Annals of Operations Research
https://doi.org/10.1007/s10479-026-07363-y

ORIGINAL RESEARCH

Forecast reconciliation of agricultural GHG emissions in Europe with fuzzy clustering

Raffaele Mattera[1] · Paolo Maranzano[2,3] · Caterina Morelli[2,5] · Germana Scepi[4]

Received: 28 April 2025 / Accepted: 16 July 2026

Abstract
We propose a novel statistical procedure which combines forecasting reconciliation techniques and unsupervised clustering algorithms to forecast annual greenhouse gases emissions at the territorial level. Specifically, we aim at predicting future emissions from the agricultural sector for Europe by aggregating bottom-level regional forecasts to the national and continental levels. Fuzzy clustering is used to define aggregations of bottom and middle-level series into clusters based on the similarity of emissions patterns; then, the fuzzy hierarchies are combined with the original administrative aggregation structure to improve the prediction accuracy of forecasting models. Empirical results show that fuzzy clustering aggregation, by leveraging uncertainty compared to crisp approaches, enhances forecast reconciliation accuracy at all levels of the hierarchy. Our approach offers a scalable solution to enhance forecast accuracy and reliability, particularly for short time series characterized by complex hierarchical structures.

Keywords Greenhouse Gases (GHG) emissions · Hierarchical forecasting · Agricultural emissions · Regional data (NUTS) for Europe · Unsupervised learning · Fuzzy k-means

1 Introduction

In recent decades, environmental degradation and climate change have emerged as pressing global concerns, primarily driven by anthropogenic emissions of greenhouse gases (GHGs) and air pollutants. GHGs in the European continent are generated by many activities, each of them contributing to climate change in specific ways (Crippa et al., 2024).

✉ Paolo Maranzano
paolo.maranzano@unimib.it

1 Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", Viale Abramo Lincoln, 5, 81100 Caserta, Italy

2 Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

3 Fondazione Eni Enrico Mattei (FEEM), Corso Magenta, 63, 20123 Milan, Italy

4 Department of Economics and Statistics, University of Naples "Federico II", Via Vicinale Cupa Cintia, 21, 80126 Naples, Italy

5 Department of Economic and Political Sciences, Università dellaValle d'Aosta - Université de la, Vallée d'Aoste Strada Cappuccini n.2A, Aosta 11100, Italy

Published online: 07 August 2026

Carbon dioxide (CO_2) is recognized worldwide to be the most relevant greenhouse due to its abundance in the atmosphere. It is predominantly released through the process of combustion of fossil fuels activated to produce energy, for transportation, and for large part of industrial activities, as well as deforestation and cement production. F-gases, which is synthetic in nature, are used in refrigeration, air conditioning, and other industrial processes like aluminum production.

Agriculture, often underrepresented in emissions discourse compared to sectors like energy and transportation, plays a pivotal yet complex role in this landscape. Not only it is a significant source of GHGs, but it is also a primary source of pollutants that directly affect both ecosystem and human health. Methane (CH_4) emissions primarily arise from agriculture, especially livestock farming and waste management, as well as fossil fuel extraction and processing. Nitrous oxide (N_2O) is largely emitted through agricultural practices, such as the use of nitrogen-based fertilizers, with smaller contributions from industrial processes and fossil fuel combustion. According to the European Environmental Agency (Agency, 2022), agriculture was the primary source of ammonia (NH_3) and methane (CH_4) in 2020, accounting for 94% and 56% of total European emissions respectively. It is worth noting that NH_3 emissions experienced only an 8% reduction in the last 20 years (the lowest decrease among all major pollutants), thus reflecting on the one side the sector's persistence in emissions and on the other the challenges of implementing effective mitigation strategies.

Agricultural emissions have deep implications beyond climate change. Exposure to particulate matter, especially fine particles ($PM_{2.5}$), has been consistently linked to adverse health outcomes, including cardiovascular and respiratory diseases. In Europe alone, $PM_{2.5}$ exposure is responsible for approximately 380,000 premature deaths annually, highlighting the urgency of addressing this issue (Thunis et al., 2021). A significant proportion of this pollution originates from secondary particles formed through atmospheric reactions involving ammonia, which is predominantly emitted by livestock farming (Fassò et al., 2023; Lovarelli et al., 2020). The health risks are exacerbated in regions with intensive agricultural practices, where empirical studies have shown a strong association between proximity to farms and respiratory conditions such as pneumonia and reduced lung function (Borlee et al., 2017; Kalkowska et al., 2018). Furthermore, increased livestock density directly correlates with elevated $PM_{2.5}$ and NH_3 levels, particularly in industrialized farming regions like the Po Valley in Northern Italy, where pollution burdens are among the highest in Europe (Lunghi et al., 2024; Otto et al., 2024). These findings emphasize the need for targeted mitigation strategies and robust policy interventions.

Addressing emissions from agriculture also entails navigating complex trade-offs. The shift towards sustainable practices–such as replacing chemical fertilizers with bio-based alternatives–can offer significant environmental benefits. However, these transitions must balance productivity, economic viability, and social equity (Kannan et al., 2025; Pimentel, 1996). Policymakers are increasingly turning to systems modeling to evaluate these trade-offs and develop targeted interventions. In addition to mitigation strategies, the accurate quantification and forecasting of emissions are essential for effective environmental governance. Yet, as highlighted by Porter and Reay (2016), data inconsistencies and the reliance on default emission factors remain major obstacles, especially in the context of food production and wastage.

Empirical forecasting of greenhouse gas (GHG) emissions in agriculture encompasses a diverse array of methodologies, reflecting the sector's complex dynamics and regional variations. Therefore, the identification of GHG emissions determinants (or at least the association with further elements) of agricultural emissions (see, for instance, Fassò et al., 2023), as well as their short-term and long-term forecasting are now pivotal in the scientific panorama.

Traditional approaches, such as GHG calculators, rely on emission factors and empirical models to estimate emissions based on minimal input data. However, these calculators often face limitations in accounting for the variability in climatic conditions, management practices, and crop rotations, leading to significant prediction uncertainties (Richards et al. 2016; Peter et al. 2017). Manure management, a critical factor in livestock emissions, further exemplifies the need for dynamic models that incorporate daily temporal variations and site-specific parameters to accurately estimate emissions from storage and field applications (Petersen (2018)). Recent advances also include integrating comprehensive tracking systems, to map the environmental impact of agricultural supply chains and develop actionable plans for carbon neutrality (El Hathat et al. 2023). While these methodologies provide valuable insights, their precision and scope often hinge on the availability of localized, high-quality data and the integration of advanced technologies to overcome inherent modeling uncertainties. Recent studies underscored the potential of statistical machine learning (ML) and deep learning (DL) techniques in forecasting GHGs emissions, especially agricultural emissions, with high spatial and temporal variability. For instance, Long short-term memory network (LSTM) models have been shown to outperform traditional regression and shallow learning models in accurately predicting CO_2 and N_2O fluxes from agricultural soils, with significantly lower prediction errors (Hamrani et al., 2020). Similarly, hybrid approaches like feed-forward neural networks (FFNN) and adaptive boosting have demonstrated high precision in forecasting emissions linked to specific agricultural practices (e.g. see Dar et al., 2024). Beyond GHGs, ML techniques have been applied to predict nitrogen pollution patterns (Fernández-Avilés et al., 2024), identifying key drivers like land use and rainfall and achieving substantial accuracy with models like Random Forest (Wang et al., 2024). Moreover, ML-based regional frameworks provide robust predictions of emission efficiency trends and actionable insights for targeted low-carbon strategies (Jiang et al., 2024). These methodologies not only enhance the understanding of emissions dynamics but also inform sustainable agricultural and energy practices, addressing challenges like energy consumption in dairy production through data-driven optimization of resources and infrastructure (Shine et al., 2020). Despite their usefulness, the adoption of these methods can be limited while analyzing high spatial resolution data, as those at the NUTS level, because the time series are of short length, and this limits the achievement of an appropriate learning for ML and DL. In this settings, indeed, low parametrized models have to be preferred (Hyndman & Athanasopoulos, 2018).

In this regard, we believe it is critical to produce reliable forecasts about the dynamics of emissions in the agricultural sector, as these enable a better anticipation of pollution trends, supporting the design of equitable mitigation strategies, and ultimately safeguarding both public health and environmental sustainability. However, to the best of our knowledge, there has been little attention paid to forecasting GHGs emissions from the agricultural sector and in particular to those associated with multiple spatial scales. Indeed, existing studies have largely focused either on individual countries or broader continental trends across Europe (e.g. Antanasijević et al., 2015; Moiceanu and Dinca, 2021). A key challenge lies in the availability and granularity of data; indeed, while relatively long time series exist at the national level, the same is not true for sub-national units (e.g., regions or municipalities), where GHG emissions data are typically available only on an annual basis from 1990 onward. This results in short time series that constrain the applicability of many advanced forecasting techniques. In addition to the previous critical issues, even the atmospheric behavior of emissions further complicates forecasting. As Nenes et al. (2020) demonstrate, the interplay between ammonia and other precursors, such as nitric acid, can modulate particulate matters concentrations in non-linear and geographically variable ways, reinforcing the necessity of regionally tailored emission models.

To introduce the novelty of our paper, we start by discussing a set of observational and empirical evidence related to the GHGs emission from the agricultural sector. First, we might observe that agricultural emissions exhibit an inherently hierarchical structure, reflecting the way data is collected and organized from local production areas (e.g., regions or provinces), to national aggregates, and ultimately to continental-level summaries. This spatial and administrative hierarchy is crucial not only for reporting purposes but also for understanding and predicting emission dynamics across scales. In this context, forecast reconciliation techniques (see Hyndman et al., 2011; Wickramasuriya et al., 2019; Panagiotelis et al., 2021) offer a powerful framework to ensure coherence between forecasts produced at different levels of aggregation. By leveraging information from fine-grained, bottom-level forecasts, reconciliation improves the accuracy and consistency of higher-level predictions, while respecting the hierarchical relationships among them.

Moreover, we notice the evidence that emission patterns in agriculture are often clustered. For instance, Morelli et al. (2026) proposes a spatiotemporal clustering approach to analyze GHG emissions across European regions, considering different sectors. Their findings reveal significant heterogeneity in emission trends, especially from agriculture, highlighting persistent hot-spots in densely farmed regions such as Northern France, Belgium, and the Netherlands. The territorial pattern in the European agricultural market should not surprise, as it is strictly connected with the noticeable variability in farm and company structures across and within regions and countries (Guarín et al., 2020), which can be explained by a combination of historical, cultural, geographical, economic, and political factors (Zimmermann et al., 2009). Also, the intersection of agricultural market concentration and environmental outcomes is gaining attention. Boccaletti et al. (2010) emphasizes the dual effects of structural transformation in European agriculture, that is, while consolidation may improve efficiency, it can also exacerbate environmental degradation, especially when linked to ammonia and methane emissions. Similarly, Cerqueti et al. (2025) identify spatial heterogeneity in agricultural production concentration across Europe, which may have important implications for pollution management and policy design. For the purposes of this study, the high degree of heterogeneity cannot be ignored as it underscores the importance of localized forecasting models that account for regional differences in agricultural practices and land use.

Motivated by the above evidence, we introduce a novel approach that addresses several of the limitations stated above by combining consolidated forecasting methods with machine learning algorithms which exploit relevant properties of spatio-temporal data, such as grouped and hierarchical structures, within a statistical machine learning framework. Specifically, we leverage on hierarchical time series forecasting methods (Hyndman et al., 2011) and we integrate unsupervised learning into forecast reconciliation to improve the predictive accuracy of agricultural GHG emissions across Europe, at both the country and regional levels. Our framework also incorporates clustering strategies to exploit structural similarities across regions, further enhancing the robustness and coherence of forecasts within a unified, multi-scale modeling setup. Recent literature on hierarchical forecasting (see, for instance, Pang et al., 2022; Mattera et al., 2024; Zhang et al., 2024) highlighted that the process of aggregation (e.g., from regions to countries and from countries to the European level) can be a relevant framework in which unsupervised learning, and in particular clustering-based reconciliation methods, might be used to build enhanced hierarchies capable of capturing the relevant underlying spatial patterns. However, in these contributions clustering is employed to generate crisp partitions of the bottom-level series, and hierarchical uncertainty is addressed through the construction, selection, or averaging of alternative discrete hierarchies.[1]

Differently, we explicitly model uncertainty in hierarchy construction within the reconciliation framework itself by introducing fuzzy clustering. This allows each bottom-level series to contribute simultaneously to multiple latent aggregates through continuous membership weights, resulting in fuzzy linear aggregation constraints rather than binary ones. To the best of our knowledge, this is the first forecast reconciliation approach in which fuzziness is an intrinsic structural component of the hierarchy, rather than a device for generating alternative crisp hierarchies. Embedding fuzzy clustering in the hierarchical forecasting problem provides a more nuanced representation of regional similarities, which is particularly relevant in the agricultural domain, where borders are porous, practices spill over across regions, and administrative divisions do not always align with emission drivers. By explicitly acknowledging and modeling this structural uncertainty, our approach enhances forecast accuracy and adaptability. This constitutes the core contribution of our work: integrating fuzzy clustering into the hierarchical reconciliation framework to reflect the complex, overlapping spatial structures that characterize agricultural emissions in Europe.

The rest of the paper is structured as follows. Section 2 introduces to the real data on GHG emissions for agricultural sector in Europe, discussing similarity patterns across countries in terms of temporal evolution of emissions. Section 3 discusses the optimal reconciliation and the proposed methodology based on fuzzy clustering, while Sect. 5 presents the empirical strategy, the obtained results and some robustness checks. Section 6 concludes with final remarks and future research directions.

2 Regional data on GHGs emissions from the European agricultural sector

The definition of Europe adopted in this paper refers to the European continent, that is, it includes the countries member of the European Union (EU), a well as non-member countries, such as Iceland, Norway, Switzerland, United Kingdom, and the Balkan countries. From the administrative standpoint, European countries are classified according to the Nomenclature of Territorial Units for Statistic (NUTS) (European Commission, 2016). The NUTS classification uses a four-strata hierarchical organization in which municipalities, also called local area units (LAUs) uniquely belong to provinces (NUTS-3), which in turn belong to regions (NUTS-2), macro-regions (NUTS-1), and countries (NUTS-0). The classification is

[1] While Pang et al. (2022) introduced the concept of clustering-based reconciliation, Mattera et al. (2024) proposed the ensemble of different hierarchies to reduce clustering uncertainty and increase performance. Zhang et al. (2024) compared hierarchical and partitional clustering approaches and investigate the motivation behind the over-performance of clustering-based reconciliation methods. In our work, we extend these approaches by leveraging fuzziness and therefore explicitly modeling uncertainty in the clustering structure.

updated periodically by the Eurostat and the areas borders can change according to national or supranational legislation.[2]

In our application, we considered the largest possible spatio-temporal dataset at the regional level, seeking to maintain a balance with respect to the environmental relevance of the agricultural sector. Thus, we preserved all the continental European regions, that is, all the NUTS-2 areas belonging to the continental Europe excluding extra-continental territories,[3] augmented by Iceland, where the agricultural sector accounts for an average of 15% of total GHG emissions. In fact, being interested in forecasting agricultural emissions in contiguous areas with a relevant importance for the agricultural sector, we aim at ensuring that the extremely high distances involving non-continental regions do not render the spatial dissimilarities among the remaining regions insignificant.

Regional data on atmospheric GHGs emissions at the sector level are provided by the European Commission through the Emissions Database for Global Atmospheric Research (EDGAR) database (Crippa et al., 2024), being part of the Annual Regional Database of the European Commission (ARDECO) (Bucciarelli, 2024) project.[4] Among the available information, we considered yearly measurements of total GHGs emissions from 1990 to 2023 (i.e., $T = 34$ time stamps) for the $N = 309$ regions in the EU-27 countries plus Albania, Bosnia-Herzogovina, Iceland, Norway, North Macedonia, Serbia-Montenegro, Switzerland, Turkey, and the United Kingdom (i.e., 36 countries involved). Data are measured in ktons of CO_2 equivalent (CO2eq) and they include fossil CO_2 only, CH_4, N_2O and F-gases. While EDGAR provides sector-specific and gas-specific information, we extract data specifically referring to the total emission of GHGs attributable to agricultural soils, agricultural waste burning, enteric fermentation, manure management, and indirect N_2O emissions from agriculture.

In Fig. 1 we show the country-level trends for GHGs emissions from agriculture in Europe since 1990. The time series confirm that the agricultural sector, despite the recent improvements in technology and farming-livestock techniques, in the last three decades experienced a slow and very modest decline in its environmental impact. With the exception of few countries, like Germany (orange line), United Kingdom (green line), and France (red line), agricultural emissions from the European countries are stable. On the contrary, Turkey (light blue line) is the only country showing a clear increasing trend with an increase of about 50% since 2010. Even without using clustering techniques as done by Morelli et al. (2026), a grouped-data structure is apparent and highlights the presence of at least two groups, that is countries with stable patterns but poor emissions in absolute terms (bottom lines) and more dynamic countries having a pivotal role in determining the overall emissions for Europe (top lines).

[2] We adopt the Eurostat NUTS-2021 classification https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/territorial-units-statistics as the consistent regional framework for the entire period under analysis. To ensure comparability over time, we apply the NUTS-2021 boundaries retrospectively to all years and disregard subsequent territorial revisions or boundary changes. A potential drawback is that, while the United Kingdom would be excluded under the more recent NUTS-2024 edition, the 2021 version still includes it, whereas some territorial units in Portugal (i.e., PT19, PT1A, PT1B, PT1C and PT1D are introduced in NUTS-2024) and the Netherlands (i.e., NL35 and NL36 are introduced in NUTS-2024) are lost due to administrative reforms. As discussed in the main text, we prioritize the NUTS-2021 edition to preserve the broadest possible spatio-temporal coverage.

[3] We dropped from the analysis the Azores (PT20) and Madeira (PT30) for Portugal, all the French territories in Central and South America (FRY), and Ceuta (ES63), Melilla (ES64), and the Canarias islands (ES70) for Spain. Notice that, for these territories, the absolute agricultural emissions are very low due to their dimensions and the sectoral contribution over total emissions never overcome 10%.

[4] Emissions data are publicly provided by JRC-EDGAR at the following link https://edgar.jrc.ec.europa.eu/dataset_ghg2024_nuts2 (Accessed on February 26th, 2026.). Data are harmonized at the regional NUTS-2 level and do not present missing values for the considered areas.

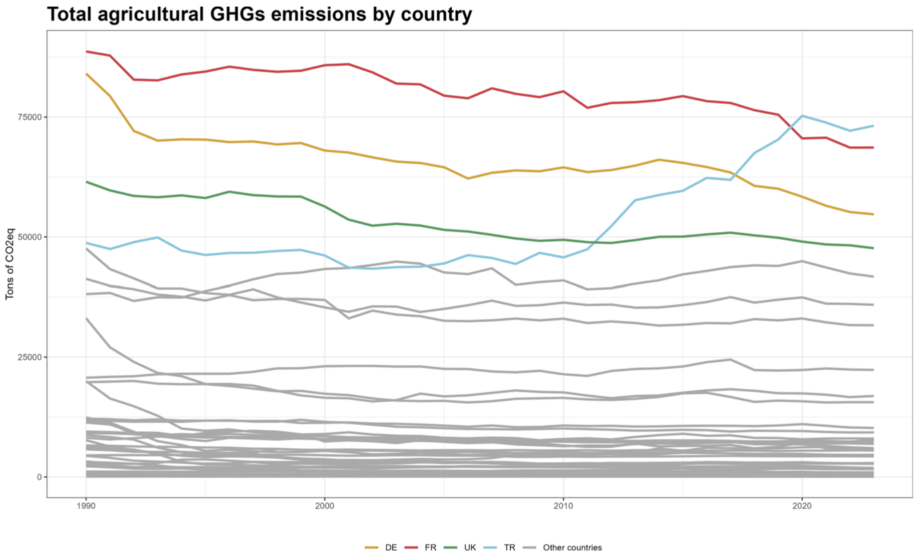


Fig. 1 Yearly total GHGs emissions from the agricultural sector for European countries (NUTS-0) from 1990 to 2023. Germany, France, UK and Turkey are shown as interesting examples, while other countries are reported with grey color

When it comes to the regional level, the picture becomes more fragmented as new local features emerge. In Fig. 2 we represent the contribution of agriculture to the overall amount of GHGs emissions for the European regions from 1990 to 2023. Interestingly, the maps suggest several co-existing patterns in the agricultural emissions. First, for many areas the share of agricultural emissions seems to decrease (colors move from reddish to greenish), especially in Central Europe, Italy, United Kingdom, and Turkey. This suggest a change in the emissions generating process, which sees the European agriculture less and less relevant compared to other sectors. Second, for many countries the intra-national distribution is heterogeneous due to the co-presence of areas with very high shares of agricultural emissions and areas having an almost null value. This empirical evidence supports the already-mentioned historical, cultural, geographical, economic, and political heterogeneity of European countries (Zimmermann et al., 2009) which affected the agricultural market structure, as well as the attitude toward its environmental sustainability.

Overall, the data reveal a relevant spatial heterogeneity at both national and local levels, partially moderated by the stability across time. Then, forecasting agricultural emissions for Europe can be challenging. On the one side, modeling and forecasting data at the national level only does not reflect the intra-national features and underestimate the actual complexity of the dataset. This aspect can be even more problematic when moving to the continental level. On the other side, as we are considering low-frequency (i.e., yearly) time series for a long-term analysis (i.e., thirty-four years), handling individual trajectories without considering aggregations or grouped structures can lead to a low quality forecasting with poorly realistic projections and lack of stability in the estimated values. Therefore, the idea of augmenting the natural hierarchy of time series with the identification of possible latent structures for forecasting purposes could lead to highly accurate future forecasts consistent with the actual economic-productive structure of Europe.

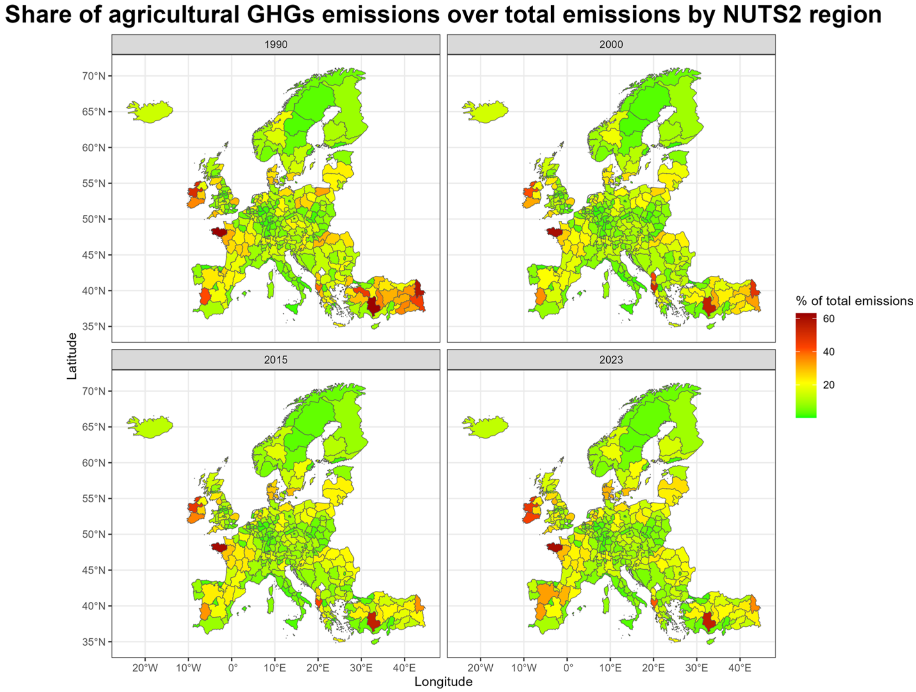


Fig. 2 Share of agricultural GHGs emissions over total emissions by NUTS-2 region for 1990, 2000, 2015, and 2023

3 Methodology

As discussed in Sect. 2, GHGs emissions at the European level can be disaggregated in terms of NUTS, meaning that the sum of the emissions in all the European NUTS exactly equals the total emissions for Europe. In general, time series adhering to such constraints are called linearly constrained (Girolimetto & Di Fonzo, 2024). Hierarchical time series are a specific type of linearly constrained time series, which adhere to a hierarchical structure.

More formally, let $\boldsymbol{b}_t \in \mathbb{R}^{n_b}$ be the vector of bottom-level time series, where $n_b = 309$ corresponds to the number of NUTS-2 regions included in the analysis. Let $\boldsymbol{a}_t \in \mathbb{R}^{n_a}$ denote the vector of aggregated series. An aggregation matrix $\boldsymbol{A} \in \mathbb{R}^{n_a \times n_b}$ encodes the linear constraints linking bottom-level NUTS-2 regions and aggregated series, that is,

$$\boldsymbol{a}_t = \boldsymbol{A}\boldsymbol{b}_t. \tag{1}$$

The full vector of time series is then $\boldsymbol{y}_t = \left(\boldsymbol{a}_t', \boldsymbol{b}_t'\right)' \in \mathbb{R}^n$ where $n = n_a + n_b$, and the corresponding summation matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n_b}$ stacks $\boldsymbol{A}$ and the identity matrix I_{n_b} as follows

$$\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{A} \\ I_{n_b} \end{bmatrix} \boldsymbol{b}_t = \boldsymbol{S}\boldsymbol{b}_t. \tag{2}$$

Different aggregation structures can be considered within the general formulation in Eqs. (1)–(2).

As a simple benchmark, one may consider a single-level aggregation where all NUTS-2 emissions directly sum to the European total. In this case, $n_a = 1$ and the aggregation matrix reduces to a vector of ones $\boldsymbol{A} = \mathbf{1}'_{n_b}$. In a baseline administrative hierarchy, the n_b NUTS-2 regions are first aggregated at the country level and subsequently to the European total. Since the $n_b = 309$ regions belong to 36 countries, the aggregation matrix $\boldsymbol{A}$ has dimension $n_a \times n_b = 37 \times 309$, where the first row of $\boldsymbol{A}$ is a vector of ones corresponding to the European total aggregate, and the remaining 36 rows correspond to country-level aggregates. Specifically, the (i, j)-th entry of $\boldsymbol{A}$ equals 1 if NUTS-2 region j belongs to the aggregate i, and 0 otherwise.

3.1 Forecast reconciliation

To forecast European agricultural emissions, we have two main options, that is, forecasting the top-level series directly or forecasting all the n hierarchical time series and obtaining the forecasts for the top-level by aggregating such bottom-level forecasts. This last approach should be preferred over the first because forecasting the NUTS-level emissions provides important information for policy purposes, both at the country and European levels. In doing so, it is natural to expect the forecasts to add up in the same way as the data in-sample. However, it is well known that forecasts obtained by n independent forecasting models do not generally fulfill this coherency property. For this reason, a flourishing literature (e.g. see Hyndman et al., 2011; Wickramasuriya et al., 2019; Panagiotelis et al., 2021; Hollyman et al., 2021; Panagiotelis et al., 2023) emerged proposing methods to ex-post adjust the base forecasts with the aim of making these coherent with the hierarchical structure (2). This ex-post adjustment task is known as forecast reconciliation.

More formally, let us define $\widehat{\boldsymbol{y}}_h$ as the vector of h-step-ahead base forecasts obtained with a generic forecasting model, separately applied to each n time series in the hierarchy (2). While time series $\boldsymbol{y}_t$ naturally aggregate according to a hierarchical structure given by $\boldsymbol{A}$, the forecasts $\widehat{\boldsymbol{y}}_h$ do not necessarily. Optimal reconciliation techniques, provide an ex-post adjustment to the forecasts $\widehat{\boldsymbol{y}}_h$ and make them coherent with the aggregation structure. More precisely, we define coherent forecasts as

$$\widetilde{\boldsymbol{y}}_h = \boldsymbol{M}\widehat{\boldsymbol{y}}_h, \tag{3}$$

where $\boldsymbol{M} = \boldsymbol{S}\boldsymbol{G}_h$ is a $n \times n$ mapping matrix, whose role is to make the base forecasts $\widehat{\boldsymbol{y}}_h$ coherent with the aggregation (2). The summing matrix $\boldsymbol{S}$ aggregates the reconciled forecasts according to the relevant grouped structure. Therefore, different reconciliation approaches can be obtained for different mapping matrices $\boldsymbol{G}_h$ and grouped structures $\boldsymbol{S}$.

The simplest reconciliation approach is called bottom-up, which is obtained by summing up the base forecasts. In this case, the mapping matrix takes the form $\boldsymbol{G} = \begin{bmatrix}\mathbf{0}_{n_b \times n_a} & \boldsymbol{I}_{n_b}\end{bmatrix}$ and does not depend on the forecast horizon h. More recently, Wickramasuriya et al. (2019) proposed an alternative approach for determining the optimal mapping matrix, obtained by minimizing the sum of the variances of all the reconciled forecasts. This approach, called MinT, leads to the following closed-form solution

$$\boldsymbol{G}_{MinT} = \left(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}. \tag{4}$$

The MinT approach is more general and encompasses both (19) and (20) as special cases. Therefore, (4) is the reconciliation approach most commonly used nowadays as it usually performs better than both (19) and (20). In the presence of high-dimensional hierarchies, shrinkage estimators (e.g., Ledoit & Wolf, 2004) are commonly used to invert $\boldsymbol{W}_h$.

3.2 Reconciliation with fuzzy clustering

There are empirical settings where the aggregation matrix $\boldsymbol{A}$, and therefore also $\boldsymbol{S}$, are not known in practice. This is the case of death counts in mortality forecasting (Li et al., 2019), energy load (Brégère & Mattera, 2025; Pang et al., 2022) and stock prices (Mattera et al., 2024). These previous research has shown that forecast reconciliation is useful and can also be applied in these settings. Indeed, one can consider a clustering of n_b time series into $k \geq 2$ groups, which can be represented by a binary membership matrix $\boldsymbol{\mathcal{C}}$ of dimension $k \times n_b$. Therefore, we can write the hierarchical time series under cluster structure as follows:

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \widehat{\boldsymbol{\mathcal{C}}} \end{bmatrix}, \tag{5}$$

where $\widehat{\boldsymbol{\mathcal{C}}}$ denotes an estimate of the clustering of the n_b bottom time series. Any time series clustering approach can be used to build $\widehat{\boldsymbol{\mathcal{C}}}$ in practice. From an interpretative perspective, clustering-based aggregation layers are territorial units exhibiting similar temporal profiles of GHG emissions and, therefore, potentially reflecting common production structures, climatic conditions, or mitigation practices. As such, clustering-based hierarchies may convey policy-relevant information that complements standard administrative aggregation structures.

The (5) can be extended to combine multiple cluster structures as follows (Mattera et al., 2024)

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \widehat{\boldsymbol{C}}_1 \\ \vdots \\ \widehat{\boldsymbol{C}}_l, \\ \vdots \end{bmatrix}, \tag{6}$$

where $\widehat{\boldsymbol{C}}_l$ represents an l-th partition of the data into k_l clusters. Then, any optimal reconciliation approach can be used by simply plug-in (6) into (1). While Mattera et al. (2024) propose to enrich the hierarchy with partitions obtained using alternative approaches, we remark that one can also dispose of *known* partitions which can be combined with estimated ones. In other words, while some of the cluster structures in (6) are known, so that we denote it as $\boldsymbol{C}_l$, some others are estimated via clustering, and we call this $\widehat{\boldsymbol{C}}_{l'}$ with $l' \neq l$. For example, in the case of the GHGs emissions presented in Sect. 2, administrative hierarchy linking NUTS regions to countries and to the EU total are always available and pre-defined. Therefore, in this paper we consider hybrid constructions where known aggregation constraints are augmented with one or more data-driven clustering layers. The resulting aggregation matrices are made explicit in Sect. 5.1. In this regard, Zhang et al. (2024) shows that forecast accuracy that accrues from using clustering does not arise from grouping similar series but from the structure of the hierarchy. Therefore, it seems to be recommended to augment known hierarchies with additional levels obtained with clustering.

However, forcing each unit into a single cluster through binary membership may be overly restrictive, especially in environmental applications where emission dynamics often display hybrid or transitional patterns across regions. Under fuzzy clustering, the additional hierarchy nodes should be interpreted as *soft aggregates*, defined as weighted combinations of bottom-level series. Unlike crisp cluster aggregates, these nodes capture overlapping similarities across regions and explicitly account for uncertainty in group membership. In this sense, fuzzy hierarchy nodes do not replace administrative or policy units, but provide an uncertaintyaware aggregation layer that complements existing hierarchies and improves the effectiveness of forecast reconciliation.

In this spirit, there is no reason for $\boldsymbol{a}_t$ to be restricted to aggregates of $\boldsymbol{b}_t$. From a modeling perspective, this interpretation naturally leads to a more general class of aggregation constraints. In other words, for linearly aggregated time series we can consider any linear combination of the bottom-level series $\boldsymbol{b}_t$, so the corresponding $\boldsymbol{A}$ and $\boldsymbol{S}$ matrices may contain any real values, not just 0 and 1 (see also Hyndman, 2022).

Starting from these intuitions, we consider a simple approach to introduce non-binary weights into (1) and we do it by leveraging the usefulness of clustering in forecast reconciliation. In particular, following abundant research in time series clustering (for a complete treatment, see Maharaj et al., 2019), assuming that a time series certainly belongs to a single cluster may be limiting in many practical domains. This is where fuzzy clustering becomes advantageous. By allowing each time series to have partial memberships across multiple clusters, fuzzy clustering captures the nuanced and overlapping dynamics often present in real-world data. Such flexibility is particularly relevant in hierarchical forecasting, where bottom-level series frequently exhibit hybrid patterns and dependencies. Fuzzy clustering enables more adaptive information sharing and can enhance coherence across hierarchical levels by avoiding rigid groupings that may obscure shared trends.

In this paper, we consider the case of fuzzy hierarchies in (5). With fuzziness, the elements of the $\boldsymbol{U}$ matrix are no longer binary, and can take values between 0 and 1, indicating the likelihood of each time series being clustered in the c-th cluster. Therefore, to obtain a reconciliation with a fuzzy hierarchy, it is required to replace (5) with the following

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{1}'_{n_b} \\ \widehat{\boldsymbol{U}} \end{bmatrix}, \tag{7}$$

with $\widehat{\boldsymbol{U}}$ being a membership matrix of dimension $n_a \times k$ where each element of $\boldsymbol{U}$ takes value of between 0 and 1. The membership matrix $\widehat{\boldsymbol{U}}$ can be estimated using fuzzy clustering algorithms. In particular, the fuzzy k-means (Bezdek et al., 1984) finds it as the iterative solution to the following minimization problem

$$\min_{u_{ic}, \boldsymbol{b}_c} J(\boldsymbol{U}, \boldsymbol{B}) = \sum_{i=1}^{n_b} \sum_{c=1}^{k} u_{ic}^m \mid\mid \boldsymbol{b}_i, \boldsymbol{b}_c \mid\mid^2, \tag{8}$$

under the constraint $\sum_{c=1}^{k} u_{ic} = 1$, where $\boldsymbol{U}$ is the membership degree matrix with generic element $0 \leq u_{ic} \leq 1$ defining the membership degree of the ith unit to the c-th cluster, $\boldsymbol{B}_c$ the set of k cluster prototypes and $\mid\mid \boldsymbol{b}_i, \boldsymbol{b}_c \mid\mid^2$ is the Euclidean distance between the ith bottom time series and the c-th cluster prototype and $m > 1$ is the fuzzifier parameter. A value $m = 1$ of the fuzzifier parameter leads to the standard k-means algorithm; the fuzzifier parameter is usually set equal to $m = 2$, but it can also be chosen optimally based on grid search. The solutions to the problem (8) are given by

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n_b} u_{ic}^m \boldsymbol{b}_i}{\sum_{i=1}^{n_b} u_{ic}^m}, \tag{9}$$

for the c-th cluster prototype and

$$u_{ic} = \left(\sum_{c'=1}^{k} \left(\frac{||\boldsymbol{b}_i - \boldsymbol{b}_c||}{||\boldsymbol{b}_i - \boldsymbol{b}_{c'}||} \right)^{\frac{2}{m-1}} \right)^{-1}, \tag{10}$$

for the membership degrees.

We highlight that an interesting property of the fuzzy k-means clustering in this domain lies on the interpretation of the nodes of the so developed hierarchy, which are proportional to the c-th cluster prototype. As shown in (9), the new c-th node of the hierarchy equals a weighted average of the bottom time series included in the c-th cluster. This meets precisely the idea behind linearly constrained time series, and therefore this approach can be adopted to define further constraints that can be used to improve the accuracy of the reconciliation. We finally notice that, consistently with the preliminary analysis shown in Sect. 2, there is a substantial difference in emissions magnitudes across the European regions. Therefore, to achieve a better clustering quality, we rely on log-emissions while clustering the NUTS emissions curves.

3.3 A toy example

To clarify the aggregation structures and reconciliation strategies discussed above, we introduce a simple toy example. Suppose that $n_b = 4$ bottom-level time series are observed and are collected in the vector $\boldsymbol{b}_t = (b_{1t}, b_{2t}, b_{3t}, b_{4t})'$. In the case of GHG emission data, we can imagine $\boldsymbol{b}_t$ as the total emissions in the year t for four NUTS-2 regions. The simplest hierarchy consists of a single aggregate obtained as the sum of all bottom-level series. The corresponding aggregation matrix is

$$\boldsymbol{A}_0 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix},$$

so that the aggregate is $a_t = \boldsymbol{A}_0 \boldsymbol{b}_t = b_{1t} + b_{2t} + b_{3t} + b_{4t}$. In other words, the total emissions are equal to the sum of emissions in each of the NUTS. Suppose now that the four bottom-level series are partitioned into two clusters, namely $\mathcal{C}_1 = \{1, 2\}$ and $\mathcal{C}_2 = \{3, 4\}$. This clustering can be represented by the binary membership matrix

$$\widehat{\boldsymbol{C}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Stacking the total aggregate and the cluster-level aggregates yields the aggregation matrix

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

This structure corresponds to a hierarchy in which the bottom-level series aggregate both to cluster totals and to an overall total. In many applications, however, forcing each series to belong to a single cluster may be overly restrictive. Under fuzzy clustering, each series can partially belong to multiple clusters. For instance, consider the membership matrix

$$\widehat{\boldsymbol{U}} = \begin{bmatrix} 0.8 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.8 & 0.9 \end{bmatrix},$$

where each column sums to one. In this case the first unit belongs to Cluster 1 with 0.8 membership and with membership 0.2 to the Cluster 2. The same applies to other units. The resulting aggregation matrix then becomes

$$\widehat{\boldsymbol{A}}^{(f)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.8 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.8 & 0.9 \end{bmatrix}.$$

Compared to $\widehat{\boldsymbol{C}}$, and therefore to the aggregation matrix $\widehat{\boldsymbol{A}}$, the fuzzy aggregation matrix $\widehat{\boldsymbol{A}}^{(f)}$, which is based on the fuzzy membership matrix $\widehat{\boldsymbol{U}}$, directly captures the uncertainty in the clustering assignment. Suppose now that, in addition to the fuzzy clustering structure estimated from the data, a known and crisp aggregation structure is available. For instance, assume that bottom-level units $\{1, 2, 3\}$ belong to the same administrative region (Country A), while unit $\{4\}$ belongs to a different region (Country B). This known hierarchy can be represented by the binary aggregation matrix

$$\boldsymbol{C}_0 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where each row corresponds to a country-level aggregate. In this case, a hybrid hierarchy can be constructed by stacking the fuzzy clustering layer and the known administrative structure. The resulting aggregation matrix takes the form

$$\widehat{\boldsymbol{A}}^{(f,0)} = \begin{bmatrix} \boldsymbol{A}_0 \\ \widehat{\boldsymbol{U}} \\ \boldsymbol{C}_0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0.8 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.3 & 0.8 & 0.9 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

In this paper, we evaluate if the use of this mixed structure (that is, combining both data-driven and known clustering structures) leads to improvements in forecast reconciliation performances. In particular, we evaluate if the combination between fuzzy clustering and known structure improves upon crisp clustering.

4 Simulation study

We investigate the performance of forecast reconciliation based on fuzzy clustering and compare it with standard hard clustering approaches typically adopted in hierarchical reconciliation frameworks. In many practical settings, bottom-level time series are first grouped into clusters in order to construct an augmented aggregation structure used for reconciliation. However, such approaches implicitly assume a crisp partition of the series, which may be unrealistic when cross-series relationships are heterogeneous or partially overlapping.

In what follows, we consider a three-level hierarchical system composed of $n_B = 5$ bottom-level time series organized into two latent groups. The top-level series corresponds to the aggregation of all bottom series. Let $\mathbf{b}_t$ denote the vector of bottom-level observations at time t. In presence of K clusters, the aggregated series are constructed through a membership matrix $\mathbf{U} \in [0, 1]^{K \times n_B}$ satisfying $\sum_{k=1}^{K} U_{ki} = 1$, for $i = 1, \ldots, n_B$.

To generate coherent yet partially overlapping cluster structures, we assume that $K = 2$ clusters exist and that the dynamics of all bottom series are influenced by two latent AR(1) components, $\{Z_{1t}\}$ and $\{Z_{2t}\}$, representing common drivers associated with the two groups. Importantly, both components affect all series, but with different intensities. This induces partial memberships across clusters rather than a strict partition. Specifically, bottom-level series are generated as

$$b_{it} = w_i Z_{1t} + (1 - w_i) Z_{2t} + \varepsilon_{it}, \qquad i = 1, \ldots, n_B,\ t = 1, \ldots, T,$$

where $\varepsilon_{it} \sim$ i.i.d. $\mathcal{N}(0, \sigma_\varepsilon^2)$ and w_i determines the degree of association of each series with the two latent groups and therefore define the true clustering structure. Rather than imposing identical distributions across series, we introduce a controlled degree of overlap by sampling

$$w_i \sim \mathsf{Unif}(u_{1,i}, u_{2,i}),$$

where the intervals $(u_{1,i}, u_{2,i})$ differ across bottom series. In particular, some series exhibit strong association with one group, others moderate association, while at least one series is generated with near-balanced exposure to both components. This design avoids both degenerate hard partitions and fully homogeneous mixtures, producing an intermediate level of fuzziness representative of realistic clustering uncertainty. Given the weights, the true membership matrix is defined as

$$\mathbf{U}^{\star} = \begin{pmatrix} w_1 & \dots & w_{n_B} \\ 1 - w_1 & \dots & 1 - w_{n_B} \end{pmatrix}.$$

The latent components follow distinct AR(1) processes,

$$Z_{kt} = \phi_k Z_{k,t-1} + \eta_{kt}, \qquad \eta_{kt} \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2_{\eta,k}), \quad k \in \{1, 2\},$$

with $|\phi_k| < 1$ and $\phi_1 \neq \phi_2$.

Our simulation experiment considers as sample sizes $T \in \mathcal{T} = \{50, 100, 200, 500, 1000\}$ and as forecast horizons $h \in \mathcal{H} = \{1, 2, 3\}$. For each pair (T, h), we perform $B = 100$ Monte Carlo replications. Within each replication, bottom-level series are generated and coherent aggregated series are constructed using the true matrix $\mathbf{U}^{\star}$. Base forecasts are obtained using correctly specified AR(1) models in order to avoid misspecification effects unrelated to reconciliation.

A fuzzy membership matrix $\widehat{\mathbf{U}}$ is then estimated from the simulated bottom series using fuzzy k-means with $k = 2$, leading to the reconciliation structure

$$\widehat{\mathbf{S}}_{\text{fuzzy}} = \begin{pmatrix} \mathbf{1}'_{n_B} \\ \widehat{\mathbf{U}} \\ I_{n_B} \end{pmatrix}.$$

A hard clustering counterpart is obtained through maximum-membership assignment, yielding

$$\widehat{\mathbf{S}}_{\text{hard}} = \begin{pmatrix} \mathbf{1}'_{n_B} \\ \widehat{\mathbf{C}} \\ I_{n_B} \end{pmatrix}.$$

We compare unreconciled forecasts (Base), fuzzy clustering-based reconciliation, and hard clustering reconciliation. Reconciled forecasts are computed using a MinT-type projection. Forecast accuracy is evaluated across Monte Carlo replications using squared forecast errors computed at the top level and the average squared forecast errors at the bottom level of the hierarchy.

Table 1 reports the average results across the B replications. Overall, the results indicate that reconciliation improves forecasting accuracy relative to unreconciled forecasts, confirming that exploiting cross-series information helps reduce system incoherence. In particular, fuzzy clustering-based reconciliation consistently provides competitive or superior performance compared with both the baseline and hard clustering approaches.

A clear pattern emerges as the forecast horizon h increases. At the top level, forecasting errors naturally grow with h for all methods, reflecting the accumulation of prediction uncertainty. However, the relative improvement delivered by fuzzy reconciliation becomes larger for longer horizons. Considering the effect of increasing sample size, the main performance patterns remain remarkably stable as T grows. While larger samples generally reduce

Table 1 Average mean squared forecast errors across $B = 100$ Monte Carlo replications for unreconciled forecasts (Base), fuzzy clustering-based reconciliation (Fuzzy), and hard clustering reconciliation (Crisp)

T	h	Top			Bottom (average)		
		Base	Fuzzy	Crisp	Base	Fuzzy	Crisp
50	1	14.8035	14.2804	15.7910	0.7058	0.7652	0.8224
50	2	20.6202	19.8479	20.7201	0.9889	1.0163	1.0508
50	3	24.3648	23.4926	23.7417	1.1804	1.1814	1.1931
100	1	13.9207	13.5238	14.9652	0.6702	0.7369	0.7892
100	2	19.6187	18.4722	19.0427	0.9395	0.9600	0.9799
100	3	22.5574	21.4490	21.7303	1.0925	1.0930	1.1022
200	1	13.6003	13.6703	15.5434	0.6527	0.7447	0.8133
200	2	20.1406	19.7412	20.1955	0.9668	1.0075	1.0233
200	3	23.3318	22.5777	22.9868	1.1214	1.1330	1.1486
500	1	13.7656	13.7552	15.6332	0.6732	0.7502	0.8190
500	2	19.2119	17.9744	19.2272	0.9113	0.9300	0.9762
500	3	22.8467	21.5380	22.3100	1.0924	1.0841	1.1132
1000	1	13.1472	13.2875	15.7980	0.6338	0.7385	0.8260
1000	2	18.8320	17.7775	19.2149	0.9023	0.9335	0.9837
1000	3	21.7542	20.3114	21.2065	1.0401	1.0425	1.0740

Results are reported for different sample sizes T and forecast horizons h. Forecast accuracy is evaluated at both the top level of the hierarchy and at the bottom level, where errors are averaged across all bottom series. Lower values indicate better forecasting performance

forecast errors for all approaches, the relative ranking of the methods does not substantially change. This suggests that the advantage of fuzzy reconciliation is not driven by small-sample variability but rather by its ability to better represent the underlying dependence structure among series. We finally remark that differences across methods appear to be smaller at the bottom level than top. Nevertheless, fuzzy reconciliation still improves compared to crisp reconciliation.

5 Application: forecasting agricultural GHGs emissions across European NUTS-2 regions

We aim at forecasting the agricultural GHGs time series described in Sect. 2, which have annual frequency, are of short length and are characterized by different trends. Therefore, given the limited data length and these characteristics, we consider exponential smoothing to forecast next emission levels, which can be estimated considering the following state-space form

$$\begin{aligned} y_t &= \ell_{t-1} + c_{t-1} + \varepsilon_t, \\ \ell_t &= \ell_{t-1} + c_{t-1} + \alpha\varepsilon_t, \\ c_t &= c_{t-1} + \beta\varepsilon_t, \end{aligned} \tag{11}$$

where ℓ_t denotes an estimate of the level of the series at time t, c_t denotes an estimate of the trend (slope) of the series at time t, α is the smoothing parameter for the level (with $0 \leq \alpha \leq 1$) and β is the smoothing parameter for the trend (with $0 \leq \beta \leq 1$). The parameters are estimated by maximum likelihood assuming that errors are normally and independently distributed (Hyndman et al., 2008). We remark that the model (11) corresponds to the innovations form of the local linear trend model (Durbin & Koopman, 2012). Other statistical models can also be considered for generating baseline forecasts, such ARIMA (with optimal order selection, e.g. see Hyndman and Khandakar, 2008).

Given base forecasts obtained with Eq. (11), we then evaluate the accuracy of alternative reconciliation approaches in forecasting both total emissions in Europe (top-level series) and the total emissions in each European NUTS-2 area (bottom-level series). In particular, we compare bottom-up reconciliation and optimal reconciliation obtained using alternative aggregation structures as discussed in the next sub-section. The proposed approach can be easily implemented using the `FoReco` `R` library (Girolimetto & Fonzo, 2022).

To compare alternative forecasts of agricultural emissions, we adopt the following rolling-window experiment. Given $T = 34$ years available, we keep $M = 24$ years as the estimation window and consider the last 10 years of observations to evaluate out-of-sample metrics. We first estimate the model from $t = 1$ to $t = M$ and generate h step ahead forecasts with model (11) for $t = M + h$ for all the time series in the hierarchy, thus obtaining $\widehat{\boldsymbol{y}}_h$. The latter are then reconciled in different manners, thus obtaining r alternative reconciliation approaches, namely $\widetilde{\boldsymbol{y}}_h(1), \ldots, \widetilde{\boldsymbol{y}}_h(r)$, where r is the number of alternative reconciliation considered. We determine the time series clustering structure from $t = 1$ to $t = M$ while implementing cluster-based reconciliation approaches for the first iteration. The rolling window procedure then continues by removing the oldest observation and including the newest, so that the second iteration involves clustering from $t = 2$ to $t = M + 1$ and forecasts h steps ahead. The procedure continues until the forecast for $t = T$ is produced and no new observation is available.

To evaluate the accuracy obtained for the top-level series, that is, the total agricultural emissions in Europe, we consider the $(T - M)$ out-of-sample forecast errors and compute both Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{N_h} \sum_{t=1}^{N_h} \left(y_{t,\text{top}} - \widehat{y}_{t,\text{top}}\right)^2}, \tag{12}$$

and Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{N_h} \sum_{t=1}^{N_h} \left|y_{t,\text{top}} - \widehat{y}_{t,\text{top}}\right|. \tag{13}$$

with $N_h = T - M - h + 1$. The evaluation of n_b bottom-level series, that is, the regional GHGs emissions, poses some additional challenges (Athanasopoulos & Kourentzes, 2023). Among the others, we have that NUTS emissions may vary significantly in their magnitude depending on territorial extension or other factors. Therefore, the average of RMSE and MAE obtained for the n_b NUTS is not a suitable choice to consider. Indeed, good forecasting for regions with lower emissions would be underrepresented using this approach, while inaccuracies of forecasts for NUTS with larger emissions would be exacerbated. For a fair comparison, we replace RMSE and MAE by measures computed considering scaled errors, which are independent of the scale of the data. In particular, we employ the scaled error as defined in Hyndman and Koehler (2006), which is less than one if the reconciled forecasts provide a better result than the average one-step naive forecast computed in-sample. We can then consider Mean Absolute Scaled Error (MASE) or Root Mean Square Scaled Error (RMSSE) for each NUTS and then obtain an overall accuracy measure by taking the cross-sectional average of the obtained MASEs and RMSSEs.

5.1 Aggregation matrices construction

Different aggregation structures lead to alternative reconciliations and, therefore, alternative forecasts. In what follows, we compare many alternatives, distinguishing between fuzzy and crisp (i.e., non fuzzy) approaches in constructing the hierarchies.

First, we consider a baseline reconciliation based on a pure spatial aggregation matrix, called $\boldsymbol{C}_0$, where the n_b NUTS time series aggregate to the respective countries and, then, all the countries aggregate to the total emissions in Europe, that is

$$\boldsymbol{A}_0 = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \boldsymbol{C}_0 \end{bmatrix}. \tag{14}$$

We have already discussed this natural aggregation, being of dimension $n_a \times n_b = 34 \times 309$. This is a known hierarchy, and it is not possible to consider a fuzzy version for this approach.

Second, we consider the following aggregation. We start considering $\boldsymbol{C}_0$ and obtaining the $(n_a - 1) = 36$ time series for each country, each equal to the sum of the emissions for each of the n_b NUTS belonging to that country, that is,

$$\boldsymbol{c}_{0t} = \boldsymbol{C}_0 \mathbf{b}_t. \tag{15}$$

We then cluster the resulting time series $\boldsymbol{c}_{0t}$ according to their shape-based similarity, thus by means of (8). Then, a second level of the hierarchy is included based on the estimated membership degree matrix $\widehat{\boldsymbol{U}}_1$ as follows

$$\widehat{\boldsymbol{A}}_1 = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \widehat{\boldsymbol{U}}_1 \boldsymbol{C}_0 \\ \boldsymbol{C}_0 \end{bmatrix}. \tag{16}$$

The matrix product $\widehat{\boldsymbol{U}}_1 \boldsymbol{C}_0$ is required to make sure that all the n_b NUTS in a country have the same membership degree. The crisp version of (16) can be simply obtained by replacing $\widehat{\boldsymbol{U}}_1$ with $\widehat{\boldsymbol{C}}_1$ which assigns each time series to the cluster c whose membership is the highest.

Another approach may look at similarities in the NUTS time series to define the fuzzy clustering rather than the aggregated country-level time series. In this case, we apply exactly the model (8) considering the bottom time series $\boldsymbol{b}_t$ which leads to the membership degree matrix $\widehat{\boldsymbol{U}}_2$ and to the following aggregation matrix

$$\widehat{\boldsymbol{A}}_2 = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \widehat{\boldsymbol{U}}_2 \\ \boldsymbol{C}_0 \end{bmatrix}. \tag{17}$$

Also in this case the crisp version can be obtained by replacing $\widehat{\boldsymbol{U}}_2$ with $\widehat{\boldsymbol{C}}_2$ which assigns each time series to the cluster c whose membership is the highest.

Finally, we can combine both structures, thus obtaining

$$\widehat{\boldsymbol{A}}_3 = \begin{bmatrix} \mathbf{1}'_{n_b} \\ \widehat{\boldsymbol{U}}_2 \\ \widehat{\boldsymbol{U}}_1 \boldsymbol{C}_0 \\ \boldsymbol{C}_0 \end{bmatrix}. \tag{18}$$

We notice that (14) and (17) can be seen as restricted versions of (18). Therefore, other restricted versions implying only clustering without a known structure like $\boldsymbol{C}_0$ could also be considered.

To obtain these structures in practice, at each iteration of the rolling window forecasting experiment, clustering is performed independently on country-level and bottom-level (NUTS2) time series using Fuzzy K-means clustering. We use the simple Euclidean distance and we opt for a raw data-based approach according to the standard taxonomy (Maharaj et al., 2019). Specifically, each unit is represented by its emission trajectory within the current training block. Therefore, clustering is performed on temporal emission profiles rather than on static attributes. We remark that raw data-based approaches are useful for measuring shape-based similarities across short time series. To avoid distortions due to highly heterogeneous emission magnitudes across regions, a logarithmic transformation is applied, preserving zeros while compressing large values. No further scaling is applied, as the log transformation mitigates scale disparities.

The number of clusters is not fixed a priori. Instead, for each rolling iteration a clustering partition is computed for candidate values $k = 2, \ldots, k_{\max}$ with $k_{\max} = 10$, and the optimal value is selected by maximizing the fuzzy silhouette index (Campello & Hruschka, 2006) at each iteration. This allows aggregation structures to adapt over time as emission dynamics evolve. The fuzzifier parameter controlling membership smoothness is set to $m = 2$ in the main experiments, while a sensitivity analysis is conducted for $m \in \{1.1, 1.5, 2, 2.5\}$. Since the optimal number of clusters may vary across rolling windows, clustering stability is evaluated through the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) computed between consecutive iterations.

5.2 Main results

The benefit of reconciliation can be appreciated at both top and bottom level forecasts. Actually, also the accuracy on middle level series can be improved by reconciliation but in our context the only relevant level from policy perspective is that associated with country-level forecasts. Therefore, in what follows we evaluate three levels of accuracy, that is, Europe (top-level), European countries (middle-level) and NUTS (bottom-level).

Top-level. Table 2 reports the relative improvements in RMSE obtained through reconciliation with respect to base forecasts across different forecasting horizons. Results clearly show that reconciliation systematically improves top-level forecasts for all considered horizons. While bottom-up reconciliation already provides substantial gains, optimal reconciliation further improves forecast accuracy, with improvements remaining large across horizons and exceeding 20% for medium and long horizons. Similar results are found considering MAE-related metrics. Among the considered aggregation structures, the combined hierarchy (18) consistently achieves the best or near-best performance.

Comparing crisp and fuzzy aggregation structures shows that fuzzy clustering systematically provides slightly larger improvements despite producing hierarchies of identical size. This suggests that soft memberships offer a more flexible aggregation structure, leading to more accurate reconciled forecasts for total agricultural GHG emissions in Europe.

Middle-level. Forecasts of the emissions at the country level are also important from a policy perspective, as they represent an administrative aggregate relevant for the EU agricultural policy. In what follows, we show the results of the middle levels of the hierarchy, excluding middle levels involving clusters. In other words, we here focus only on the accuracy of the reconciliation considering country aggregates. Table 3 shows the average out-of-sample

Table 2 Relative improvement (%) in terms of RMSE versus Base forecasts for top level series using ETS baseline forecasts across horizons. Optimal MinT reconciliation (4) is adopted

Method	h = 1	h = 2	h = 3
Bottom-Up	9.74	22.33	34.29
$\boldsymbol{A}_0$	11.90	22.48	33.89
$\widehat{\boldsymbol{A}}_1$ (crisp)	12.50	22.60	33.84
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	12.80	22.71	33.89
$\widehat{\boldsymbol{A}}_2$ (crisp)	12.88	22.70	33.84
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	13.35	22.83	33.89
$\widehat{\boldsymbol{A}}_3$ (crisp)	13.28	22.76	33.79
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	13.91	22.97	33.89

Table 3 Relative improvement (%) versus Base forecasts in terms of MASE across horizons using ETS baseline forecasts, reported for bottom and middle aggregation levels. Optimal MinT reconciliation (4) is adopted

	Middle-level			Bottom-level		
Method	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3
$\boldsymbol{A}_0$	−0.39	−0.38	−0.29	0.10	0.13	0.00
$\widehat{\boldsymbol{A}}_1$ (crisp)	−0.30	−0.36	−0.33	0.17	0.12	−0.05
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	−0.21	−0.32	−0.30	0.23	0.16	−0.02
$\widehat{\boldsymbol{A}}_2$ (crisp)	−0.22	−0.37	−0.42	0.23	0.13	−0.08
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	−0.13	−0.31	−0.39	0.26	0.15	−0.07
$\widehat{\boldsymbol{A}}_3$ (crisp)	−0.18	−0.38	−0.45	0.26	0.10	−0.14
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	−0.02	−0.28	−0.39	0.34	0.17	−0.09

accuracy for middle and bottom levels time series, considering the MASE accuracy metric. As discussed previously, the average MASE across countries provides a good measure of overall accuracy as the MASE does not reflect differences in terms of scale in the total emissions. As in Table 2, the results are reported as percentage improvement compared to baseline forecasts. The first column shows the results for the country-level aggregates and the second NUTS-level ones.

Table 3 shows that reconciliation does not provide improvements at short horizons for country-level aggregates, and that performance tend to decrease for longer horizons. Across aggregation structures, fuzzy clustering typically yields slightly better results than crisp clustering, although differences remain limited in magnitude. Therefore, reconciliation does not contributes to accuracy improvements at this level.

Bottom-level. In the second column of Table 3, we focus on the accuracy of the forecasts at the bottom level of the hierarchy, which corresponds to the NUTS-level emissions. As in the case of the middle level, the fuzzy clustering-based reconciliation also achieves the best performance in terms of out-of-sample accuracy. Improvements at the bottom (i.e., regional) level are smaller than those observed for aggregated series but remain positive for short horizons in most aggregation structures. Fuzzy clustering-based reconciliation consistently performs at least as well as, and often slightly better than, crisp alternatives. These results suggest that reconciliation mainly redistributes information across series, providing larger benefits for aggregated series while still offering moderate gains at the most granular level of the hierarchy.

Overall, reconciliation effectively provides the largest benefits at the top level, with improvements increasing with the forecasting horizon, making it particularly effective for medium- and long-term projections of total GHG emissions in Europe. At bottom levels, gains are more moderate and mainly concentrated at short horizons. Therefore, given that reconciliation significantly improves aggregate forecasts without deteriorating accuracy at more disaggregated levels, we find that the proposed approach is suitable for policy-oriented forecasting applications.

We also investigate the stability of the clustering structure across rolling-window iterations in order to assess the practical usefulness of dynamic cluster re-estimation. At the country level, the average number of selected clusters is equal to 4.3 with standard deviation 2.75, indicating that the optimal number of clusters moderately varies across rolling windows. The corresponding average Adjusted Rand Index (ARI, Hubert and Arabie, 1985) between consecutive partitions is 0.70 (standard deviation 0.35), suggesting a reasonably stable clustering structure, given that comparison in terms of ARI for two solutions with different number of clusters naturally lead to a lower score. At the bottom level, the clustering structure is even more stable. Indeed, the selected number of clusters remains constant across iterations (average 5.0, standard deviation 0.0), while the average ARI between consecutive partitions reaches 0.99 (standard deviation 0.01). This indicates that the clustering structure at the most disaggregated level is highly stable over time.

Fig. 3 reports in detail the spatial distribution of the out-of-sample MASE for different forecasting horizons. The left panel refers to the bottom-level series, whereas the right panel presents the country-aggregated (middle-level) results.

Two main patterns emerge. First, at the bottom level there is no clear deterioration in forecast accuracy as the forecasting horizon increases. The spatial distribution of the errors remains relatively stable across different values of h, suggesting that the model preserves its predictive performance even at longer horizons for disaggregated series. In contrast, at the country-aggregate level a progressive deterioration in accuracy is observed as the forecasting horizon increases. This effect, however, is not uniform across Europe. Forecast accuracy remains satisfactory for most Eastern European countries, even at longer horizons. The deterioration is primarily concentrated in a subset of countries, notably Germany, France, Spain, Portugal, Ireland, Iceland, and Turkey, where aggregated forecast errors become comparatively larger.

An important aspect highlighted by Fig. 3 is that this worsening performance at the country-aggregate level is not mirrored at the bottom level within the same countries. The NUTS-level series belonging to Germany or France, for example, do not display systematically lower forecast accuracy compared to other European regions. On the contrary, their bottom-level performance is broadly comparable to that of other countries. The deterioration therefore emerges mainly after aggregation. This finding suggests that the overall country-level results are driven by specific national aggregation effects rather than by uniformly poor local predictive performance.

From a policy perspective, this has relevant implications. While the proposed modeling strategy performs satisfactorily for most countries, the evidence indicates that a uniform modeling approach may not be equally suitable for all national aggregates. In particular, for the countries exhibiting persistent deterioration at higher horizons, alternative model specifications or country-specific calibration strategies may be required to improve aggregated forecast accuracy.

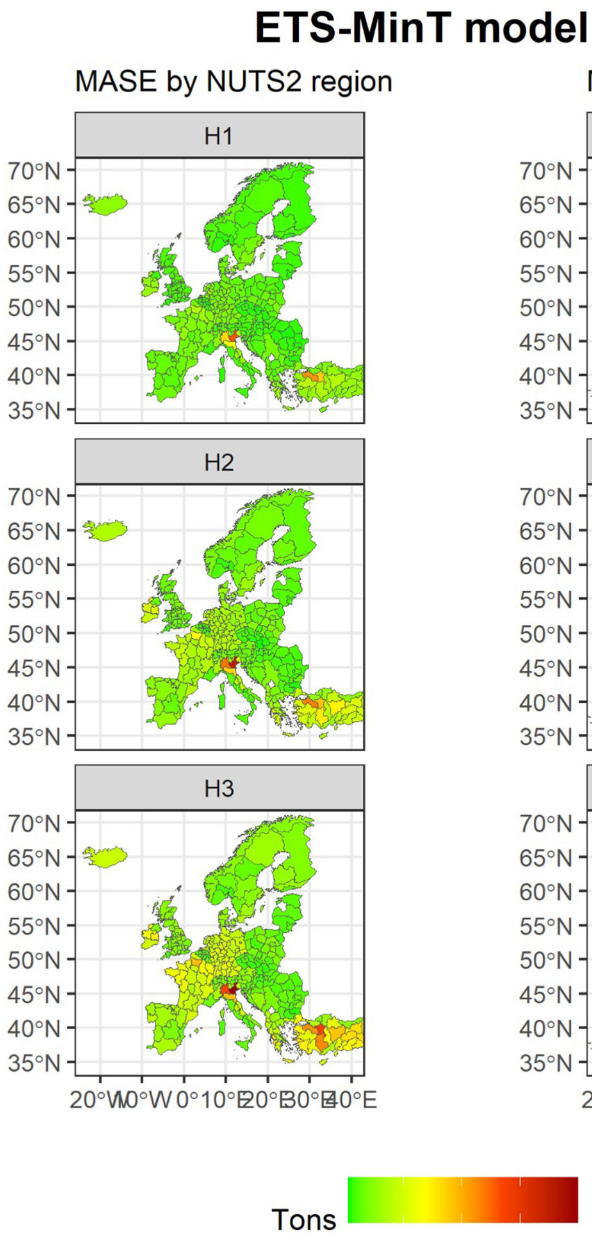


Fig. 3 Spatial distribution of the obtained out-of-sample MASE for different forecasting horizons h at the regional level (left column) and at the country level (right column)

Table 4 Relative improvement (%) in terms of RMSE versus Base forecasts for top level series using ETS baseline forecasts across horizons

Method	OLS reconciliation			WLS reconciliation		
	h=1	h=2	h=3	h=1	h=2	h=3
$\boldsymbol{A}_0$	16.82	24.22	34.69	12.16	23.24	34.77
$\widehat{\boldsymbol{A}}_1$ (crisp)	17.37	24.53	34.87	12.66	23.36	34.77
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	17.29	24.48	34.84	12.77	23.42	34.81
$\widehat{\boldsymbol{A}}_2$ (crisp)	17.11	24.39	34.79	12.99	23.50	34.81
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	17.01	24.32	34.74	13.25	23.51	34.78
$\widehat{\boldsymbol{A}}_3$ (crisp)	17.51	24.62	34.92	13.36	23.59	34.81
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	17.36	24.52	34.85	13.65	23.65	34.82

OLS and WLS reconciliation approaches are compared

5.3 Robustness and sensitivity

5.3.1 Alternative optimal reconciliation approaches

The MinT reconciliation is the most widely used strategy in empirical studies. However, several alternative approaches have been proposed for finding optimal mapping matrices $\boldsymbol{G}$; thus, we believe it is important to compare alternative reconciliation approaches as the best method can be different case-by-case. In what follows, we consider the results in terms of two well-known reconciliation approaches, namely the Ordinary Least Squares (OLS) and the Weighted Least Squares (WLS). Moreover, such robustness is relevant to evaluate the usefulness of clustering and in particular of fuzzy clustering as proposed in the paper.

The OLS approach has been proposed in Hyndman et al. (2011) and considers an optimal projection matrix equal to

$$\boldsymbol{G}_{OLS} = \left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1}\boldsymbol{S}, \tag{19}$$

obtained by regressing the base forecasts $\widehat{\boldsymbol{y}}_h$ with the summation matrix $\boldsymbol{S}$.

The main drawback of the OLS approach (19) is that it may be slow for high-dimensional hierarchies and that all the series in the grouped structure are equally weighted. Although efficient approaches to handling computational complexity have been proposed (e.g. see Ashouri et al., 2022), the assumption of equal weights limits the out of sample performance of this reconciliation approach. Hyndman et al. (2016) overcomes this issue by proposing a variance-scaling, namely the WLS approach, such that

$$\boldsymbol{G}_{WLS} = \left(\boldsymbol{S}'\boldsymbol{\Lambda}^{-1}\boldsymbol{S}\right)^{-1}\boldsymbol{S}\boldsymbol{\Lambda}^{-1}, \tag{20}$$

with $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{W}_h)$ being the diagonal of the covariance matrix of the h-steps ahead forecast errors, $\boldsymbol{W}_h$. According to (20), forecasts with larger variance are weighted less.

Tables 4 and 5 offer a comprehensive comparison of reconciliation strategies applied to hierarchical forecasts of agricultural emissions, focusing respectively on the total European emissions and the disaggregated middle and bottom levels using OLS and WLS reconciliation approaches. As in other tables, results are provided as improvements compared to baseline (non-reconciliated) forecasts.

Several key findings emerge from the table. First, it confirms that reconciliation-based methods improve top-level forecasts accuracy, with increasing effect with forecasting horizon h. Second, fuzzy clustering-based reconciliation outperforms its crisp counterparts across for the WLS reconciliation. In contrast, for OLS-based reconciliation the crisp approach performs

Table 5 Relative improvement (%) versus Base forecasts across horizons at middle and bottom levels using ETS baseline forecasts

	OLS reconciliation			WLS reconciliation		
Method	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3
Panel A: Middle level						
$\boldsymbol{A}_0$	−10.99	−5.30	−3.58	−0.22	−0.14	−0.06
$\widehat{\boldsymbol{A}}_1$ (crisp)	−8.76	−6.51	−5.46	−0.08	−0.12	−0.08
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	−7.43	−5.36	−4.26	−0.07	−0.07	−0.05
$\widehat{\boldsymbol{A}}_2$ (crisp)	−9.29	−4.87	−3.52	−0.05	−0.09	−0.08
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	−9.56	−5.17	−3.65	0.02	−0.06	−0.09
$\widehat{\boldsymbol{A}}_3$ (crisp)	−7.92	−5.99	−5.19	0.06	−0.07	−0.10
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	−6.81	−4.97	−4.10	0.12	−0.00	−0.08
Panel B: Bottom level						
$\boldsymbol{A}_0$	−37.55	−20.18	−13.97	0.19	0.12	0.09
$\widehat{\boldsymbol{A}}_1$ (crisp)	−35.65	−19.66	−14.23	0.23	0.13	0.09
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	−34.42	−18.76	−13.56	0.24	0.14	0.10
$\widehat{\boldsymbol{A}}_2$ (crisp)	−47.23	−26.96	−19.39	0.25	0.15	0.09
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	−46.66	−26.16	−18.35	0.25	0.14	0.08
$\widehat{\boldsymbol{A}}_3$ (crisp)	−44.17	−25.76	−19.11	0.28	0.16	0.09
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	−41.94	−23.95	−17.46	0.28	0.16	0.09

Results are reported for both OLS and WLS reconciliation approaches

better. In general, it is interesting to note that improvements in forecast accuracy for the top-level series are slightly larger (although comparable) using OLS and WLS reconciliation than MinT.

Table 5 reports the performance of OLS and WLS reconciliation across middle and bottom aggregation levels.

The general behavior observed under MinT reconciliation remains largely visible: forecast accuracy deteriorates when moving from top to more disaggregated levels, and reconciliation gains become smaller and more heterogeneous across horizons. However, important differences emerge across reconciliation schemes. In particular, OLS reconciliation systematically deteriorates forecast accuracy at both middle and bottom levels, producing large loss of accuracy compared to the baseline across all horizons h. This behavior reflects a well-known limitation of OLS reconciliation, which assigns equal weights to all series and ignores differences in forecast error variability. In large hierarchies with heterogeneous noise levels, such uniform weighting tends to propagate errors from noisy series to more stable aggregates, resulting in substantial performance degradation, especially at highly disaggregated levels.

A further aspect concerns the comparison between crisp and fuzzy aggregation structures. Under WLS reconciliation, fuzzy aggregation generally performs as well as or slightly better than its crisp counterpart, especially for the combined clustering structures, indicating that allowing partial memberships helps capture overlapping regional dynamics. In contrast, under OLS reconciliation the differences between crisp and fuzzy variants become marginal, suggesting that when the reconciliation scheme itself is poorly specified, improvements in clustering structure have limited impact on final forecast accuracy. Overall, these results still favor reconciliation methods, particularly when combined with clustering. In fact, we

Table 6 Relative improvement (%) versus Base forecasts using ARIMA baseline forecasts across horizons at top, middle, and bottom levels

	Top level			Middle level			Bottom level		
Method	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3
$\boldsymbol{A}_0$	−7.06	4.24	12.58	0.88	1.33	2.58	0.08	−0.05	−1.18
$\widehat{\boldsymbol{A}}_1$ (crisp)	−8.45	3.18	11.52	0.36	0.26	1.68	−0.34	−0.63	−1.87
$\widehat{\boldsymbol{A}}_1$ (fuzzy)	−8.75	4.40	14.69	−0.48	−0.01	1.17	−0.48	−0.86	−2.01
$\widehat{\boldsymbol{A}}_2$ (crisp)	−4.19	7.16	15.46	0.44	1.07	2.18	0.23	0.15	−1.29
$\widehat{\boldsymbol{A}}_2$ (fuzzy)	0.73	9.93	14.46	0.87	0.98	0.90	0.13	0.20	−2.06
$\widehat{\boldsymbol{A}}_3$ (crisp)	−5.61	6.01	14.24	0.08	0.40	1.52	−0.06	−0.21	−1.77
$\widehat{\boldsymbol{A}}_3$ (fuzzy)	−1.32	9.53	15.73	−0.30	0.13	0.08	−0.34	−0.39	−2.62

evidence a practical value of fuzziness, thus confirming the theoretical advantage of fuzzy clustering in modeling uncertainty in the clustering structure.

5.3.2 Results with ARIMA baseline forecasts

For robustness purposes, we repeat the experiment by replacing the ETS baseline forecasts with ARIMA models estimated through the automatic selection procedure `auto.arima` (Hyndman & Khandakar, 2008) in each iteration. Table 6 reports the relative forecast improvements obtained through reconciliation across aggregation levels and forecast horizons.

Results confirm that reconciliation remains beneficial also under this alternative baseline specification. In particular, at the top level, reconciliation consistently improves forecast accuracy, with gains becoming more pronounced as the forecast horizon increases. Improvements are modest at short horizon $h = 1$, but become clearly relevant for $h = 2$ and especially for $h = 3$, indicating that the contribution of cross-sectional coherence becomes more relevant at longer horizons. At the middle aggregation level, improvements are present but generally smaller in magnitude, suggesting that country-level forecasts already incorporate part of the relevant information at the baseline stage. Still, reconciliation typically yields positive gains for medium and longer horizons. Finally, at the bottom level, improvements are more limited and in some cases slightly negative. This pattern is consistent with those shown in Table 3.

To further assess robustness, the entire experiment of Sect. 5.3.1, in which OLS and WLS reconciliation strategies were employed, is repeated using the ARIMA models to generate the baseline forecasts. The results, showed in Table 7, confirm patterns already observed with ETS models.

Indeed, reconciliation yields the largest and most consistent improvements at the top aggregation level, with gains increasing as the forecasting horizon grows. At middle and bottom levels, improvements are generally smaller and sometimes negligible, but reconciliation based on MinT and WLS rarely deteriorates forecast accuracy compared to unreconciled forecasts. However, OLS reconciliation exhibits unstable behavior in disaggregated hierarchies. Although it may still improve top-level forecasts, it systematically deteriorates accuracy at middle and bottom levels. This effect is particularly strong under ARIMA baseline forecasts, confirming that OLS is not suitable for large and heterogeneous hierarchies where series display markedly different error variances. Across reconciliation schemes and forecasting models, fuzzy clustering-based aggregation structures remain consistently among the

Table 7 Relative improvement (%) versus Base forecasts using ARIMA baseline forecasts across horizons at top, middle, and bottom aggregation levels under OLS and WLS reconciliation

	Top level			Middle level			Bottom level		
Method	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3	h = 1	h = 2	h = 3
Panel A: OLS reconciliation									
A_0	0.52	0.95	1.39	−121.18	−158.51	−195.55	−154.98	−190.38	−216.51
$\widehat{A}_1$ (crisp)	−1.77	1.38	3.20	−68.02	−57.77	−64.98	−109.80	−123.24	−131.62
$\widehat{A}_1$ (fuzzy)	−2.10	2.12	5.10	−80.52	−77.39	−93.22	−118.99	−134.70	−144.20
$\widehat{A}_2$ (crisp)	0.27	2.60	4.72	−107.37	−143.58	−177.71	−195.71	−230.23	−266.32
$\widehat{A}_2$ (fuzzy)	0.92	2.38	3.14	−110.27	−147.92	−184.13	−164.88	−187.61	−201.04
$\widehat{A}_3$ (crisp)	−1.72	2.55	5.51	−60.94	−44.59	−49.67	−186.72	−210.04	−241.59
$\widehat{A}_3$ (fuzzy)	−1.45	3.07	5.96	−73.54	−69.91	−86.08	−156.06	−164.15	−170.79
Panel B: WLS reconciliation									
A_0	−11.31	3.81	14.49	1.34	2.16	3.45	−0.02	−0.09	−0.05
$\widehat{A}_1$ (crisp)	−9.97	4.95	15.75	1.44	1.88	3.71	0.02	−0.01	0.10
$\widehat{A}_1$ (fuzzy)	−10.36	5.65	17.27	1.16	2.23	4.19	−0.04	−0.00	0.12
$\widehat{A}_2$ (crisp)	−10.30	4.76	15.77	1.35	2.17	3.75	0.10	0.02	0.10
$\widehat{A}_2$ (fuzzy)	−8.28	6.38	16.99	1.64	2.39	3.75	0.15	0.12	0.20
$\widehat{A}_3$ (crisp)	−9.22	5.63	16.69	1.34	1.80	3.78	0.14	0.07	0.20
$\widehat{A}_3$ (fuzzy)	−7.94	7.44	18.74	1.28	2.27	3.94	0.11	0.16	0.31

best-performing solutions, with the combined fuzzy hierarchy often achieving the highest or near-highest improvements, especially at longer horizons.

In synthesis, with respect the forecasting horizon, the reconciliation effectiveness tends to increase as h increases, highlighting the growing importance of cross-sectional information as forecasting uncertainty increases. Moreover, fuzzy aggregation structures generally perform better, or at least as well as, their crisp counterparts, particularly at longer horizons. This confirms that allowing series to contribute to multiple aggregation groups provides a more flexible representation of cross-sectional relationships, which translates into more robust reconciliation performance. The lack of considerable improvements at the bottom suggests that reconciliation mainly redistributes forecast adjustments required to achieve coherence across higher levels, which may not always translate into bottom-level accuracy gains. However, we remark that the good performance for the top-level series validates reconciliation as a valid tool to support environmental policies at the aggregate level.

A similar spatial analysis is reported for the ARIMA baseline forecasts. Figure 4 indeed displays the distribution of out-of-sample MASE at both the bottom level and the country-aggregate (middle) level.

The geographical patterns are largely consistent with those observed under exponential smoothing, that is Fig. 3. At the bottom level, no pronounced concentration of high errors emerges, apart from limited localized effects (see, for instance, the Po Valley in Northern Italy). Overall, bottom-level forecast accuracy remains relatively homogeneous across Europe. At the country-aggregate level, however, the same pattern of deterioration is visible: the countries previously identified—such as Germany, France, Spain, Portugal, Ireland, Iceland, and Turkey—again exhibit comparatively higher forecast errors compared to other countries. Importantly, this occurs despite the use of a different baseline forecasting model. Although ARIMA combined with reconciliation leads to a relative improvement over its own baseline—particularly at the middle level, where performance gains are more pronounced compared to the ETS case—the spatial concentration of higher errors remains essentially unchanged.

This result is informative, as it indicates that the observed weaknesses at the country-aggregate level are not primarily driven by the choice of the baseline forecasting model or by the reconciliation procedure itself. Even when ARIMA reconciliation improves upon the unreconciled ARIMA baseline forecasts, the same national aggregates remain comparatively more difficult to predict. In other words, this suggests that, while reconciliation improves accuracy relative to the respective baseline, certain country aggregates exhibit intrinsic forecasting complexity that is robust to changes in the baseline specification.

5.3.3 Sensitivity to fuzzifier paremter *m*

To further assess the robustness of the proposed reconciliation approach, we investigate the sensitivity of the results with respect to the fuzzifier parameter m, which controls the degree of softness in the clustering memberships used to construct fuzzy aggregation structures. In fuzzy k-means clustering, the fuzzifier regulates how strongly each time series belongs to the different clusters. When $m \to 1$, the membership matrix $\boldsymbol{U}$ converges to a crisp partition, where each series is assigned almost entirely to a single cluster. Conversely, for larger values of m, memberships become increasingly diffuse and series contribute more uniformly to multiple clusters. Therefore, the choice of m directly affects the aggregation constraints used in reconciliation and may influence forecasting performance. Although $m = 2$ is commonly adopted in the literature and is used for our main experiments, it is important to verify whether the observed gains depend critically on this choice. Figure 5 reports percentage

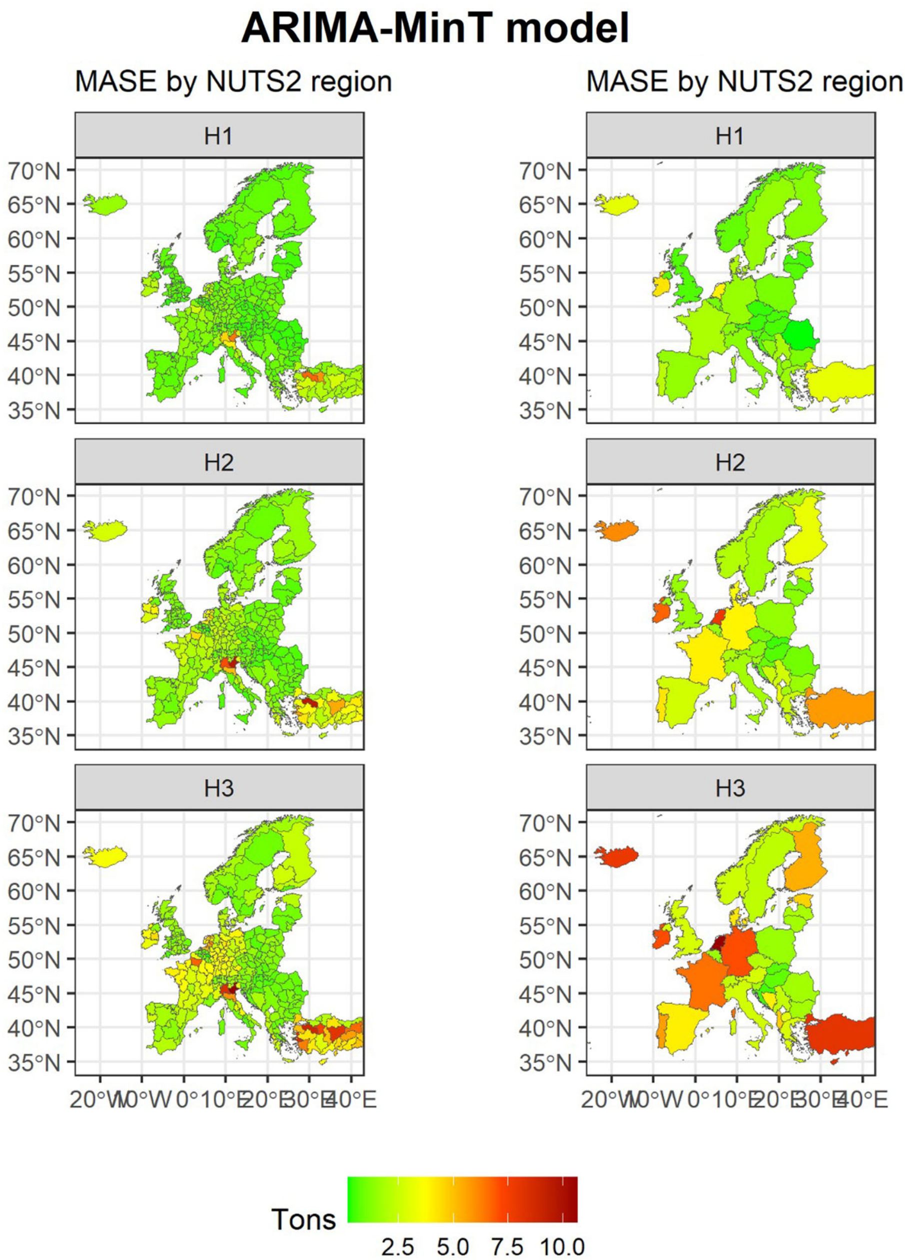


Fig. 4 Spatial distribution of the obtained out-of-sample MASE for different forecasting horizons h with ARIMA baseline forecasts: country-level (left) and NUTS-level (right)

improvements relative to baseline forecasts across forecast horizons and fuzzifier values $m \in \{1.1, 1.5, 2, 2.5\}$ for both ETS and ARIMA baseline models and for top, middle, and bottom aggregation levels. At the top level, improvements are consistently observed across all tested values of m and forecast horizons. Results are stable across fuzzifier choices, showing that reconciliation gains are not driven by a particular fuzziness configuration. For ARIMA baselines, improvements are smaller or even slightly negative at horizon $h = 1$, in agreement with the robustness results reported in Table 7, but become clearly positive as the forecast horizon increases. In contrast, ETS-based forecasts show stable improvements across all horizons.

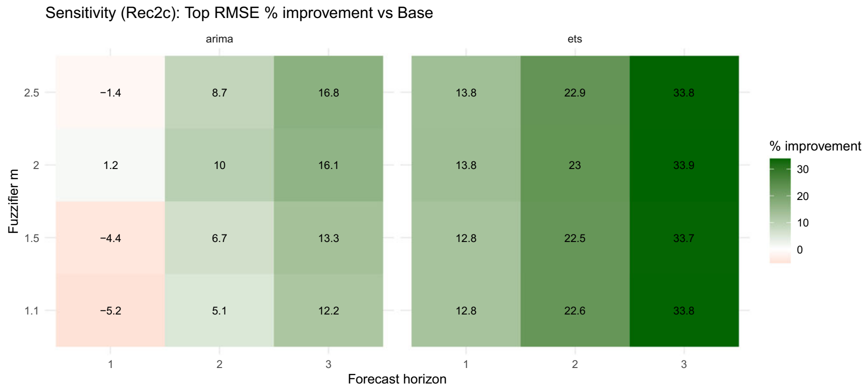
Overall, reconciliation effectiveness tends to increase with forecast horizon, indicating that cross-sectional constraints become more beneficial as forecast uncertainty grows. At the middle aggregation level, results remain broadly consistent across fuzzifier values. Both ARIMA and ETS baselines display similar patterns across horizons, and improvements (or mild deterioration) are largely insensitive to the exact value of m. This indicates that the clustering-induced aggregation layers provide stable information regardless of the degree of fuzziness used in constructing memberships. At the bottom level, improvements are generally modest and in several cases slightly negative, confirming that reconciliation primarily benefits higher aggregation levels, while gains at the most disaggregated level are limited. Importantly, however, the qualitative pattern remains stable across all tested values of m, showing that results do not hinge on a specific fuzzifier choice even when improvements are small.

6 Conclusions

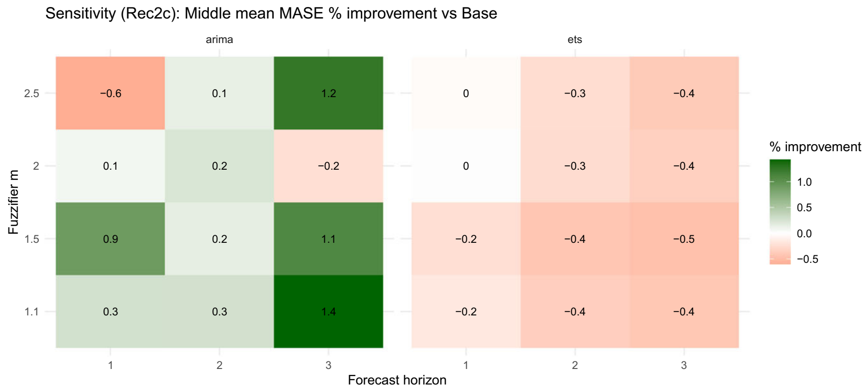
Temporal forecasting of environmental variables featuring the impact of human-related activities, such as pollutant emissions, are getting large attention from both private citizens, political institutions, and researchers. Despite data are nowadays largely available from public sources, they are often of limited length, especially when referred to territorial levels rather than aggregated at the national level. However, from the policy maker perspective, forecasting at the sub-national level might be a primary objective in order to cope with the obvious territorial heterogeneity and adopt locally-adapted mitigation strategies. In this context, the use of classical time series forecasting techniques which ignores the hierarchical structure at which the data are naturally recovered can be misleading and provide poorly reliable future predictions.

To address the above-mentioned issues, in this paper we propose a novel statistical approach which leverages statistical machine learning algorithms to forecast annual GHGs emissions at the territorial level. Specifically, we were interested in predicting emissions from the agricultural sector, which is a primary source of greenhouse gases, across the European regions. Specifically, we considered data on agricultural emissions for European regions (NUTS-2) nested within countries (NUTS-0) to predict the overall emissions for Europe as a whole. Our methodology enhances hierarchical forecasting reconciliation techniques with fuzzy clustering, which defines alternative aggregation structures of bottom and middle-level series into clusters, which are defined on the basis of the shape similarity of emissions patterns rather than solely on an administrative basis. In doing so, we also account for the inherent uncertainty behind this hierarchy construction step, by letting all the time series belong to all the clusters with a given degree of membership. We then combine the fuzzy hierarchies with the administrative aggregation structure.

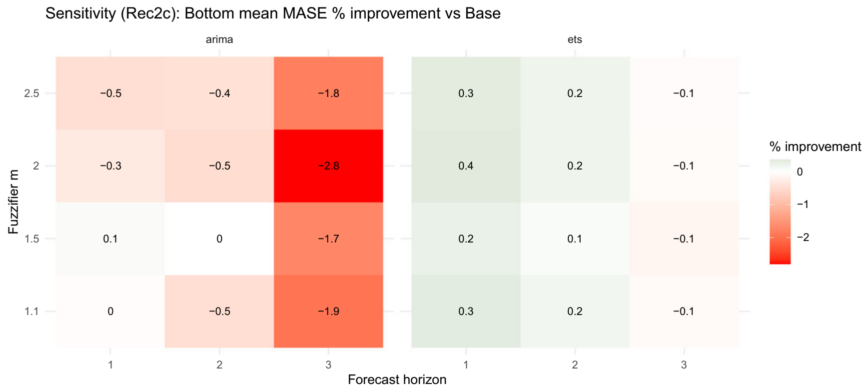
The empirical results demonstrate that the use of fuzzy clustering in forecast reconciliation leads to systematic gains over crisp approaches in terms of accuracy at any level of the



(a) Top level



(b) Middle level



(c) Bottom level

Fig. 5 Sensitivity analysis with respect to the fuzzifier parameter m and forecast horizon h. Panels **A–C** report percentage improvements of the MinT reconciliation with fuzzy hierarchy $\widehat{\boldsymbol{A}}_3$ relative to baseline forecasts at top, middle, and bottom aggregation levels, respectively hierarchy. Moreover, we also highlight that MinT aggregation is the most accurate and reliable reconciliation strategy in this setting and that OLS aggregation is not suitable for complex hierarchies due to its simplistic assumptions and poor scalability.

Future research should address some of the additional issues discussed in the manuscript, such as the presence of relevant spatial patterns in the time series of agricultural emissions. In particular, spatial patterns could be integrated in our approach within either the clustering step or the reconciliation step. For instance, regarding the former, spatial and spatio-temporal clustering techniques should be considered to group time series according to their trajectories and the geographical distance (e.g. Morelli et al., 2026; Brégère and Mattera, 2025); with respect to the latter, instead of computing binary matrices to reconcile the bottom-level forecasts, one could consider continuous weighting matrices based on the distance among units as in a spatial regression framework. Also, future research could extend the present analysis by jointly considering alternative clustering algorithms within the proposed fuzzy reconciliation framework, thus combining the methodological focus of this paper with the comparative perspective adopted in Mattera et al. (2024).

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement. Part of the authors were funded by the University of Milano-Bicocca (Italy) through the project "SCARFACE - Sequestering CARbon through Forests, AgriCulture, and land usE" (2024-ATEQC-0048).

Data Availability All results presented in this paper can be reproduced using the R statistical software. The codes were developed entirely by the authors. For reproducibility purposes, all the scripts and the data are made available for the public through the following dedicated GitHub repository https://github.com/raffmattera/fuzzyreco_GHG.git.

Declarations

Conflict of interest/Conflict of interest None

References

Agency, E. E. (2022). Air pollution section. Sources of air pollution in Europe. European Environmental Agency portal. https://www.eea.europa.eu/publications/air-quality-in-europe-2022/sources-and-emissions-of-air

Antanasijević, D., Pocajt, V., Ristić, M., & Perić-Grujić, A. (2015). Modeling of energy consumption and related ghg (greenhouse gas) intensity and emissions in Europe using general regression neural networks. *Energy, 84*, 816–824.

Ashouri, M., Hyndman, R. J., & Shmueli, G. (2022). Fast forecast reconciliation using linear models. *Journal of Computational and Graphical Statistics, 31*(1), 263–282.

Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. *International Journal of Forecasting, 39*(4), 1502–1511.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences, 10*(2–3), 191–203.

Boccaletti, S., Maranzano, P., & Viegas, M. (2024). Inequality and concentration in farmland production and size: Regional analysis for the European Union from 2010 to 2020. arXiv preprint arXiv:2409.00111

Borlee, F., Joris Yzermans, C., Aalders, B., Rooijackers, J., Krop, E., Maassen, C. B. M., Schellevis, F., Brunekreef, B., Heederik, D., & Smit, L. A. M. (2017). Air pollution from livestock farms is associated with airway obstruction in neighboring residents. *American Journal of Respiratory and Critical Care Medicine, 196*(9), 1152–1161. https://doi.org/10.1164/rccm.201701-0021OC

Brégère, M., & Mattera, R. (2025). Spatio-temporal clustering and reconciliation for regional electricity demand forecasting. hal-04853137.

Bucciarelli, G. (2024). ARDECO: Annual Regional Database of the European Commission (ARDECO). https://CRAN.R-project.org/package=ARDECO, r package version 2.1.0.

Campello, R. J., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems, 157*(21), 2858–2875.

Cerqueti, R., Maranzano, P., & Mattera, R. (2025). Spatially-clustered spatial autoregressive models with application to agricultural market concentration in Europe. *Journal of Agricultural, Biological, and Environmental Statistics*. https://doi.org/10.1007/s13253-024-00672-4

Crippa, M., Guizzardi, D., Solazzo, E., Muntean, M., Schaaf, E., Monforti-Ferrario, F., Banja, M., Olivier, J., Grassi, G., & Rossi, S. (2024). Ghg emissions of all world countries - 2024. *Report, Publications Office of the European Union, Luxembourg*. https://doi.org/10.2760/40028977,JRC138862. https://publications.jrc.ec.europa.eu/repository/handle/JRC138862

Dar, A. A., Chen, Z., Rodríguez-Rodríguez, S., Haghighat, F., & González-Rosales, B. (2024). Assessing greenhouse gas emissions in Cuban agricultural soils: Implications for climate change and rice (*Oryza sativa* l.) production. *Journal of Environmental Management*. https://doi.org/10.1016/j.jenvman.2024.120088

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

El Hathat, Z., Sreedharan, V. R., Venkatesh, V. G., Zouadi, T., Arunmozhi, M., & Shi, Y. (2023). Modelling and analyzing the ghg emissions in the vuca world: Evidence from tomato production in Morocco. *Journal of Cleaner Production*. https://doi.org/10.1016/j.jclepro.2022.134862

European Commission. (2016). NUTS – Nomenclature of territorial units for statistics. https://ec.europa.eu/eurostat/web/nuts

Fassò, A., Rodeschini, J., Moro, A. F., Shaboviq, Q., Maranzano, P., Cameletti, M., Finazzi, F., Golini, N., Ignaccolo, R., & Otto, P. (2023). Agrimonia: A dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data*. https://doi.org/10.1038/s41597-023-02034-0

Fernández-Avilés, G., Mattera, R., & Scepi, G. (2024). Factor-augmented autoregressive neural network to forecast nox in the city of madrid. *Socio-Economic Planning Sciences, 95*, Article 101958.

Girolimetto, D., & Di Fonzo, T. (2022). Foreco: Point forecast reconciliation R package v0 2

Girolimetto, D., & Di Fonzo, T. (2024). Point and probabilistic forecast reconciliation for general linearly constrained multiple time series. *Statistical Methods & Applications, 33*(2), 581–607.

Guarín, A., Rivera, M., Pinto-Correia, T., Guiomar, N., Šūmane, S., & Moreno-Pérez, O. M. (2020). A new typology of small farms in Europe. *Global Food Security*. https://doi.org/10.1016/j.gfs.2020.100389

Hamrani, A., Akbarzadeh, A., & Madramootoo, C. A. (2020). Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2020.140338

Hollyman, R., Petropoulos, F., & Tipping, M. E. (2021). Understanding forecast reconciliation. *European Journal of Operational Research, 294*(1), 149–160.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Hyndman, R. J. (2022) Notation for forecast reconciliation. Retrieved October 17, 2023, from https://robjhyndman.com/hyndsight/reconciliation-notation.html

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice. OTexts.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis, 55*(9), 2579–2589.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, 27*, 1–22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer.

Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis, 97*, 16–32.

Jiang, H., Yin, J., Wei, D., Luo, X., Ding, Y., & Xia, R. (2024). Industrial carbon emission efficiency prediction and carbon emission reduction strategies based on multi-objective particle swarm optimization-backpropagation: A perspective from regional clustering. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2023.167692

Kalkowska, D. A., Boender, G. J., Smit, L. A., Baliatsas, C., Yzermans, J., Heederik, D. J., & Hagenaars, T. J. (2018). Associations between pneumonia and residential distance to livestock farms over a five-year period in a large population-based study. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0200813

Kannan, D., Pati, R. K., Padhi, S. S., & Bai, C. (2025). Policy analysis in agrochemical supply chain: A system dynamics approach. *Annals of Operations Research, 344*(1), 533–561. https://doi.org/10.1007/s10479-024-06113-2

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88*(2), 365–411.

Li, H., Li, H., Lu, Y., & Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics and Economics, 86*, 122–133.

Lovarelli, D., Conti, C., Finzi, A., Bacenetti, J., & Guarino, M. (2020). Describing the trend of ammonia, particulate matter and nitrogen oxides: The role of livestock activities in northern Italy during Covid-19 quarantine. *Environmental Research*. https://doi.org/10.1016/j.envres.2020.110048

Lunghi, J., Malpede, M., & Aleluia Reis, L. (2024). Exploring the impact of livestock on air quality: A deep dive into ammonia and particulate matter in Lombardy. *Environmental Impact Assessment Review*. https://doi.org/10.1016/j.eiar.2024.107456

Maharaj, E. A., D'Urso, P., & Caiado, J. (2019). *Time series clustering and classification*. Chapman and Hall/CRC.

Mattera, R., Athanasopoulos, G., & Hyndman, R. (2024). Improving out-of-sample forecasts of stock price indexes with forecast reconciliation and clustering. *Quantitative Finance, 24*(11), 1641–1667.

Moiceanu, G., & Dinca, M. N. (2021). Climate change-greenhouse gas emissions analysis and forecast in Romania. *Sustainability, 13*(21), 12186.

Morelli, C., Maranzano, P., & Otto, P. (2026). Spatiotemporal clustering of ghgs emissions in Europe: Exploring the role of spatial component. *Spatial Statistics, 73*, Article 100960. https://doi.org/10.1016/j.spasta.2026.100960

Nenes, A., Pandis, S. N., Weber, R. J., & Russell, A. (2020). Aerosol ph and liquid water content determine when particulate matter is sensitive to ammonia and nitrate availability. *Atmospheric Chemistry and Physics, 20*(5), 3249–3258. https://doi.org/10.5194/acp-20-3249-2020

Otto, P., Fusta Moro, A., Rodeschini, J., Shaboviq, Q., Ignaccolo, R., Golini, N., Cameletti, M., Maranzano, P., Finazzi, F., & Fassò, A. (2024). Spatiotemporal modelling of PM 2.5 concentrations in Lombardy (Italy): A comparative study. *Environmental and Ecological Statistics, 31*(2), 245–272. https://doi.org/10.1007/s10651-023-00589-0

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting, 37*(1), 343–359.

Panagiotelis, A., Gamakumara, P., Athanasopoulos, G., & Hyndman, R. J. (2023). Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research, 306*(2), 693–706.

Pang, Y., Zhou, X., Zhang, J., Sun, Q., & Zheng, J. (2022). Hierarchical electricity time series prediction with cluster analysis and sparse penalty. *Pattern Recognition, 126*, Article 108555.

Peter, C., Helming, K., & Nendel, C. (2017). Do greenhouse gas emission calculations from energy crop cultivation reflect actual agricultural management practices? - a review of carbon footprint calculators. *Renewable and Sustainable Energy Reviews, 67*, 461–476. https://doi.org/10.1016/j.rser.2016.09.059

Petersen, S. O. (2018). Greenhouse gas emissions from liquid dairy manure: Prediction and mitigation. *Journal of Dairy Science, 101*(7), 6642–6654. https://doi.org/10.3168/jds.2017-13301

Pimentel, D. (1996). Green revolution agriculture and chemical hazards. *Science of the Total Environment, 188*(SUPPL. 1), S86–S98. https://doi.org/10.1016/0048-9697(96)05280-1

Porter, S. D., & Reay, D. S. (2016). Addressing food supply chain and consumption inefficiencies: potential for climate change mitigation. *Regional Environmental Change, 16*(8), 2279–2290. https://doi.org/10.1007/s10113-015-0783-4

Richards, M., Metzel, R., Chirinda, N., Ly, P., Nyamadzawo, G., Duong, V., & Q., De Neergaard, A., Oelofse, M., Wollenberg, E., Keller, E., Malin, D., Olesen, J. E., Hillier, J., Rosenstock, T. S. (2016). Limits of agricultural greenhouse gas calculators to predict soil n2o and ch4 fluxes in tropical agriculture. *Scientific Reports*. https://doi.org/10.1038/srep26279

Shine, P., Upton, J., Sefeedpari, P., & Murphy, M. D. (2020). Energy consumption on dairy farms: A review of monitoring, prediction modelling, and analyses. *Energies*. https://doi.org/10.3390/en13051288

Thunis, P., Clappier, A., Beekmann, M., Putaud, J. P., Cuvelier, C., Madrazo, J., & de Meij, A. (2021). Non-linear response of PM2.5 to changes in NOx and NH3 emissions in the Po basin (Italy): Consequences for air quality plans. *Atmospheric Chemistry and Physics, 21*(12), 9309–9327. https://doi.org/10.5194/acp-21-9309-2021

Wang, X., Ji, X., Xu, Y. J., Mao, B., Jia, S., Wang, C., Liu, Z., & Lv, Q. (2024). Multi-machine learning methods to predict spatial variation characteristics of total nitrogen at watershed scale: Evidences from the largest watershed (Yangtze River Watershed), Asian. *Science of the Total Environment*. https://doi.org/10.1016/j.scitotenv.2024.175144

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association, 114*(526), 804–819.

Zhang, B., Panagiotelis, A., & Li, H. (2024). Constructing hierarchical time series through clustering: Is there an optimal way for forecasting? *International Journal of Forecasting, 41*(3), 1022–1036.

Zimmermann, A., Heckelei, T., & Domínguez, I. P. (2009). Modelling farm structural change for integrated ex-ante assessment: Review of methods and determinants. *Environmental Science and Policy, 12*(5), 601–618. https://doi.org/10.1016/j.envsci.2009.01.014