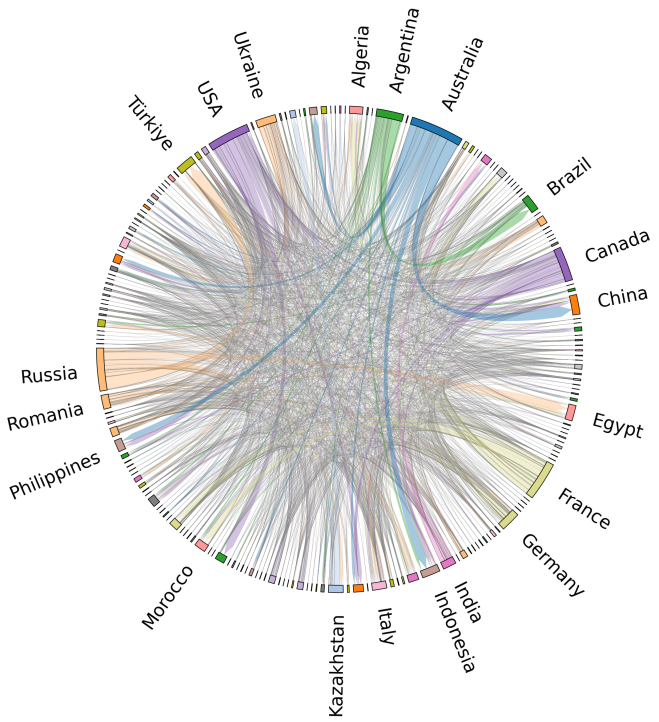
(b) Rice trade (2007)



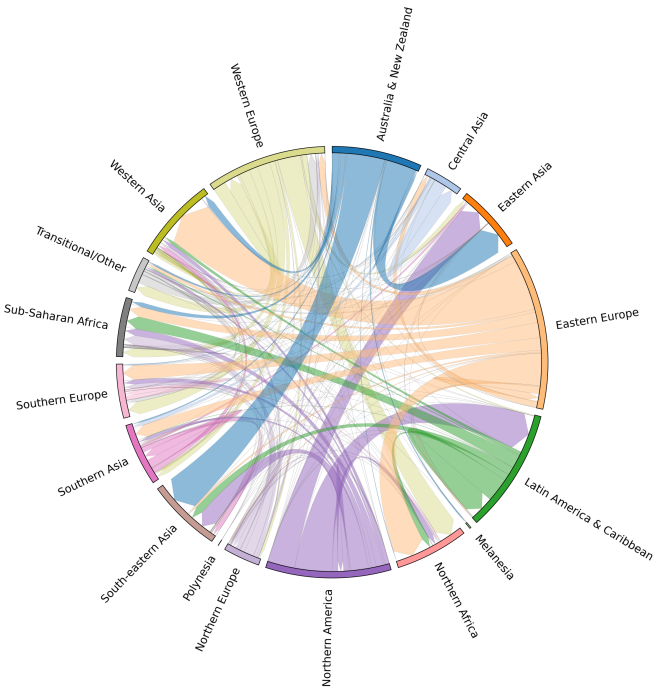

(c) Wheat trade (2006)

(d) Wheat trade (2007)

Figure 18: Sankey diagram to visualize major trade changes in consecutive years. In each diagram, the left and right columns represent exporters and importers, respectively. Importers. Top: Rice Shock 2006-07 where India's exports reduced by 62.1%. Bottom: Wheat Shock 2006-07 where the gap due to a drought shock in Australia was primarily filled by Russia.




(a) Country-level




(b) Region-level

Figure 19: Global wheat trade (2022).