University of Helsinki
Dissertationes Universitatis Helsingiensis
429/2026

Learning from past mistakes: Uncertainty in demographic forecasting

Ricarda Duerst

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Social Sciences of the University of Helsinki in lecture hall Tekla Hultin, University Main Building, on 25 September 2026, at 1 p.m.

Helsinki 2026

Academic Advisors

Professor **Mikko Myrskylä**, Ph.D.
Institute for Demography and Population Health
University of Helsinki
Max Planck Institute for Demographic Research

Jonas Schöley, Ph.D.
Max Planck Institute for Demographic Research

Pre-Examiners

Associate Professor **Marie-Pier Bergeron Boucher**, Ph.D.
Interdisciplinary Centre on Population Dynamics
University of Southern Denmark

Associate Professor **Jason Hilton**, Ph.D.
Centre for Population Change
University of Southampton

Custos

Professor **Mikko Myrskylä**, Ph.D.
Institute for Demography and Population Health
University of Helsinki

Opponent

Carlo-Giovanni Camarda, Ph.D.
Institut national d'études démographiques

Publisher: University of Helsinki
Series: Dissertationes Universitatis Helsingiensis 429/2026
Cover Image: Ricarda Duerst
ISBN 978-952-84-2975-3 (paperback)
ISBN 978-952-84-2974-6 (PDF)
ISSN 2954-2898 (print)
ISSN 2954-2952 (online)
PunaMusta, Joensuu 2026

Contents

Acknowledgements i
Abstract ii
Tiivistelmä iii
Zusammenfassung iv
List of original publications vi
Report on author contributions vii

1 Introduction **1**

2 Data and Methods **5**
2.1 Human Mortality Database 5
2.2 Human Fertility Database 7
2.3 Learning from past mistakes 8
2.3.1 Empirical prediction intervals 8
2.3.2 Calibration, validation, and application 10
2.3.3 Quantile-mapping 13
2.3.4 Producing the central forecasts 13

3 Results **17**
3.1 Paper 1: The contribution of forecast uncertainty to lifespan uncertainty . 17
3.2 Paper 2: Empirical prediction intervals applied to short-term mortality forecasts and excess deaths 18
3.3 Paper 3: Calibrating probabilistic forecast paths on past forecast errors: An application to the Finnish total fertility rate 20

4 Discussion **23**

5 Conclusion **25**

Acknowledgements

This thesis was conducted at the Max Planck Institute for Demographic Research (MPIDR) and in affiliation with the University of Helsinki (UH). To start with, I would like to thank my supervisors at the MPIDR. Great thanks go to Mikko Myrskylä for inviting me to join the MPIDR as a PhD student and for his pragmatic support throughout my PhD journey. Then, I would like to express my heartfelt thanks to my supervisor Jonas Schöley, without whom this thesis would not have come together. Thank you for your optimism, your understanding, your support, your interest and expertise, and for every lunch we spent together. I have been lucky to have had you as my supervisor. Further, I would like to thank my previous supervisor Christina Bohk-Ewald for introducing me to the world of demographic forecasting and method validation, and for her ever understanding and encouraging way of supervision.

I would also like to thank my colleagues at the Institute for Demography & Population Health and at the Centre for Social Data Science at the University of Helsinki. You have made my research stay and all other visits to Helsinki a warm and joyful experience. Also, thanks go to all of my colleagues at the MPIDR for your feedback, help, and expertise.

I have also been fortunate to participate as a core student in the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). I would like to thank the administrative team, especially Heiner Maier and Christina Westphal, for making the program possible. I also wish to thank all the student and faculty members I have had the privilege to meet during our seminars.

Special thanks go to the friends I made along the way. Thank you Lauren, Josephine, Liann, Kelsey, Joonas, Julia, Liina, Rok, Peter, Steffen, Angela, and Philipp, the best roommate in the world, for all the shared sorrows and joys. It means a lot to me to be able to call you my friends. Further, I wish to thank the whole PhD student community at the MPIDR for being a source of endless creativity, fun activities and communal support.

Finally, I would like to thank all of my family members and friends for being there for me. Thank you for listening, for asking about my progress, and for not asking about my progress. You gave me the security I needed to persevere. Last, but certainly not least, heartfelt thanks go to my dear Henning, who exhausted his repertoire of motivational techniques on me. This would not have been possible without your love, trust and support.

Seeing this thesis completed makes me happy, proud, and relieved. This would not have been possible without the help, mentorship, and support of all the people mentioned above. I would like to close with a poem by Rabindranath Tagore that best describes my feelings about seeing this thesis finished:

It dances today, my heart,
like a peacock it dances,
it dances.
It sports a mosaic of passions like a peacock's tail,
It soars to the sky with delight, it quests,
Oh wildly, it dances today, my heart,
like a peacock it dances.

Abstract

Demographic forecasts are predictions about future population size and structure, and the processes that shape these. Communicating the uncertainty around these predictions is essential because the forecasts are used by decision-makers to inform policies and regulations that affect society. Therefore, probabilistic forecasts that assign a probability to each possible outcome are useful, because they allow for an assessment of the likely future while, at the same time, enabling us to prepare for the unlikely but impactful outcomes. Thus, improving the estimation of forecast uncertainty is important.

The common practice for estimating forecast uncertainty is to employ a model-based estimation that communicates the statistical uncertainty of the forecast model parameters and the outcome. However, other non-model-related sources of uncertainty are not covered. The analysis of the empirical forecast error, which is the error we retrieve by comparing historical data to our forecast for that period, can provide meaningful insights for the estimation of forecast uncertainty. By using the information on past forecast performance, we can inform the uncertainty of the forecast for the future based on the notion that a forecast is only as good as similar forecasts in the past turned out to be.

This thesis highlights the advantages of forecast error analysis and empirical prediction intervals, such as broad applicability and empirical nature, for use in demographic forecasting to improve probabilistic forecasts. Further, I showcase the application of this methodology for mortality and fertility forecasting in three research articles that make up this thesis. In addition, by combining existing techniques, I extend the methodology of empirical forecast uncertainty estimation.

The analyses and results of the research articles are based on open-access, population-level data of high quality and are fully reproducible. The first research article disentangles the different sources of uncertainty in the context of individual lifespan uncertainty and cohort mortality forecasting. We quantify the contribution of the forecast uncertainty to the total lifespan uncertainty of birth cohorts over time via a variance decomposition. Employing an empirical forecast error for the estimation of the forecast uncertainty and comparing it to model-based uncertainty allows us to derive results for two possible scenarios. On the one hand, we consider a future in which mortality develops without extreme shocks, and on the other hand, a future with the possibility of drastic mortality events.

In the second study, forecast error analysis is the key methodology that reveals seasonal patterns in the uncertainty of expected deaths forecasts used for the estimation of excess mortality during the COVID-19 pandemic. We derive empirical prediction intervals from the forecast errors to improve the expected deaths forecasts.

The third research article combines the methodology of empirical prediction intervals and quantile-mapping to provide forecasts of the Finnish total fertility rate in the form of individually simulated forecast paths that have been calibrated to empirical forecast errors.

In conclusion, this thesis showcases how the use of forecast error analysis and empirical prediction intervals improves demographic forecasts while adding to the existing methodology.

Tiivistelmä

Väestöennusteet ovat arvioita tulevasta väestömäärästä ja -rakenteesta sekä niitä muovaavista prosesseista. Ennusteisiin liittyvän epävarmuuden ilmaiseminen on olennaista, koska päätöksentekijät käyttävät niitä yhteiskuntaan vaikuttavien politiikkatoimien ja sääntelyn päätösten tukena. Siksi todennäköisyyspohjaiset ennusteet, jotka liittävät jokaiseen mahdolliseen kehityspolkuun todennäköisyyden, ovat hyödyllisiä: ne mahdollistavat todennäköisen tulevaisuuden arvioinnin ja samalla auttavat varautumaan epätodennäköisiin mutta vaikutuksiltaan merkittäviin tulemiin. Siksi ennusteisiin liittyvän epävarmuuden määrittämisen parantaminen on tärkeää.

Yleinen käytäntö ennuste-epävarmuuden arvioimisessa on mallipohjainen lähestymistapa, joka kuvaa ennustemallin parametrien tilastollista epävarmuutta. Tämä ei kuitenkaan kata mallista riippumattomia epävarmuuden lähteitä. Empiirisen ennustevirheen analyysi - eli historiallista dataa ja sitä vastaavaa ennustetta vertaamalla saatu virhe – voi tarjota olennaista tietoa ennuste-epävarmuuden arviointiin. Aiempien ennusteiden toteutuneen tarkkuuden perusteella voidaan arvioida tulevien ennusteiden epävarmuutta. Lähtökohtana on se, että ennuste on vain niin hyvä kuin vastaavat ennusteet ovat aiemmin olleet. Väitöskirja korostaa ennustevirheanalyysin ja empiiristen ennustevälien etuja, erityisesti niiden laajaa sovellettavuutta ja empiiristä luonnetta, sekä niiden käyttöä probabilististen väestöennusteiden parantamiseksi. Lisäksi esittelen menetelmien soveltamista kuolleisuus- ja hedelmällisyysennusteisiin kolmessa väitöskirjaan kuuluvassa osatutkimuksessa. Laajennan empiirisen ennuste-epävarmuuden estimointimenetelmää hyödyntämällä ja yhdistämällä olemassa olevia lähestymistapoja.

Osatutkimuksien analyysit ja tulokset perustuvat avoimesti saatavilla olevaan, korkealaatuiseen väestötason aineistoon ja ovat täysin toistettavissa. Ensimmäisessä osatutkimuksessa eritellään epävarmuuden eri lähteitä yksilön eliniän epävarmuuden sekä kohorttikuolleisuuden ennustamisen kontekstissa. Arvioimme ennuste-epävarmuuden osuutta syntymäkohorttien kokonaiseliniän epävarmuudessan ajan kuluessa varianssihajotelman (eng. variance decomposition) avulla. Hyödyntämällä empiiristä ennustevirhettä ennuste-epävarmuuden estimoinnissa ja vertaamalla sitä mallipohjaiseen epävarmuuteen voimme tarkastella kahta vaihtoehtoista skenaariota: tulevaisuuden ilman äärimmäisiä kuolleisuushäiriöitä/-shokkeja sekä tulevaisuuden, jossa voimakkaat kuolleisuuden muutokset ovat mahdollisia.

Toisessa osatutkimuksessa ennustevirheanalyysi toimii keskeisenä menetelmänä. Sen avulla tunnistetaan kausivaihtelua COVID-19-pandemian aikaisen ylikuolleisuuden arvioinnissa käytettyjen odotettujen kuolemien ennusteiden epävarmuudessa. Ennustevirheiden perusteella muodostetaan empiiriset ennustevälit, joiden avulla odotettujen kuolemien ennusteita parannetaan. Kolmannessa osatutkimuksessa yhdistetään empiiristen ennustevälitien menetelmä ja kvantiilikartoitus (eng. quantile-mapping). Näitä hyödyntämällä tuotetaan Suomen kokonaishedelmällisyysluvun ennusteita simuloituina ennustepolkuina. Ennustepolut on kalibroitu empiiristen ennustevirheiden perusteella.

Tämä väitöskirja osoittaa, kuinka ennustevirheanalyysin ja empiiristen ennustevälien käyttö parantaa väestöennusteita ja samalla kehittää olemassa olevaa ennustemetodologiaa.

Zusammenfassung

Demographische Prognosen sind Vorhersagen über die zukünftige Größe und Struktur von Bevölkerungen sowie über die Prozesse, die diese prägen. Die Vermittlung der mit diesen Vorhersagen verbundenen Unsicherheit ist von entscheidender Bedeutung, da demographische Prognosen von Entscheidungsträgern als Grundlage für politische Maßnahmen und Regulierungen herangezogen werden. Diese haben Auswirkungen auf die Gesellschaft. Probabilistische Prognosen, die jedem möglichen Ergebnis eine Wahrscheinlichkeit zuweisen, bieten hier den Vorteil, Einschätzungen über die wahrscheinliche Zukunft vorzunehmen. Gleichzeitig erlauben sie es, sich auf unwahrscheinliche, aber folgenreiche Entwicklungen vorzubereiten. Daher ist es wichtig, die Schätzung der Prognoseunsicherheit zu verbessern.

Die gängige Praxis zur Schätzung der Prognoseunsicherheit besteht darin, eine modellbasierte Schätzung zu verwenden, die die statistische Unsicherheit der Parameter und des Resultats des Prognosemodells wiedergibt. Andere, nicht modellbezogene Unsicherheitsquellen werden dabei jedoch nicht berücksichtigt. Die Analyse des empirischen Prognosefehlers, also des Fehlers, den wir ermitteln, indem wir historische Daten zurückhalten und mit unserer Prognose für diesen Zeitraum vergleichen, kann dabei aussagekräftige Erkenntnisse liefern. Durch die Nutzung der Informationen über die vergangene Prognosegüte kann die Unsicherheit der Prognose für die Zukunft geschätzt werden, basierend auf der Idee, dass eine Prognose nur so gut ist, wie sich ähnliche Prognosen in der Vergangenheit bewährt haben.

Diese Arbeit beleuchtet die Vorteile der Analyse von Prognosefehlern und empirischer Vorhersageintervalle, wie beispielsweise ihre breite Anwendbarkeit und ihren empirischen Charakter, für den Einsatz in demographischen Prognosen zur Verbesserung probabilistischer Vorhersagen. Darüber hinaus wende ich in den drei Forschungsartikeln, aus denen diese Arbeit besteht, diese Methoden auf die Prognose von Sterblichkeit und Fertilität an. Durch die Kombination bestehender Verfahren erweitere ich zudem die Methodik zur empirischen Schätzung von Prognoseunsicherheit.

Die Analysen und Ergebnisse der Forschungsartikel basieren auf frei zugänglichen, qualitativ hochwertigen Bevölkerungsdaten und sind vollständig reproduzierbar. Der erste Forschungsartikel entwirrt die verschiedenen Unsicherheitsquellen im Zusammenhang mit der Unsicherheit in der individuellen Lebenserwartung und der Prognose von Kohortensterblichkeit. Mittels einer Varianzzerlegung quantifizieren wir den Beitrag der Prognoseunsicherheit zur gesamten Unsicherheit über die Lebenserwartung von Geburtskohorten im Zeitverlauf. Durch die Verwendung eines empirischen Prognosefehlers zur Schätzung der Prognoseunsicherheit und dessen Vergleich mit der modellbasierten Unsicherheit können wir Ergebnisse für zwei mögliche Szenarien ableiten. Einerseits betrachten wir eine Zukunft, in der sich die Mortalität ohne extreme Schocks entwickelt, und andererseits eine Zukunft, in der drastische Mortalitätsereignisse möglich sind.

In der zweiten Studie ist die Analyse von Prognosefehlern die zentrale Methode, mit der saisonale Muster in der Unsicherheit über Prognosen zu erwarteten Todesfällen aufgezeigt werden, die zur Schätzung der Übersterblichkeit während der COVID-19-Pandemie herangezogen wurden. Aus den Prognosefehlern leiten wir empirische Vorher-

sageintervalle ab, um die Prognosen zu verbessern.

Der dritte Forschungsartikel kombiniert die Methodik empirischer Vorhersageintervalle mit Quantil-Mapping, um Prognosen zur finnischen zusammengefassten Geburtenziffer in Form individuell simulierter Prognosepfade zu erstellen, die anhand empirischer Prognosefehler kalibriert wurden.

Zusammenfassend zeigt diese Arbeit, wie der Einsatz von Prognosefehleranalysen und empirischen Vorhersageintervallen demographische Prognosen verbessert und erweitert gleichzeitig die bestehende Methodik.

List of original publications

This thesis is based upon the following publications:

- Paper 1: Duerst, R. & Schöley, J. (2026). *The contribution of forecast uncertainty to lifespan uncertainty.* Demographic Research 55, 431-450.
- Paper 2: Duerst, R. & Schöley, J. (2024). *Empirical prediction intervals applied to short term mortality forecasts and excess deaths.* Population Health Metrics 22:34.
- Paper 3: Duerst, R., Hellstrand, J., Schöley, J. & Myrskylä, M. *Calibrating probabilistic forecast paths on past forecast errors: An application to the Finnish total fertility rate.* Submitted to the Journal of the Royal Statistical Society - Series A.

Report on author contributions

The papers that come together to form this thesis are the result of the collaborative work with my supervisors Jonas Schöley (JS) and Mikko Myrskylä (MM), and my co-author Julia Hellstrand (JH). I am the leading first author on each of the three papers and was solely or together with my co-authors responsible for the conceptualization of the papers, drafting of the first version of each paper, finalizing each paper, and communication during the publication process. Further, I presented the research findings at several international conferences. The following report summarizes each author's contribution to the final papers.

Paper 1: The contribution of forecast uncertainty to lifespan uncertainty

Conceptualization: I conceptualized the research idea together with JS, who supervised this project.

Methodology and analysis: I was responsible for methodology, software and analysis, together with JS. I have performed the formal analyses with support from JS.

Visualization of results: I have created the visualizations in the manuscript with support from JS.

Preparation and writing of the manuscript: I have written the initial draft of the manuscript. JS has commented on all versions of the manuscript. MM has read and commented on an early version of the draft. I have edited the manuscript according to my supervisor's comments.

Paper 2: Empirical prediction intervals applied to short-term mortality forecasts and excess deaths

Conceptualization: The research idea was developed by JS in response to a question that was raised by journalists during the COVID-19 pandemic. I conceptualized the research idea together with JS, who supervised this project.

Methodology and analysis: I was responsible for methodology, software and analysis, together with JS. I have performed the formal analyses with support from JS.

Visualization of results: I have created the visualizations in the manuscript with support from JS.

Preparation and writing of the manuscript: I have written the initial draft of the manuscript. JS has commented on all parts of this draft and I have edited it accordingly.

Paper 3: Calibrating probabilistic forecast paths on past forecast errors: An application to the Finnish total fertility rate

Conceptualization: The research idea was developed by me, JS and MM after MM was approached by the Finnish Centre for Pensions that requested forecasts of Finnish fertility. JS and MM have supervised this project. I was responsible for the conceptualization to which JS, MM and JH contributed.

Methodology and analysis: I was responsible for methodology and software development to which JS and JH contributed, and performed the formal analyses with contributions from JS.

Visualization of results: I have created the visualizations in the manuscript.

Preparation and writing of the manuscript: I have written the initial draft of the manuscript. JS has commented on all parts of this draft. MM and JH have commented on the draft. I have edited the manuscript according to my supervisor's and co-author's comments.

Chapter 1

Introduction

> I just wonder if current forecasters lack in intellectual curiosity or if they are intentionally ignoring forecast errors.
>
> *Nassim Nicholas Taleb*
> *The Black Swan*

Why do we make mistakes? So we can learn from them. It is a simple idiom, but that does not mean that there is no truth to be found, both for our individual lives and for science. In this thesis, I propose to learn from our past mistakes by taking a specific error into account: The forecast error. This is the error that we observe in hindsight when comparing our prediction for the future with the actual outcome of what we wanted to predict. By paying attention to this forecast error, we can learn about the uncertainty of our forecast. Specifically, I focus on the uncertainty in demographic forecasting and how the analysis of forecast errors and the resulting insights on forecast uncertainty can improve the forecasts themselves. This thesis is a methodological one. I am more concerned with *how* and a little less with *what* to forecast. Nevertheless, the results of the methodological applications presented in the papers that make up this thesis are relevant. However, my scientific curiosity lies in the methodology used to obtain these results.

Demographic forecasts are predictions about future populations. Depending on the subject of study, these forecasts can concern a multitude of outcomes: From population size, broken down into subgroups by, for example, sex, age, or educational attainment, to the processes that shape a population, such as births, deaths, or migration streams. Similarly, the forecast horizon – that is, the length of the period for which the forecast is made – may range from days and weeks to years and centuries. Demographic forecasting has a long lasting tradition and the scientific community is making advances in methodology up to this day. The demand for forecasts of population size and structure, as well as the processes shaping these, drives the interest in this field. The purpose behind demographic forecasts and the explanation for their demand is to have a better chance at planning for

the future. Thus, the forecasts do not exist in isolation but have real life applications. The results of demographic forecasts are used as input for further modeling for the social security systems and as justification for policy decisions that aim to plan for the future that is to come. For example, setting the retirement age of a country's population is not an arbitrary choice. Careful planning should be considered in light of the life-changing consequences that such a decision has for an entire population and their economy. Other implications are showcased by the rather recent occurrence of the COVID-19 pandemic. The reader might remember how planning for resources such as respiratory ventilators and intensive care units, or even hospital beds and medical staff was dependent on estimations of the future development of case numbers and deaths.

Communicating the uncertainty that comes with demographic forecasts is essential, in light of the consequences that decisions based on these forecasts have for society, the economy, social security systems, etc. Providing forecast results without, or with false or misleading measures of certainty attached to them is, at the very least, bad scientific practice. At worst, it can have major consequences and be unethical.

Therefore, when forecasting, looking at the mistakes we have made offers a chance to improve current forecasts. However, to do so, we cannot wait 50 years to finally see whether the 50 year forecast we made was actually correct. Luckily, we can apply a trick: Withholding historical data to mimic the forecast experience one would have had in the past opens the route to what is called empirical error analysis. Early works that sparked interest in forecast error analysis and its use for constructing uncertainty intervals around forecasts were published by Williams and Goodman (1971), Stoto (1983), Cohen (1986), and Smith and Sincich (1988). The reader will learn the details of the methodology used to retrieve uncertainty intervals for forecasts from empirical forecast errors in the Methods section of this thesis. The underlying principle of these empirical prediction intervals is that a forecast is only as good as previous similar forecasts have turned out to be. This assumption enables us to use the empirical forecast errors to inform the uncertainty of the actual forecast. A key feature of the empirical uncertainty lies in its outcome-based focus. In contrast to conventional, parameter-based uncertainty, empirical uncertainty is able to capture all sources of uncertainty. While parameter-based uncertainty is limited to sources of uncertainty that are included in the model, empirical uncertainty allows us to include the uncertainty from non-model-related sources, too. In other words, the uncertainty from all things that can affect the forecast outcome of interest, regardless of their inclusion in the forecast model, is captured in the empirical uncertainty without the need for extensive modeling and high data demands.

This thesis consists of three articles that all deal with demographic forecasting and employ forecast error analysis. With each paper, the use of this methodology becomes more elaborate and shapes the outcomes of the studies more. The first paper disentangles the different sources of uncertainty in the context of individual lifespan uncertainty and cohort mortality forecasting. Next to life expectancy, lifespan variance or lifespan variability is a measure used to describe the mortality experience of a population. It captures the inequality in the lifespans of the members of a population. Smaller lifespan variability means less inequality in the length of lives – that is, more similar ages at death. In the demographic literature, the population-level mortality measure lifespan variance is

sometimes interpreted as an indicator of the uncertainty an individual faces about their length of life (Nepomuceno et al., 2022; Van Raalte et al., 2018; Sasson, 2016; Edwards, 2013; Edwards and Tuljapurkar, 2005). In paper 1, we ask whether this is an admissible interpretation. We argue that this individual lifespan uncertainty is a function of the variability in the ages at death in a cohort life table and the forecast uncertainty coming from the completion of cohort lifetables. To answer the question, we quantify the contribution of the forecast uncertainty to the total lifespan uncertainty of birth cohorts over time via a variance decomposition as introduced by Caswell (2023). A small contribution of forecast uncertainty to the overall uncertainty would support the interpretation of lifespan variance as individual lifespan uncertainty. In that case, it would be the cohort lifetable lifespan variance – that is, the differences in the lifespans between the cohort members – that mainly determines the individual's uncertainty about their time of death, and not the uncertainty that comes from not knowing which death rates apply to them in the future – that is, from completing their cohort lifetable via forecasting. Using a variance decomposition, we assess the importance of both components of lifespan uncertainty for cohorts born from 1901 to 2019.

In paper 2, forecast error analysis is the key methodology that reveals uncertainties in expected deaths forecasts used for estimations of excess mortality during the COVID-19 pandemic. Excess deaths are the deaths that would have occurred if a specific event, in this case the COVID-19 pandemic, did not happen. Estimations of excess deaths were used to assess the mortality impact or mortality burden of the pandemic. We derive empirical prediction intervals from the forecast errors to improve the expected deaths forecasts. We showcase how the forecast uncertainty can be calibrated by including the empirical knowledge gained from the forecast error analysis, resulting in better estimates of excess deaths with more realistic uncertainty.

In the third paper, the forecast error analysis and the empirical prediction intervals derived from it merge together with the forecast model itself. The uncertainty of forecasts of the Finnish total fertility rate is calibrated to the historical errors and then, by combining existing methodology, the forecast results become calibrated probabilistic forecast paths. Probabilistic forecasts, in contrast to deterministic or point forecasts, account for uncertainty by providing a range of possible outcomes and attaching a probability to each of them. Commonly, probabilistic forecasts are provided in the form of prediction intervals. However, we propose and showcase the construction of forecast paths whose full distribution has been calibrated using empirical forecast errors.

This thesis brings the analysis of forecast errors back into the consciousness of forecasters and those doing research on forecasting. Further, I showcase how the use of empirical prediction intervals improves demographic forecasts and highlight that the value of a forecast lies in its careful calibration of the uncertainty to increase the usefulness of the forecast results in light of the consequences of its use for decision-making. Last but not least, in this thesis, I advocate for the use of forecast error analysis to improve forecasting. Due to its empirical nature, the implementation of forecast error analysis is applicable to all sorts of forecast models and outcomes, demographic or not.

Because this is a cumulative thesis, the introductory chapter serves as an overview and preparation for the reader. First, I will introduce the main data and methods used in the

papers in more natural language. The technical details and formalizations can be found in the respective papers and their appendices in the following chapters. After summarizing the main results of the papers, a quick discussion and conclusion mark the end of the introductory chapter. Following this, the reader may delve into the heart of this thesis by reading the research articles.

Chapter 2

Data and Methods

Data quality and coverage are fundamental to producing any forecast; and so is the choice of methods. In the following, I will first introduce the main data sources and datasets used in the papers of this thesis. Second, the reader may familiarize themselves with those methods that build the core of the analyses. For a detailed and formal description of the methods, see the methods sections and appendices of the respective papers.

All data used are available free of charge from the respective websites. The only necessary step before downloading the data is to create an account with an email address. Further, in support of the open science idea to enable efficient research that is transparent, accessible, and reproducible (see UNESCO, 2022, for an introduction to open science), the data I have created with my research – that is, the codebase written for the analyses and the published research articles – are also available free of charge online (see the data availability statements of the respective papers for links to the code repositories).

2.1 Human Mortality Database

The mortality forecasts presented in this thesis are based on data from the Human Mortality Database (HMD). This joint effort of the Max Planck Institute for Demographic Research (Germany), the University of California, Berkeley (USA), and the French Institute for Demographic Studies (France) provides high-quality data on death counts, population exposures, mortality rates, and other metrics of human longevity by different categories of age, period, cohort, and sex. The database covers 56 highly industrialized countries around the world.

The quality of the HMD data lies in its careful consideration and preparation. Thus, only countries whose death registration and census data "cover the entire population and that are very nearly complete" (Wilmoth et al., 2021, p. 4) are included in the database.

The HMD provides raw data as well as estimates of the different mortality measures. The raw data is taken directly from the original data sources such as national censuses and death registrations from the national statistical offices and is only cleaned of obvious

processing errors. The data cleaning is applied to deaths, births and population sizes that are used as input data for the calculation of the other provided datasets. These include, among others, population exposures (i.e., estimates of the population exposed to the risk of death), death rates and life tables. Some of the estimation methodology applied to the raw data includes the splitting of aggregated data into single-age categories, the distributing of deaths with unknown age across the age at death distribution, and the estimation of population sizes for ages above 80. All these methods are applied to harmonize the data so that analyses comparing countries or pooling data across countries are possible. See the methods protocol of the HMD (Wilmoth et al., 2021) for an in-depth look at the methodology used to achieve data conformity.

For the analyses in Paper 1, death counts and population exposures were sourced for the male population of Denmark, England and Wales, France, Iceland, the Netherlands, Norway, and Sweden for the years 1871 to 2019. These countries have the longest time series available, which qualifies them for the forecast error analysis.

For England and Wales, the HMD differentiates between the "civilian" and the "total" population. The two datasets differ for the years 1912 to 1920 and 1939 to 1950 during which the civilian dataset excludes military deaths. For our forecast analyses, we want the entire population to be reflected. Excluding military deaths means excluding a part of the population whose mortality is exceptionally vulnerable to conflicts and wars. Wars are a major source of forecast uncertainty in two ways. On the one hand, including periods of elevated mortality as input data for forecasting time series of mortality can bias the forecast results compared to a forecast without the inclusion of excess deaths due to wars. On the other hand, when empirical prediction intervals are applied, the inclusion of the forecast errors caused by this bias is desirable for estimating the future forecast uncertainty. Mortality forecast models have a hard time forecasting unforeseen events (unforeseen at least for the time-series model), such as wars or pandemics that affect the mortality. Further, we desire comparability between the countries in our analyses and the majority of country datasets in the HMD do not differentiate between the civilian and the total population.

The higher vulnerability to conflicts and wars is the reason we chose the male population for our analysis of lifespan uncertainty in Paper 1. Not only is the life expectancy of males generally lower compared to females. The variance in the age at death of males is also higher. This gives us more room in the form of higher lifespan uncertainty to perform our variance decomposition. In light of higher mortality due to extreme events and its effects on forecast uncertainty, we have decided to take a two-sided approach: On the one hand, we perform the variance decomposition for a scenario in which we do not expect extreme mortality due to wars or epidemics. We do so by excluding the war-affected years from the mortality forecasting procedure to complete the cohort mortality lifetables. On the other hand, in a second scenario, we do allow for the possibility of extreme mortality by using empirical forecast error quantification to estimate the forecast variance.

For Paper 2, we use another data source provided by the HMD. The Short-Term Mortality Fluctuations data series (STMF) is a sub-dataset that was created in response to the data needs of scientists and the public that emerged together with the COVID-19 pandemic starting in 2020. The STMF provides harmonized weekly death counts for 38 countries

by week, year, age, and sex. The input data for the dataset consists of annual population exposures and weekly death counts sourced from the national statistical offices. As with the other data in the HMD, the same data quality requirements and methods of data harmonization are applied. In order to be included in the dataset, countries' death registrations and census data must be complete. Further, deaths with unknown age are distributed across the age at death distribution and aggregated data is split into finer groups.

A detailed description of the data, its sources, and quality aspects, such as coverage and timeliness, is documented in detail in the Metadata (Short-Term Mortality Fluctuations data series (STMF). Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France), 2025b).

We have chosen the 23 countries (Austria, Belgium, Bulgaria, Croatia, Estonia, Finland, France, Germany, Hungary, Israel, Latvia, Lithuania, Luxembourg, the Netherlands, Norway, Poland, Portugal, Scotland, Slovakia, Slovenia, Spain, Sweden, and Switzerland) whose data series on weekly deaths starts in the year 2000 or before.

2.2 Human Fertility Database

The fertility forecasts in paper 3 are created using data from the Human Fertility Database (HFD) prepared by the Max Planck Institute for Demographic Research (Germany) and the Vienna Institute of Demography (Austria). The HFD provides a variety of fertility measures, such as fertility rates, mean age at birth, and birth counts by age, period, and cohort for 34 countries. The database follows the same principles of data access, quality, and comparability as the Human Mortality Database. The data is accessible free of charge after prior registration. Only data from countries whose birth registrations are complete are included, and a set of applied methods ensures the conformity of the data across countries.

Similarly to the HMD, the fertility measures provided are calculated from birth counts and population counts used as input data, which are mainly sourced from the national statistical offices. For a detailed description of the data, sources and methods applied, see the HFD methods protocol (Jasilioniene et al., 2021).

For paper 3, we use the age-specific fertility rates that have been grouped into 5-year age groups from the HFD. The outcome of the paper is probabilistic fertility forecasts for Finland that have been calibrated on historical forecast errors. We also use the fertility data of Sweden, Norway, and Denmark for the calibration of these forecasts. The four Nordic countries have had a similar fertility development during the last century (Andersson et al., 2009). This joint movement is also known as the *Nordic fertility regime*. Even if, in recent years, slightly diverging patterns between the countries have been observed (Hellstrand et al., 2021), the long lasting joint fertility development qualifies the Nordic countries for the use in the forecast error analysis.

2.3 Learning from past mistakes

The foundation of any forecast error analysis is the forecast errors themselves. Depending on their use in the papers, they are further developed into empirical prediction intervals for the respective forecasts. The basic forecast error is derived by comparing the forecast value $\hat{y}$ with the actual observed value y of the outcome of interest at time point t. Thus, it can only be derived in retrospect. The simplest way of comparing these two values is to take the difference, resulting in the forecast error:

$$e_t = y_t - \hat{y}_t. \tag{2.1}$$

Other error measures, also called error scores, have been developed to allow for more comparability by reducing scale dependency, aggregating forecast errors over time, or relativizing extreme outliers (see Hyndman and Koehler, 2006, for an overview). They all measure the accuracy of deterministic or point forecasts. Other measures used for the assessment of probabilistic forecasts will be introduced in chapter 2.3.3.

Throughout the papers, two different error scores for deterministic forecasts are used. In Paper 1, we use the mean square error (MSE) pooled across all 7 countries r and 33 cohorts c to compare the observed cohort life expectancy $e^{\text{observed}}_{0_{c,r}}$ with the mean of the simulations of Lee-Carter forecasts of cohort life expectancy $\overline{e_{0_{c,r}}}$:

$$\text{MSE} = \frac{1}{33 \times 7 - 1} \sum_{c \times r} (e^{\text{observed}}_{0_{c,r}} - \overline{e_{0_{c,r}}})^2. \tag{2.2}$$

The MSE plays a role in the variance decomposition of total lifespan uncertainty into lifetable lifespan uncertainty and forecast uncertainty.

In paper 2 and 3 we employ the log-relative error u at time t, which is the logged ratio between the observed and the forecast outcome of interest:

$$u_t = ln(\frac{y_t}{\hat{y}_t}). \tag{2.3}$$

This relative error score reduces scale dependency and asymmetry. This is of advantage for modeling the forecast error distribution used to derive empirical prediction intervals.

2.3.1 Empirical prediction intervals

Probabilistic forecasts, in contrast to deterministic point forecasts, allow to distinguish between likely and extreme scenarios by assigning probabilities to possible outcomes (see e.g., Lee, 1998; Keilman, 2018). Usually, this is done by providing upper and lower bounds – that is, a prediction interval – in which the forecast outcome will lie with a given probability, for example, 95%. The quality criterion of probabilistic forecasts are their calibration: Are the forecast probabilities of the outcome consistent with the observed relative frequencies of that outcome? Or, in other words, do the forecast rare events occur rarely, and do the forecast likely events occur regularly? An empirical approach to ensure

good calibration of probabilistic forecasts is the use of empirical prediction intervals. These have originated from the analysis of forecast errors in demographic forecasting already in the 1970s and enjoyed some popularity in the 1980s and early 2000s among forecasters (Williams and Goodman, 1971; Stoto, 1983; Cohen, 1986; Smith and Sincich, 1988; Keilman and Pham, 2004a; Alho and Spencer, 2005; Alho et al., 2008; Rayer et al., 2009). In more recent years, however, the approach has become less common in demographic forecasting, although it continues to be used (Wilson, 2013, 2024).

The basic idea of empirical prediction intervals is that forecast errors are used to estimate a forecast error distribution that is then used to derive prediction intervals for the actual forecast. Usually, the forecast errors are derived from applying the forecast method that is used for the actual forecast to past periods in an out-of-sample design. Interestingly, the forecast errors do not have to stem from the application of the forecast method of interest. For example, Keilman and Pham (2004b) constructed empirical prediction intervals for Nordic fertility forecasts by estimating interval widths from the standard deviation of errors observed in historical forecasts published by statistical agencies that used a variety of methods to forecast.

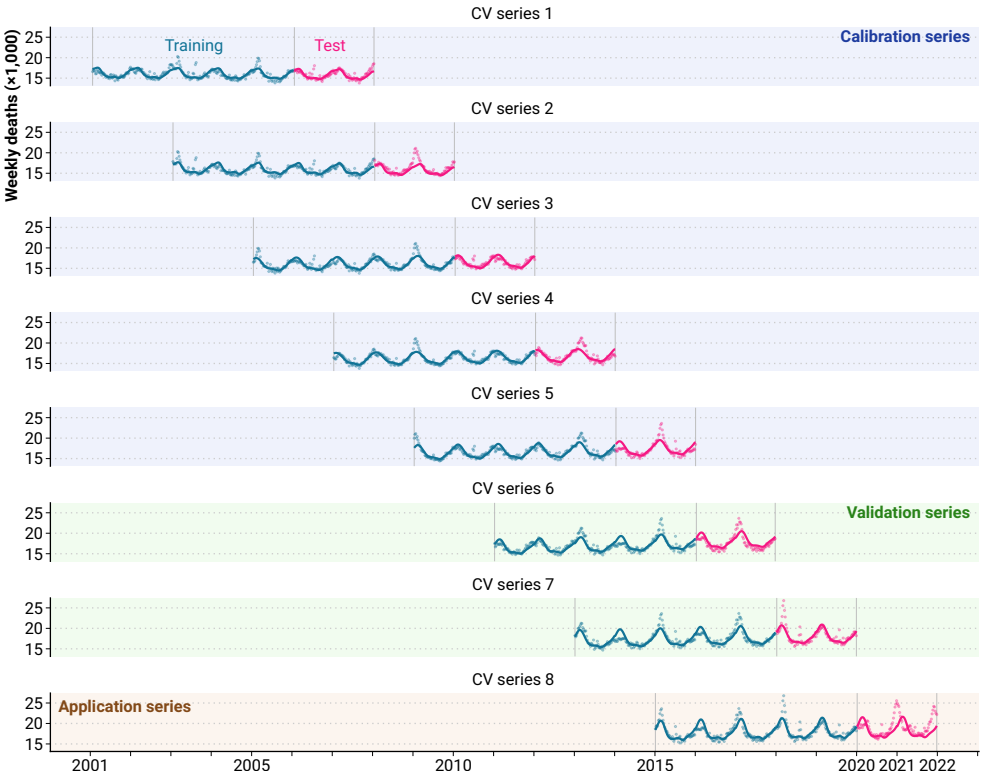
Due to its empirical nature, forecast error analysis is, obviously, not bound to demographic forecast outcomes. Other fields of research are interested in using forecast errors for the calibration of prediction intervals, too. First introduced by Gammerman et al. (1998), conformal prediction (CP, Shafer and Vovk, 2008; Fontana et al., 2023) is a line of research in machine learning that uses out-of-sample errors to generate calibrated probabilistic forecasts. Researchers have developed advanced CP methodology for different applications, including, in more recent publications, time series forecasting (Papadopoulos et al., 2007; Fontana et al., 2023; Angelopoulos et al., 2024). Further, in climate modeling, calibration of predictions using historical data is done under the term empirical quantile-mapping as a form of bias-correction of forecast distributions (Cannon, 2018; Qian and Chang, 2021).

The existing methodology provides probabilistic forecasts calibrated on historical forecast errors in the form of prediction intervals. However, existing research has not yet developed approaches to generate calibrated time series trajectories that align with these interval bounds. Such simulated outcome trajectories are used as inputs for downstream modeling, for example, to analyze and plan the determinants of social security systems. In paper 3, we therefore propose a combination of existing methodology to provide probabilistic forecasts of the Finnish total fertility rate in the form of simulated trajectories whose distribution follows an empirical forecast error distribution.

Data splitting

Before the methodology of deriving and applying empirical prediction intervals is introduced, I will detail how the data needs to be prepared. Precisely, the data containing the demographic measure of interest needs to be split or cut into smaller, partially overlapping data series in the fashion of a cross-validation (CV). The data splitting is visualized in Figure 2.1 on the example of weekly death counts used for the analysis in Paper 2. The aim of this data splitting is to maximize the amount of data points used for the forecast error analysis, to increase generalizability and to reduce over-fitting. Each of these CV

Figure 2.1: Data splitting set-up for the forecast error analysis of short-term forecasts of weekly death counts.



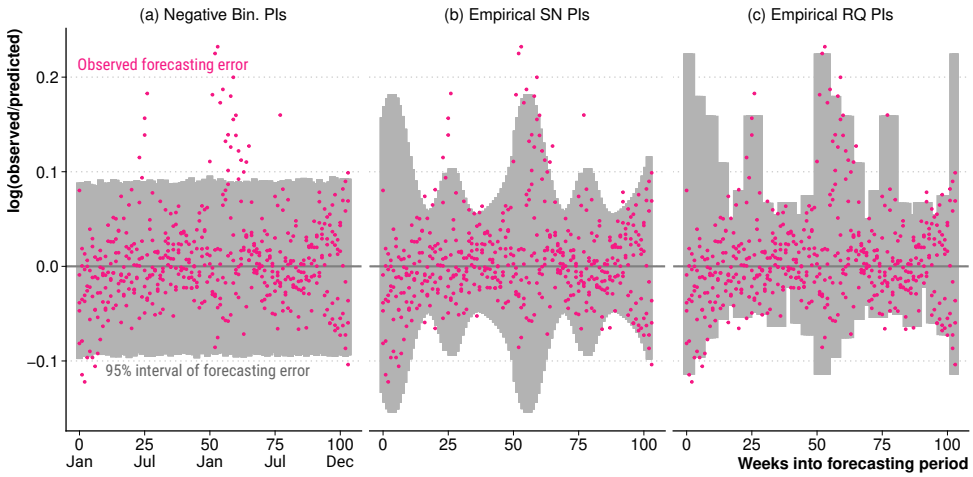
series consists of a training part (blue) and a testing part (pink). The training data is used to fit the forecast model (input data). Then, using these model fits, forecasts are produced for the time span of the testing data. Thus, the observed values of the testing data can be compared to the forecast values in order to calculate the forecast errors.

To derive empirical prediction intervals and apply them to the actual forecast, we divide the CV series into three types of data series that are needed for different methodological steps: Calibration data, validation data, and application data.

2.3.2 Calibration, validation, and application

Calibration The calibration data series are used to calculate a forecast error measure based on the comparison of the observed with the forecast outcome of interest. This is done for each CV series that is part of the calibration data and for each time step of the testing part of the series. Plotting each forecast error by the time step of the testing period (forecast length) results in a forecast error distribution that reflects the historical forecast performance of the model of choice. In the next step, this error distribution can be modeled

Figure 2.2: Observed forecast errors in the calibration data series with modeled forecast error distributions for weekly death forecasts.



using a variety of parameterizations. On the example of forecasts of weekly deaths from paper 2, Figure 2.2 shows the distribution of forecast errors modeled using three different models. While panel a) shows the 95% error interval derived from the negative-binomial variation, panels b) and c) show empirical estimates of the forecast error distribution. For panel b), a time-varying skew-normal distribution was used to capture the skewness and seasonal patterns in the empirical error distribution. The result is a smooth, time-varying interval. Panel c) shows a simpler but rougher approach to modeling by deriving the raw quantiles of the observed error distribution.

Once a forecast error distribution is estimated, for example, by using a time-varying skew-normal model, a distribution for the forecasts themselves can be constructed. This forecast distribution is calibrated to the error distribution in a way that the quantiles of the forecast distribution are derived from the corresponding quantiles of the error distribution. For example, if a given relative error is exceeded 10% of the time, the corresponding forecast distribution is equal to the central forecast times the 90% quantile of the forecast error.

Validation Now that we have an empirical forecast error distribution, we can turn our attention to validating it. Therefore, we employ the validation data series. As seen in Figure 2.1, the validation series are also separated into training and testing periods. Again, our forecast model of choice is fitted to the training periods, and central forecasts are created for each time step of the testing periods. To create empirical prediction intervals around these central forecasts, we derive a calibrated forecast distribution from the quantiles of the error distribution so that the following relationship holds: If a given relative forecast error is exceeded, for example, 10% of the time, the 90% quantile of the calibrated forecast distribution is equal to the central forecast times the 90% quantile of the relative forecast error. This can be applied in the same way to other p-quantiles of

the error distribution. The results are central forecasts with empirical prediction intervals (empirical PIs) of varying widths, for example, 80%, 90% and 95% intervals, which have been calibrated to the empirical forecast errors.

These empirical PIs need to be validated and compared to other PIs to check their performance. Different calibration metrics have been developed to assess the quality of prediction intervals and probabilistic forecasts. I will introduce four metrics that are used throughout the papers of this thesis.

The most intuitive way of checking the quality of a prediction interval is to asses its coverage in a validation setting. By comparing the actual coverage to the nominal coverage we assess which fraction of observed data points is covered by the prediction interval. A prediction interval with perfect coverage is, for example, a 95% prediction interval (nominal coverage) that contains 95% of the observed data points (actual coverage).

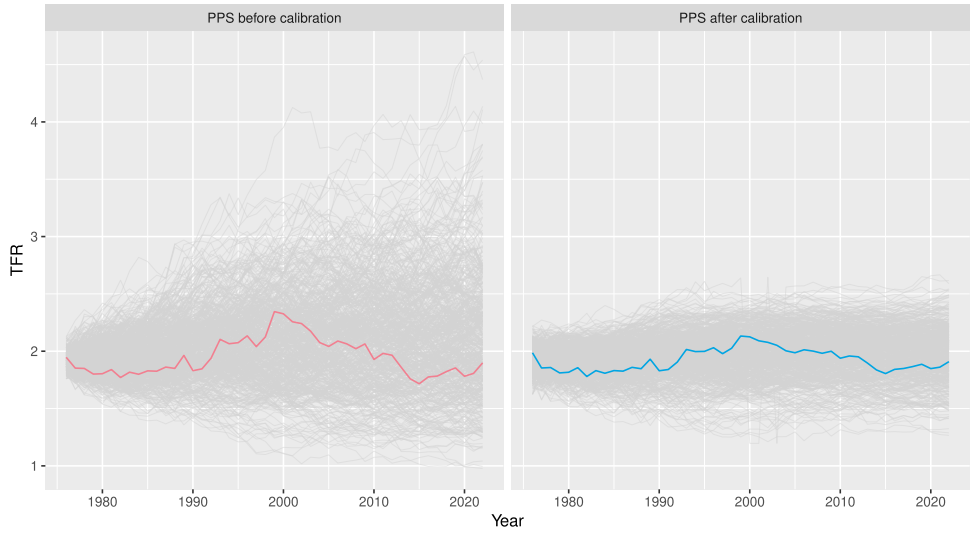
Another main quality criterion is the sharpness of the prediction interval, which can be measured using the mean interval width. The sharpness competes with the coverage. Prediction intervals that are too wide on average render the forecast meaningless, as the results are too imprecise. However, a very narrow interval width may communicate a sense of certainty about the forecast outcome that is unrealistic and could, therefore, negatively impact the coverage.

To simultaneously assess the coverage and the mean interval width, a measure that combines and balances the two criteria can be used. Gneiting and Raftery (2007) propose the Mean Interval Score (MIS) as a proper scoring rule for validating prediction intervals. The MIS punishes prediction intervals that are too wide where they do not have to be, and rewards prediction intervals that are narrow enough to maintain good coverage. For example, when comparing two prediction intervals with equal coverage, the one that is on average narrower will have the better (lower) MIS.

In paper 3, the probabilistic forecast outcome is a set of simulated forecast paths instead of central forecasts surrounded by prediction intervals. Thus, in addition to validating the prediction intervals, we further assess the quality of the entire forecast distribution. We use the Continuous Ranked Probability Score, another proper scoring rule (Gneiting and Raftery, 2007), to comparatively evaluate the distribution of forecast paths between different models.

Application After the validation analysis, if the empirical PIs are deemed of high quality, the actual forecasts of interest can be created. The procedure is similar to the steps taken during the validation part. Using the single application data series which only consists of training data (see Figure 2.1), we first produce the central forecasts for the future. Then, we derive the calibrated forecast distribution using the empirical error distribution estimated from the calibration data. The final result is a probabilistic forecast that has been calibrated to empirical forecast errors.

Figure 2.3: Forecast paths of Finnish TFR from a forecast model (PPS), before (left panel) and after quantile-mapping (right panel).



2.3.3 Quantile-mapping

The methodological steps to create forecasts with empirical PIs described above are used in paper 2 to analyse the forecast error distribution of forecasts of weekly deaths late in the COVID-19 pandemic. The creation of probabilistic forecasts of the Finnish Total Fertility Rate (TFR) in paper 3 deviates from this methodology while maintaining the same core. Instead of deriving empirical PIs of various widths around a central deterministic forecast, the outcome is a full forecast distribution that has been calibrated to the forecast errors. In this case, the central forecast consists of a set of simulated forecast trajectories or paths whose distribution is manipulated to match the empirical error distribution. To do so, we employ quantile-mapping and combine it with the methodological steps of the empirical PIs. After deriving the calibrated forecast distribution from the errors of the calibration cv-series, we map the p-quantiles of any single forecast path of the central forecast to the corresponding p-quantile of the calibrated forecast distribution. Figure 2.3 illustrates the quantile-mapping on the example of forecast paths of the Finnish TFR coming from a model called Postponement Time series model (PPS) (see section 2.3.4 for more details on the fertility modeling and forecasting). From 500 individual TFR paths, a single path is highlighted in color in both panels to illustrate how quantile-mapping transforms the individual forecast paths.

2.3.4 Producing the central forecasts

As described in the section above, before any forecast error analysis can be performed, first, we need to produce the forecasts itself. These central forecasts are not limited to a

specific type of forecast methodology, for example, to extrapolative methods. Any type of forecast model could be employed to produce the central forecasts, such as forecasts based on expert opinions or scenario approaches. That is all under the condition of having the necessary data coverage and requirements to apply the forecast method of choice to the cross-validation series. In the following, I introduce the methodology used to derive the central forecasts of each of the papers that make up this thesis.

Lifetables and their variances The aim of the first paper is to disentangle the contribution of forecast variance and lifetable lifespan variance to the total individual lifespan variance. We denote by X the random age at death of a newborn and by $\boldsymbol{\mu}$ the random vector of age-specific cohort death rates parameterizing the distribution of X. How much of the uncertainty about the newborns eventual lifespan is driven by within-lifetable stochasticity, and how much is driven by the uncertainty about which future lifetable will apply to the newborn? The law of total variance, as demonstrated in the context of lifetables by Caswell (2023), allows us to answer this question by decomposing the total variance in the newborns age at death, $\mathrm{Var}(X)$, into variance due to uncertainty in future death rates, $\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})]$, and the expected variance inherent to any lifetable, $\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X|\boldsymbol{\mu})]$:

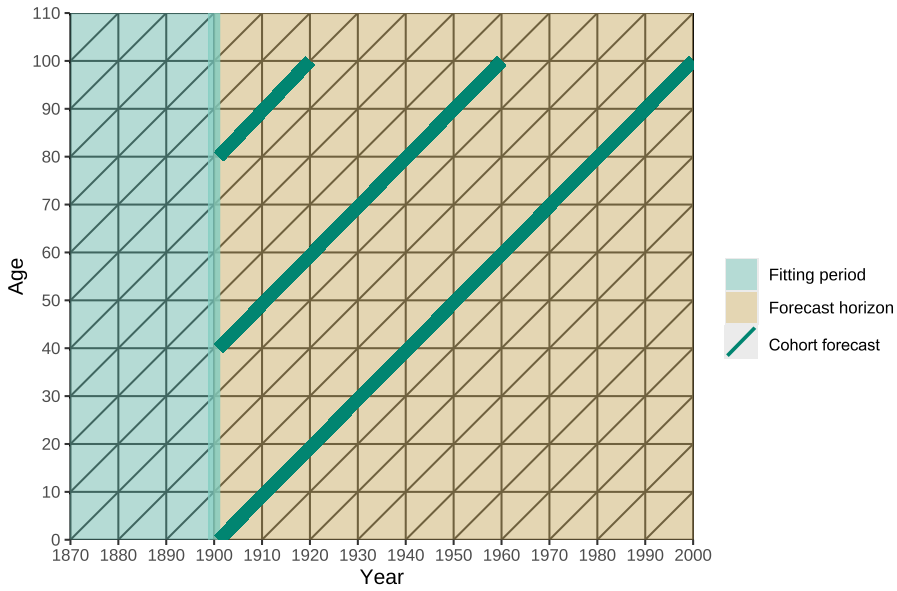
$$\underbrace{\mathrm{Var}(X)}_{\substack{\text{Total lifespan variance}}} = \underbrace{\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X|\boldsymbol{\mu})]}_{\substack{\text{Within-lifetable variance /}\\ \text{Expected lifespan variance}}} + \underbrace{\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})]}_{\substack{\text{Between-lifetable variance /}\\ \text{Forecast variance}}} . \tag{2.4}$$

The lifetable is the classic tool for demographers to portray and analyze living and dying in a population. The cohort lifetable describes the mortality experience of a birth cohort until its extinction. We can also construct lifetables with a period perspective to describe the experience of a synthetic cohort that undergoes the mortality of a specific period, for example, a year. Given the availability of age-specific death and population counts, several mortality measures can be derived using the lifetable. These are, for example, age-specific mortality and survival rates, age-specific life expectancy including life expectancy at birth, and lifespan variance. For birth cohorts that have already gone extinct, deriving these measures is straightforward, as all mortality information is available. But what about cohorts with members that are still alive? In this case, the respective cohort lifetable needs to be *completed* by forecasting the mortality experience of those cohort members that are still alive.

One way of retrieving cohort lifetable forecasts is to exploit the relationship between age, period, and cohort. This relationship can be visualized using a Lexis diagram. On the horizontal axis, we find the period in years, and on the vertical axis, the age in single increments. Thus, a birth cohort member moves through time on the diagonal of this plane. Therefore, if we forecast the period mortality across ages and years, we can extract the mortality forecast for a birth cohort by following the diagonal. This is illustrated in Figure 2.4 for period forecasts starting in 1900 and cohorts that were either born, have turned 40, or have turned 80 in this year.

We chose the standard Lee-Carter model (Lee and Carter, 1992) to forecast the period lifetables. The Lee-Carter model is a simple extrapolative model commonly used for

Figure 2.4: Visualization of the diagonal extraction of lifetable forecasts for the birth cohort of 1900.



mortality and fertility forecasting. We fit the Lee-Carter model on the 30 years of data prior to the birth of a cohort and then perform 100-year period forecasts starting at each cohort's birth year. This allows us to extract a 100-year cohort diagonal from the period-age surface. Further, we extract the mortality of cohorts that have turned 40 and 80 at the beginning of the forecasts. This allows us to perform the variance decomposition of the lifespan uncertainty faced by an individual at different ages.

Expected and excess deaths We derive the short-term mortality forecasts for the estimation of excess deaths in paper 2 using a methodology that is also employed by the World Health Organization (WHO, Weinberger et al., 2020). We forecast expected deaths $\hat{y}$ at time $T + h$ using an overdispersed count regression,

$$\hat{y}^{E}_{T+h} = \exp\left(\alpha + \beta_h h + s_w\left(w\left[h\right]\right)\right), \tag{2.5}$$

where T is the last time-point in the training data and h is the number of weeks into the forecasting horizon, $\beta_h h$ captures the long term trend of deaths, and $s_w\left(w\left[h\right]\right)$ is a cyclical penalized spline over the week of the year to reflect the seasonality of death counts. We assume the weekly deaths under the expected scenario to be distributed as

$$Y^{E}_{T+h} \sim \text{Neg.Binomial}(\hat{y}_{T+h}, \phi). \tag{2.6}$$

Following this, the estimation of excess deaths results from the difference between the observed weekly death counts and the expected deaths forecast.

Finnish fertility The fertility forecasts for Finland are derived from two scenario-based approaches that use age-specific fertility rates to forecast the total fertility rate (TFR). The main model, called the Postponement Time Series Model (PPS), is based on the assumption that the fertility postponement that took place in Finland would continue, but gradually slow down and eventually stop (Nisén et al., 2020). We forecast time series of 5-year age group fertility rates using a random-walk-with-drift (RWD) model

$$\ln(\hat{y}_{x,t}) = \beta_{x,t} + \ln(y_{x,t-1}) + \epsilon_{x,t}, \epsilon_{x,t} \sim N(0, \sigma_x^2), \tag{2.7}$$

where x is the age group, t is the calendar year, $\beta_{x,t}$ is the model drift, and $\epsilon_{x,t}$ is the error term. The drift term $\beta_{x,t}$ is forecast under two calculation assumptions. First, the increase in the average age at birth slows down and the TFR approaches the tempo-adjusted TFR of the last observed year. Thereafter, there is no drift term: $\beta_{x,t} = 0$. Finally, the TFR is obtained by adding up the age-group-specific fertility rates and multiplying it by the width of the age-groups. The second and simpler scenario serves as a naïve baseline for the later comparison of forecast performance. This model also forecasts the age-specific fertility rates using a random walk but assumes that the fertility rates are fixed at the level of the last observed year.

Chapter 3

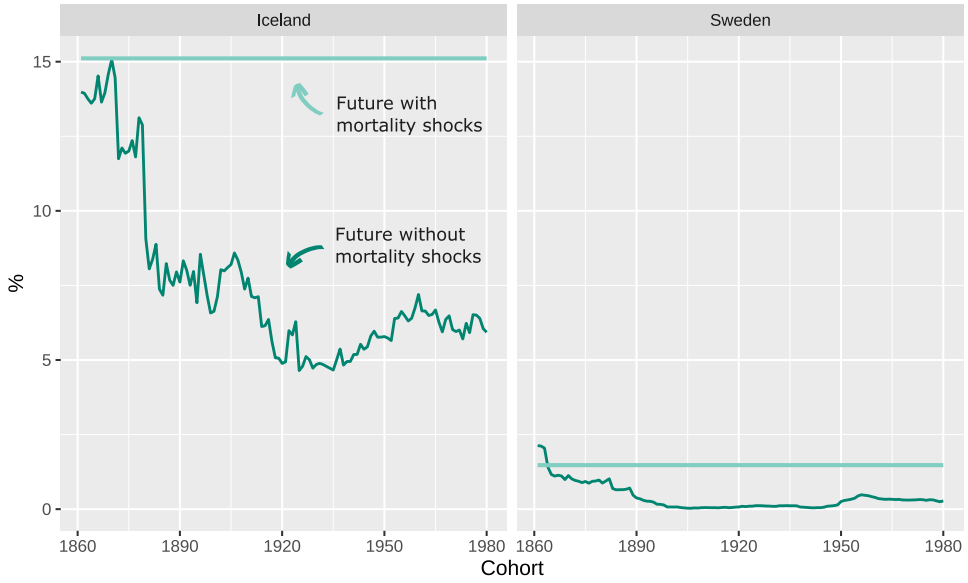
Results

In the following, I will summarize the aims and main findings of the papers that make up this thesis. The results were derived using the data and methods described above.

3.1 Paper 1: The contribution of forecast uncertainty to lifespan uncertainty

In this paper, we decompose the total lifespan uncertainty of an individual cohort member into forecast uncertainty and lifetable lifespan uncertainty. Doing so, we can either back-up or challenge the interpretation of lifespan variance at the population-level as an indicator of individual lifespan uncertainty, depending on the results. In this section, I present the results using the example of the male birth cohorts of Sweden and Iceland. Figure 3.1 shows the contribution of the forecast variance to the overall lifespan uncertainty for scenario A (dark green) and scenario B (light green) at age 40 over the cohorts from 1861 to 1980. The forecast uncertainty of scenario A stems from the Lee-Carter forecasts used to complete the cohort lifetables and corresponds to a future without mortality shocks. Scenario B depicts a future where extreme mortality is a possibility. Here, the forecast uncertainty is derived from the empirical forecast error using the Mean Square Error (MSE). A development of the share over cohorts cannot be assessed for scenario B because the MSE was aggregated over all cohorts. The two countries showcase the strongest differences in our sample of countries. In the case of Sweden, the share of forecast variance is very low (below 2%) in both scenarios across cohorts. For the Icelandic cohorts, we observe a reduction in the share starting at around 14% and reaching a level around 6% since cohort 1920 in scenario A. Further, the MSE for Iceland (15.1%) is higher than for Sweden (1.5%). Summarizing these results and the additional analysis outlined in the corresponding paper, we conclude that the forecast variance contributes a low share to the overall lifespan variance under scenario A – that is, under the assumption that future mortality develops without extreme shocks. When the possibility of such drastic mortality events due to, for example, wars or pandemics, is taken into account, the share of forecast

Figure 3.1: Share of forecast variance on total cohort lifespan variance at age 40 under a scenario without future mortality shocks and under a scenario with mortality shocks among male birth cohorts of Iceland and Sweden.



Note: The estimate of the share of forecast variance under the scenario with mortality shocks is constant across cohorts. It is based on pooling the forecast errors across all cohorts and simulations.

variance is still low, with exceptions such as Iceland. The findings for Iceland stand out due to the small population size, which translates into more volatile mortality rates that are prone to higher forecast uncertainty. Further, we show that the observed decrease in overall lifespan variance over time is primarily driven by the decrease in the lifetable lifespan variance and not the forecast variance. Thus, these findings generally support the interpretation of lifespan variability as an indicator of individual lifespan uncertainty.

3.2 Paper 2: Empirical prediction intervals applied to short-term mortality forecasts and excess deaths

The motivating research question of this paper was how unusual it is to observe 10% excess deaths throughout the year at the then current stage of the Covid-19 pandemic in 2022. In other words: What is the significance of elevated mortality given the uncertainty in statistical estimations? Using empirical prediction intervals, we show that commonly used conventional prediction intervals for the estimation of expected deaths are usually too narrow and thus give a false impression of certainty around excess death estimates. In the following, I will summarize the findings of paper 2 on the example of Germany.

From the analysis of the empirical forecast errors, it is clear that the forecast error of

the expected deaths model used is positively skewed. In other words, it underestimated the actual mortality more times than it overestimated it. The observation of skewness justifies the use of the skew-normal model as one of the options to model the forecast error distribution. In addition, we can see a strong seasonal pattern in the distribution of the forecast error: The forecast errors are larger in winter and also slightly elevated in summer. Compared to the other approach of using the raw quantiles of the distribution, modeling it with the skew-normal function results in a smoother representation. However, both models are able to capture the seasonality in the forecast error. When comparing these results with the commonly used negative-binomial approach to estimate the uncertainty in expected deaths, it becomes evident that this approach is not able to capture the seasonality and assumes a constant error over the year.

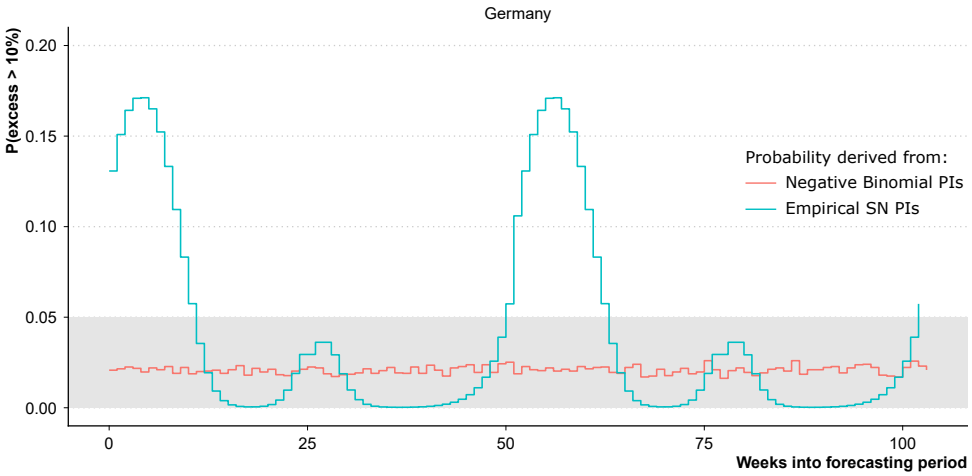
These results from the analysis of the forecast error directly translate into the prediction intervals derived from them. Both types of empirical prediction intervals are wider in winter compared to the negative binomial PIs. If one were to judge from the negative-binomial PIs, excess deaths in Winter would be considered unusual, as they lie outside of the rather narrow PI. However, judging from the empirical PIs that take the historically observed uncertainty into account, one would come to the conclusion that observing excess deaths in winter is not that unusual. Further, the raw quantiles approach causes a rougher interval that is not as smooth as the PI modeled with the skew-normal function.

The validation of the calibration of the different types of prediction intervals was performed on data from all countries of the analysis. Comparing different calibration metrics for a nominal coverage of 95% reveals that all PIs tend to be too narrow, but the PIs constructed using the negative binomial and the empirical PIs modeled with the skew-normal model come close to the nominal coverage. Across all metrics, the empirical PIs using the raw quantiles approach perform the worst. This highlights the need for more elaborate modeling of the forecasting error distribution to achieve smoother and therefore generalizable prediction intervals and avoid overfitting. Considering the coverage metric on an annual basis, the negative-binomial PIs perform best. However, this hides the seasonality of the forecast uncertainty. When looking at the coverage by season, in autumn and winter, the best coverage is achieved by the empirical skew-normal intervals. Further, it becomes obvious that the negative-binomial PIs are too narrow in winter and too wide in autumn. When taking this width into account by using the mean interval score metric, the empirical skew-normal PIs are best calibrated annually and in each season except summer, when the negative-binomial PIs perform better.

Regarding the better performance of the empirical prediction intervals compared to those based on the negative-binomial model, it should be noted that seasonality can also be incorporated into negative-binomial prediction intervals, for example, by using time-varying overdispersion (Pan et al., 2018). However, such approaches are not commonly used and require more elaborate model specifications than empirical prediction intervals.

Coming back to the original motivating research question, if negative binomial PIs are used, in the case of Germany, one would come to the conclusion that observing 10% excess deaths is just as unusual in winter as in summer. However, using empirical prediction intervals reveals the seasonality in this uncertainty. The likelihood of exceeding the threshold of 10% excess deaths varies by season (see Figure 3.2). This applies to all 23

Figure 3.2: Weekly probability for 10% excess deaths given pre-pandemic mortality derived from empirical prediction intervals and negative-binomial prediction intervals for Germany.



countries in the analysis. However, the periods for which 10% excess is not unusual to observe differ by country.

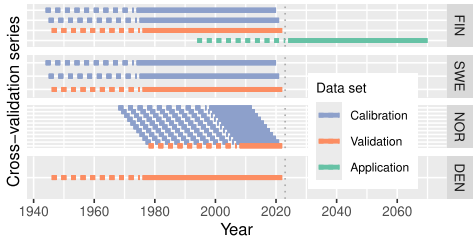
3.3 Paper 3: Calibrating probabilistic forecast paths on past forecast errors: An application to the Finnish total fertility rate

Paper 3 of this thesis is providing scenario-based forecasts of the Finnish TFR that are calibrated on historical forecast errors. By combining the methodology of modeling the forecast error distribution with the quantile-mapping approach, we calibrate probabilistic forecasts paths of the TFR in Finland up to 2070. Figure 3.3 guides the reader through the methodological steps and summarizes the results of the paper.

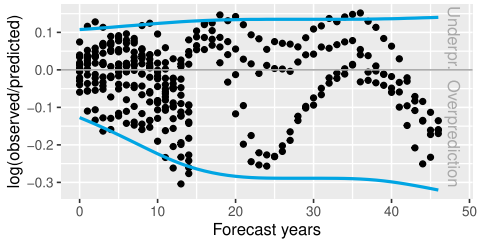
Analyzing the forecast errors of the postponement scenario model that has been applied to the calibration data shows that the model tends to overestimate the actual observed TFR in the four Nordic countries (panel 2). This overestimation is apparent from the earliest forecast years. This bias justifies the use of the skew-normal function to model the error distribution to increase smoothness and generalizability. Further, it is necessary to address the data scarcity in the later forecast years. After calibrating the forecast paths stemming from the PPS model using the model of the historical error distribution (panel 3), the 95% interval is narrower toward the end of the forecast horizon and wider in the beginning, compared to uncalibrated PPS forecast paths. Further, the negatively skewed error distribution results in a shift of the median forecast to a lower value (TFR

Figure 3.3: Graphical summary of the methodological steps and results of the analysis in Paper 3.

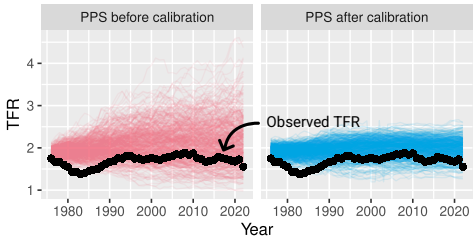
1 Split data into sliding window cross-validation sets
Split data into calibration, validation and application data sets. Divide each cross-validation series (cv-series) into a training period (dotted) and a testing period (solid).



2 Estimate time-varying distribution of forecast errors
Apply the postponenment scenario model (PPS) to the training period of all calibration data series to calculate the **forecast errors** for the testing period. Pool the forecast errors across all cv-series and model the empirical forecast error distribution.



3 Make calibrated forecasts on validation series
Apply the PPS model to the training period of all validation data series. Then calibrate the resulting forecast paths using the model of the empirical forecast error distribution. The graphs show the forecasts for the Danish validation series.

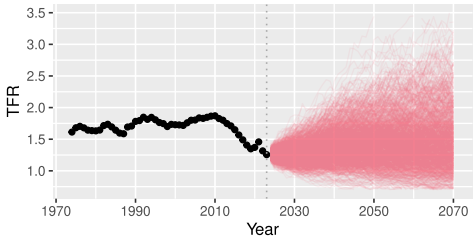


4 Calculate calibration scores
Calculate calibration scores for the testing period of all validation data series and compare the forecast performance of the PPS model before and after the calibration with the empirical forecast error distribution.

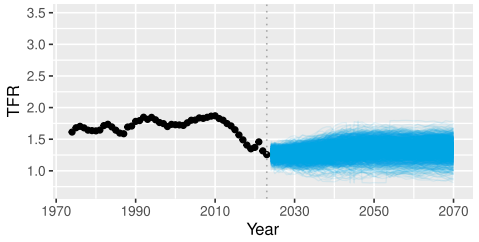
Model	Coverage	Mean Interval Score	Average width
PPS before calibration	**0.915**	1.728	1.340
PPS after calibration	0.901	**1.087**	**0.710**
naive freeze-rates model	**0.915**	1.590	1.240

The table shows error scores for the 95% quantile, pooled across all validation series. Best scores highlighted in **bold**.

5 Generate forecasts for Finland up to 2070
Apply the PPS model to the training period of the application data series to produce probabilistic forecast paths for **Finland** up to 2070.



6 Calibrate Finnish forecasts using error distribution
Calibrate the PPS forecast for the Finnish application series using the empirical forecast error distribution to produce the final forecast of the Finnish TFR up to 2070.



of 2.07 before calibration, and 1.99 after calibration at the end of the forecast horizon). The forecast validation analysis shows a better performance of the forecast paths from the PPS model that have been calibrated on the forecast errors across several metrics compared to the uncalibrated PPS paths and the naive model (panel 4). These results are mainly consistent for three different nominal coverages (80%, 90%, and 95%), for the full forecast horizon and for different forecast lengths, and across several error measures. The application of the empirical forecast error distribution to the actual forecast of Finland's TFR up to 2070 results in a slimmer forecast distribution with a median of 1.35 in 2070 (panel 6), compared to the forecast before the calibration (panel 5).

Chapter 4

Discussion

There are different sources of uncertainty in demographic forecasting. One source is the sub-optimal data quality of the input-data due to mis-reporting in, for example, censuses, data gaps that need to be filled via estimation, lack of coverage – be it across time or across the population – or other signs of low-quality data. Next, uncertainty in demographic forecasting stems from our inability to forecast the likelihood, timing, and magnitude of extreme events that heavily impact demographic measures. Examples are wars, pandemics, or (geo-)political changes. Further, an important source of uncertainty is inherent to the forecasting methodology itself. Depending on the basic forecast type (extrapolative, scenario-based, expert opinion, or a combination of these), there is uncertainty around model parameters due to statistical randomness, and uncertainty stemming from inaccurate model choice and specification. In addition, there is stochastic or aleatoric uncertainty, due to the inherent randomness in the outcome. From these sources, only the model-based types of uncertainty and the aleatoric uncertainty, if specified, are expressed in conventional prediction intervals. However, the empirical forecast error captures all of these sources of uncertainty and expresses them in a neat, easily understandable metric.

In this thesis, I have shown that the analysis of forecast errors provides meaningful insights for demographic forecasting. By combining forecast error analysis with further methodological steps, such as modeling the forecast error distribution, deriving empirical prediction intervals, and mapping forecast paths to a distribution that has been calibrated to forecast errors, I presented improved forecasts in the papers that make up this thesis.

The results of these papers are based on an assumption of continuity. It is the main assumption that allows us to derive empirical prediction intervals from observed forecast errors. I call it the Big Assumption: The forecast errors of the future will resemble the forecast errors of the past. This essential assumption is more conservative than its counterfactual, the belief that the future will be vastly different from the past. Such an assumption would need much more justification because it would mean that we expect demographic forecasts to suddenly change drastically in their performance in a way that does not resemble any changes in the past that are included in the calibration data of the empirical prediction intervals. Thus, as Alho et al. (2008) argue, it is necessary to provide

arguments for this counterfactual in order to challenge the assumption of continuity.

The Big Assumption is challenged, one might think, by the use of scenario-based forecast models as their applicability is bound to a specific demographic regime in time off of which they are based. However, in the third paper of this thesis, we have shown how scenario-based forecasting can be combined with empirical prediction intervals even if the period of applicability is small. By including additional data from countries with similar fertility regimes and over-lapping the cross-validation series, we were able to prove the applicability and better performance of this methodology compared to standard ways of estimating forecast uncertainty.

The main limiting factor for application of forecast error analysis and empirical prediction intervals in demographic forecasting is data availability. This also challenges the theoretically general applicability of empirical PIs across scientific fields and forecast model types. Due to the empirical PIs hunger for long data series, our selection of countries in the analyses of the papers is limited to high-income, low-mortality countries which have longer time series and high data quality provided by the HMD and HFD. The cause for the high data demand is the cross-validation design with its separation of data into calibration, validation and application series. However, the efficiency of data usage can be improved by overlapping of the cross-validation series, as shown in paper 3. This holds the risk of over-fitting and reduced generalizability of the results, but can be mitigated by strict hold-out datasets for the validation part of the analysis. There is another interesting approach that might be worth investigating in cases when the cross-validation design is not applicable due to data constraints or the nature of the chosen forecast model, but empirical prediction intervals are still desired: Prediction intervals based on errors in actual published historical forecasts. Instead of applying the forecast model to the cross-validation design, the forecast errors come from the comparison of observed values with actual forecasts created in the past by, for example, national statistical agencies. Keilman and Pham (2004b) introduced this method to determine the forecast error of fertility forecasts for the Nordic Countries.

Neither in paper 2 nor in paper 3 are the empirical prediction intervals perfectly calibrated, as judged by the calibration scores in the validation part of the analyses. Hence, empirical prediction intervals are not a perfect solution to estimate the uncertainty of our forecasts. Still, they perform well and better than other methods of estimating forecast uncertainty, while being based on an empirical and thus intuitive and easily communicable notion: We check how wrong our forecasts would have been if we had applied them in the past and use these insights to inform the uncertainty around our actual forecasts. At the same time, the results of these papers highlight that further methodological work is needed to account for sources of uncertainty that are not adequately captured by existing approaches. One such area is seasonal uncertainty, which remains under-researched in demographic forecasting.

Chapter 5

Conclusion

What is the aim of science? To produce knowledge. It seems paradoxical to write this thesis and come to the conclusion that the biggest contribution I made was to advocate for more uncertainty. But in the same way as saying, "I know that I don't know", it is of utmost importance to acknowledge our uncertainties, to put a number on how certain we are about our research results. Especially with demographic forecasting, it is not only the humble but also the ethically right decision to communicate that our results are nothing more than an educated guess. Looking back at our past results reveals how uncertain the future actually is and how limited our ability to predict it is. Therefore, can there be a better way to improve our predictions for the future than by learning from our past mistakes?

Bibliography

Alho, J. M., H. Cruijsen, and N. Keilman (2008). Empirically based specification of forecast uncertainty. In J. M. Alho, S. E. Hougaard Jensen, and J. Lassila (Eds.), *Uncertain demographics and fiscal sustainability*, pp. 34–54. Cambridge: Cambridge University Press.

Alho, J. M. and B. D. Spencer (2005). Uncertainty in demographic forecasts: Concepts, issues, and evidence. In *Statistical demography and forecasting*, pp. 226–268. Springer.

Andersson, G., M. Rønsen, L. B. Knudsen, T. Lappegård, G. Neyer, K. Skrede, K. Teschner, and A. Vikat (2009). Cohort fertility patterns in the nordic countries. *Demographic research 20*, 313–352.

Angelopoulos, A., E. Candes, and R. J. Tibshirani (2024). Conformal pid control for time series prediction. *Advances in neural information processing systems 36*.

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics 50*(1), 31–49.

Caswell, H. (2023). The contributions of stochastic demography and social inequality to lifespan variability. *Demographic Research 49*, 309–354.

Cohen, J. E. (1986). Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. *Demography 23*(1), 105–126.

Edwards, R. D. (2013). The cost of uncertain life span. *Journal of population economics 26*(4), 1485–1522.

Edwards, R. D. and S. Tuljapurkar (2005). Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Population and Development Review 31*(4), 645–674.

Fontana, M., G. Zeni, and S. Vantini (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli 29*(1), 1–23.

Gammerman, A., V. Vovk, and V. Vapnik (1998). Learning by transduction. arXiv:1301.7375.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Hellstrand, J., J. Nisén, V. Miranda, P. Fallesen, L. Dommermuth, and M. Myrskylä (2021). Not just later, but fewer: Novel trends in cohort fertility in the nordic countries. *Demography 58*(4), 1373–1399.

Human Fertility Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) (2025). Available at: https://humanfertility.org/.

Human Mortality Database (HMD). Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) (2025). Available at: https://mortality.org/.

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

Jasilioniene, A., D. A. Jdanov, T. Sobotka, E. M. Andreev, K. Zeman, and V. M. Shkolnikov (2021). Methods protocol for the human fertility database. Technical report, Max Planck Institute for Demographic Research, Rostock (Germany) and Vienna Institute of Demography (Austria).

Keilman, N. (2018). Probabilistic demographic forecasts. *Vienna Yearbook of Population Research 16*, 25–36.

Keilman, N. and D. Q. Pham (2004a). Empirical errors and predicted errors in fertility, mortality and migration forecasts in the european economic area. *Dicussion Papers 386*. Statistics Norway, Research Department.

Keilman, N. and D. Q. Pham (2004b). Time series based errors and empirical errors in fertility forecasts in the nordic countries. *International Statistical Review 72*(1), 5–18.

Lee, R. D. (1998). Probabilistic approaches to population forecasting. *Population and Development Review 24*, 156–190.

Lee, R. D. and L. R. Carter (1992). Modeling and forecasting us mortality. *Journal of the American statistical association 87*(419), 659–671.

Nepomuceno, M. R., Q. Cui, A. Van Raalte, J. M. Aburto, and V. Canudas-Romo (2022). The cross-sectional average inequality in lifespan (cal†): A lifespan variation measure that reflects the mortality histories of cohorts. *Demography 59*(1), 187–206.

Nisén, J., J. I. S. Hellstrand, P. Martikainen, and M. Myrskylä (2020). Hedelmällisyys ja siihen vaikuttavat tekijät suomessa lähivuosikymmeninä [Fertility and factors affecting it in Finland in the coming decades]. *Yhteiskuntapolitiikka 85*(4), 358–369.

Pan, A., S. E. Sarnat, and H. H. Chang (2018). Time-series analysis of air pollution and health accounting for covariate-dependent overdispersion. *American Journal of Epidemiology 187*(12), 2698–2704.

Papadopoulos, H., V. Vovk, and A. Gammerman (2007). Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Volume 2, pp. 388–395. IEEE.

Qian, W. and H. H. Chang (2021). Projecting health impacts of future temperature: a comparison of quantile-mapping bias-correction methods. *International journal of environmental research and public health 18*(4), 1992.

Rayer, S., S. K. Smith, and J. Tayman (2009). Empirical prediction intervals for county population forecasts. *Population Research and Policy Review 28*(6), 773–793.

Sasson, I. (2016). Trends in life expectancy and lifespan variation by educational attainment: United states, 1990–2010. *Demography 53*(2), 269–293.

Shafer, G. and V. Vovk (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research 9*, 371–421.

Short-Term Mortality Fluctuations data series (STMF). Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) (2025a). available at: https://mortality.org/.

Short-Term Mortality Fluctuations data series (STMF). Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) (2025b). Metadata.

Smith, S. K. and T. Sincich (1988). Stability over time in the distribution of population forecast errors. *Demography 25*(3), 461–474.

Stoto, M. A. (1983). The accuracy of population projections. *Journal of the American Statistical Association 78*(381), 13–20.

UNESCO (2022). Understanding open science. *UNESCO open scince toolkit* (36).

Van Raalte, A. A., I. Sasson, and P. Martikainen (2018). The case for monitoring life-span inequality. *Science 362*(6418), 1002–1004.

Weinberger, D. M., J. Chen, T. Cohen, F. W. Crawford, F. Mostashari, D. Olson, V. E. Pitzer, N. G. Reich, M. Russi, L. Simonsen, et al. (2020). Estimation of excess deaths associated with the covid-19 pandemic in the united states, march to may 2020. *JAMA internal medicine 180*(10), 1336–1344.

Williams, W. H. and M. L. Goodman (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association 66*(336), 752–754.

Wilmoth, J. R., K. Andreev, D. A. Jdanov, D. A. Glei, and T. Riffe (2021). Methods protocol for the human mortality database. Technical report, Max Planck Institute for Demographic Research, Rostock (Germany) and University of California, Berkeley (USA) and French Institute for Demographic Studies (France).

Wilson, T. (2013). Quantifying the uncertainty of regional demographic forecasts. *Applied Geography 42*, 108–115.

Wilson, T. (2024). Visualizing the shelf life of population forecasts: A simple approach to communicating forecast uncertainty. *Journal of Official Statistics 40*(1), 38–56.

The contribution of forecast uncertainty to lifespan uncertainty

Authors: Ricarda Duerst[1,2], Jonas Schöley[1]

Affiliations:
[1] Max Planck Institute for Demographic Research; Rostock, Germany
[2] University of Helsinki; Finland

Abstract

Background
Period lifespan variability is sometimes interpreted as an indicator of lifespan uncertainty for individual cohort members. Is this an admissible interpretation for living cohort members, given that future death rates are not precisely known?

Objective
To answer this question, we express individual lifespan uncertainty as a function of variation in ages at death in a cohort lifetable and uncertainty about future death rates. We then quantify the importance of each component to total individual lifespan variability of a birth cohort. A small contribution of the forecast variance would support the interpretation of lifespan variability as individual lifespan uncertainty.

Methods
Using Human Mortality Database data for seven European countries, we decompose the total variance in the remaining lifespan at birth, age 40, and age 80 of an individual cohort member into cohort lifetable variance and forecast variance for cohorts born 1871-2019. We do so under Lee-Carter cohort forecast variance and under an empirical forecast variance measure which allows for mortality shocks.

Results
Even when accounting for the possibility of future mortality shocks, the forecast variance component tends to contribute less than 10% to the total lifespan variance of an individual. Exceptions are France and the small and thus volatile population of Iceland. Conditioning on the absence of future mortality shocks, the forecast variance contributes around 1% to total lifespan variance.

Contribution
We give credibility to the interpretation of period lifetable lifespan variance as individual lifespan uncertainty by showing that the lifespan variance of a cohort is mainly determined by expected lifetable variance with smaller contributions from forecast variance.

Key words
Variance decomposition; cohort mortality; mortality forecasting; lifespan variation; lifespan inequality

1 Introduction

The variability in the age at death – that is, lifespan inequality – is the focus of ongoing research (Caswell 2023; Aburto et al. 2020; Van Raalte et al. 2018; Colchero et al. 2016; Seligman et al. 2016; Vaupel et al. 2011; Edwards and Tuljapurkar 2005) including the development of new measures (Nepomuceno et al. 2022), a focus on regional (Wilson et al. 2020) and socio-demographic disparities (Permanyer et al. 2018; Sasson 2016; Brown et al. 2012) and lifespan inequality in relation to economic decision making (Edwards 2013; Hurd et al. 2004). Like life expectancy, lifespan variability is estimated from either a cohort or a period lifetable and is a summary measure of the lifetable distribution of deaths.

Some research has interpreted lifetable lifespan variability as individual uncertainty (or predictability) of a person's eventual age at death (Nepomuceno et al. 2022; Van Raalte et al. 2018; Sasson 2016; Edwards 2013; Edwards and Tuljapurkar 2005). However, such a personal interpretation is contentious as the individual has access to different information about their future lifespan than is given by the lifetable alone. We distinguish three sources of information about individual lifespan: what the individual knows about themselves, what they know about others, and what they know about the future.

1. Personal information: Different strata of a population experience different levels of lifespan uncertainty. This has for example been shown across socioeconomic groups (Permanyer et al. 2018; Sasson 2016; Brown et al. 2012). Individual-level information on factors such as education, income, lifestyle, and medical histories can reduce the individual lifespan uncertainty. For example, a life-time smoker knows about the detrimental effects of their health behaviour on their expected length of life. Therefore, the individual's lifespan uncertainty is not the same as the lifespan uncertainty researchers can estimate with limited information. However, Caswell (2023) shows that even with extreme levels of heterogeneity in the population, only a few percent of variance in the length of life can be attributed to inequality between groups. Instead, it is the within-lifetable variance that explains most of the lifespan variability.

2. Interpersonal information: Another source of individual uncertainty or certainty about the lifespan is the mortality experience of others. While the period lifetable is a source of such interpersonal information on lifespans, it is only a snapshot at a single point in time. Nepomuceno et al. (2022), argue that individuals base their assessment of their lifetime uncertainty not only on the current mortality experience of others (period mortality), but also on the past mortality experience of contemporary cohorts: "it is possible that people adjust their survival expectations on the basis of the survival histories of a broader mixture of family members, colleagues, and neighbors [...] who were from mixed birth cohorts" (Nepomuceno et al. 2022, p188). Thus, their proposed cross-sectional measure of lifespan variability $\text{CAL}^{\dagger}$ takes both the period and the cohort perspective into account to allow for more informative monitoring of mortality changes.

3. Information about the future: Further, an individual has uncertainty about which death rates will apply to them. Only in a population in which age-specific death rates do not change with time would they know what their future mortality rates will look like. But mortality rates are subject to constant fluctuations and sudden mortality shocks, for example, due to pandemics, other natural disasters, and wars (Schöley et al. 2022). Additionally, it is uncertain to what degree long-term mortality trends will continue into the future as illustrated by slowing rates of mortality improvements in some countries (Hum et al. 2015; Ford et al. 2019; Murphy 2021). Thus, an individual has uncertainty about future death rates – that is, how their lifetable will be completed. This "forecast uncertainty" adds to the individual's total lifespan uncertainty.

We focus on the contribution of forecast uncertainty to total lifespan uncertainty and ask whether the interpretation of lifespan variability as individual lifespan uncertainty is admissible for living members of a cohort. A small contribution of the forecast variance to the total lifespan variance would support the interpretation of lifespan variability as individual lifespan uncertainty.

To separate the uncertainty faced by an individual about the length of its lifespan we divide it into the lifetable component of variability – that is, uncertainty *due to* the shape of the survival curve (lifetable lifespan variance) – and the forecast component of variability – that is, uncertainty *about* the shape of the survival curve (lifetable forecast variance) – as the cohort lifetable has to be completed using forecasts of age-specific mortality rates. For an individual cohort member, due to the law of total variance:

Total Lifespan Variance =

Expected Lifetable Lifespan Variance + Variance of Lifetable Forecasts.

We estimate each of the above components across a range of birth cohorts in seven countries. First, we perform stochastic Lee-Carter (Lee and Carter 1992) lifetable forecasts across birth cohorts. Adopting a methodology proposed by Caswell (2023), we then perform a variance decomposition of the total cohort lifespan variance into a lifetable variance and a forecast variance component. This allows us to examine the share of each component, and observe their changes and the development of the total lifespan variance over time.

We obtain results for the variance decomposition that relate to two different scenarios: (A) We expect a linear future without mortality shocks or trend changes. In this scenario, we exclude the mortality data of the two world wars from the Lee-Carter model estimation. Because the uncertainty estimated via the original Lee-Carter model is known to be overly optimistic (Lee and Miller 2001; Booth et al. 2006; Li et al. 2009) and because we excluded the war years, we look at a second scenario (B): a future with mortality shocks and trend changes. For this scenario, we use an empirical measure of forecast variance from a cross-validation exercise to estimate the forecast variance component – the Mean Square Error (MSE). The MSE is based on observed forecast errors. Thus, using this empirical measure, overly optimistic variance estimates and increased uncertainty due to extreme mortality events or trend changes are accounted for.

2 Data and methods

2.1 Data

We use data from the open-access Human Mortality Database (Max Planck Institute for Demographic Research, University of California, Berkeley, and French Institute for Demographic Studies 2023, HMD) for our analyses. The HMD provides high-quality population and mortality data. We source male death counts and population exposures by single age for Denmark, England and Wales, France, Iceland, the Netherlands, Norway, and Sweden, for the years 1871 through 2019. The countries were chosen based on long available data series that are necessary for both the cross-validation of the cohort mortality forecasts and the desire for the analysis of multiple cohorts. Further, we perform the variance decomposition for forecasts of remaining life expectancy at different ages (age 0, age 40, and age 80), thus increasing the demand for longer data series. We have chosen to perform our analysis on male mortality to estimate the upper bound of lifespan uncertainty. Male death counts have historically been more volatile due to higher male participation in military conflicts. This volatility increases both the lifetable lifespan variance and the lifetable forecast variance. Similarly, we have included Denmark and Iceland whose mortality trends are more volatile due to smaller population sizes.

Cohort forecast specification

We forecast age specific mortality rates for birth cohorts 1901 through 2020 given the data available during the 30 year period preceding a cohort. Thus, the fitting window $p = 1$ corresponding to birth cohort $c = 1901$ is 1871 through 1900 and the fitting window $p = 120$ informing forecasts for birth cohort $c = 2019$ is 1990 through 2019.

Model specification: We fit a Poisson-Lee-Carter model where death counts at single age x and single year t are distributed $D_{x,t} \sim \text{Pois}(\mu_{x,t}\, E_{x,t})$, with $E_{x,t}$ being the person-years of exposure and $\mu_{x,t}$ the death rates. We use the original Lee-Carter specification $\mu_{x,t} = \exp(a_x + b_x k_t)$ with identifiability constraints $\sum_t k_t = 0$ and $\sum_x b_x = 1$.

Model fit: The model is fit via maximum-likelihood with quadratic penalties on the second differences of the age effects (ages zero to 100). These roughness penalties improve model convergence and ensure that forecasts for neighboring age groups are similar – an important consideration when deriving cohort life

tables from period diagonals (Van Raalte et al. 2023). To further improve the stability of the forecasts, we penalize the squared difference between parameter estimates of two neighboring fitting periods, ensuring that forecasts for neighboring cohorts are similar. A ridge penalty on the age effects completes the set of penalties. The penalized log-likelihood function over parameters $\boldsymbol{\theta}^{(p)} = \{a_x, b_x, k_t : x \in \mathcal{X}, t \in \mathcal{T}^{(p)}\}$ for fitting period p then is

$$\ell(\boldsymbol{\theta}^{(p)}) = \left[\sum_{x\in\mathcal{X}}\sum_{t\in\mathcal{T}^{(p)}} D_{x,t}\log\left(\mu_{x,t}^{(p)}\right) - E_{x,t}\mu_{x,t}^{(p)}\right] - \left(\lambda_1 p_{\text{a}_x\text{roughness}}^{(p)} + \lambda_2 p_{\text{b}_x\text{roughness}}^{(p)} + \lambda_3 p_{\text{deviation}}^{(p)} + \lambda_4 p_{\text{ridge}}^{(p)}\right), \quad (1)$$

where

$$p_{\text{a}_x\text{roughness}}^{(p)} = \sum_{x=1}^{\omega-2}(a_{x+2}^{(p)} - 2a_{x+1}^{(p)} + a_x^{(p)})^2 \quad (2)$$

$$p_{\text{b}_x\text{roughness}}^{(p)} = \sum_{x=1}^{\omega-2}(b_{x+2} - 2b_{x+1} + b_x)^2 \quad (3)$$

$$p_{\text{deviation}}^{(p)} = \sum_{p=2}^{30}\sum_{x=1}^{\omega}(a_x^{(p)} - a_x^{(p-1)})^2 + (b_x^{(p)} - b_x^{(p-1)})^2 \quad (4)$$

$$p_{\text{ridge}}^{(p)} = \sum_{x=1}^{\omega}(a_x^{(p)})^2 + (b_x^{(p)})^2. \quad (5)$$

Parameter uncertainty: We quantify parameter uncertainty via an approximation to the parametric bootstrap where we sample 2,500 $\boldsymbol{\theta}^{(p)}$ parameter vectors from a multivariate Normal distribution with mean equal to the maximum likelihood parameter estimates and covariance equal to the inverse of the corresponding negative Hessian (King et al. 2000; Haberman and Renshaw 2009).

Model specification uncertainty: We chose a penalized Lee-Carter model to forecast the mortality based on extensive sensitivity checks. The results from these model comparisons on the influence of parameter uncertainty and on the importance of model specification can be found in the additional material.

Forecast: In order to forecast $\mu_{x,t}^{(p)}$ past the fitting period, we estimate the drift and the variance of the estimated year-on-year $k_t^{(p)}$ changes and simulate a corresponding random walk with drift over 100 years for each of the 2,500 parameter draws. The resulting forecasts $\hat{k}_t^{(p)}$ for forecast horizon $t \in \mathcal{H}^{(p)}$ are then used to derive a corresponding surface of age-period mortality rates $\hat{\mu}_{x,t}^{(p)} = \exp\left(a_x^{(p)} + b_x^{(p)}\hat{k}_t^{(p)}\right)$.

We convert these period-age forecasts into cohort-age forecasts by extracting the corresponding cohort-diagonals. Let $J^{(p)}$ denote the jump-off year for the forecast corresponding to the fitting window p. For each period-age forecast $\hat{\mu}_{x,t}^{(p)}$ we then extract the forecast mortality rates for those cohorts being infants during the jump-off year, $\hat{\mu}_{x,t=J(p)+x}^{(p)}$, for those cohorts who turned 40 in the jump-off year, $\hat{\mu}_{x,t=J(p)-40+x}^{(p)}$, and for those cohorts who turned 80 in the jump-off year, $\hat{\mu}_{x,t=J(p)-80+x}^{(p)}$. Forecasts are done independently for each country. For a visual representation of the extraction of cohort-diagonals, see the additional material.

The years 1914–1921 and 1939–1945 have been excluded from the estimation of a_x and b_x, and from the k_t drift and variance estimates. This conditions the forecast variance on a future without extreme mortality shocks. For an estimate of the forecast variance allowing for the possibility of extreme shocks, we calculate the mean squared forecast error of our cohort life expectancy forecasts.

Cohort variance decomposition We denote by X the random age at death of a cohort member and with $\boldsymbol{\mu}$ the random vector of age-specific cohort death rates parameterizing the distribution of X. How much of the uncertainty about the individual's eventual lifespan is driven by within-lifetable stochasticity and how much is driven by the uncertainty about which future lifetable will apply to them? The law of total variance, as demonstrated in the context of lifetables by Caswell (2023), allows us to answer this question by decomposing the total variance in the individual's age at death, $\mathrm{Var}(X)$, into variance due to uncertainty in future death rates, $\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})]$, and the expected variance inherent to any lifetable, $\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X|\boldsymbol{\mu})]$:

$$\underbrace{\mathrm{Var}(X)}_{\substack{\text{Total lifespan variance}}} = \underbrace{\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X|\boldsymbol{\mu})]}_{\substack{\text{Within-lifetable variance /}\\ \text{Expected lifespan variance}}} + \underbrace{\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})]}_{\substack{\text{Between-lifetable variance /}\\ \text{Forecast variance}}} . \tag{6}$$

We employ the stochastic cohort lifetable forecasts $\boldsymbol{\mu}_s$ to decompose the total individual uncertainty about the remaining lifespan at birth. For brevity, we write $e_{0_s} = \mathrm{E}(X|\boldsymbol{\mu}_s)$ for the expected age at death (life expectancy) and $\sigma_s^2 = \mathrm{Var}(X|\boldsymbol{\mu}_s)$ for the variance in the age at death (lifespan variance) of a cohort member under mortality rates $\boldsymbol{\mu}_s$. Both quantities are calculated from stochastic cohort death rate forecasts $\boldsymbol{\mu}_s$ using the matrix lifetable equations in Caswell (2023).

For scenario (A), a future without mortality shocks, we calculate the between-lifetable variance – that is, the forecast variance of the cohort life expectancies – as

$$\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})] = \frac{1}{2500-1}\sum_s (e_{0_s} - \overline{e_0})^2, \tag{7}$$

where $\overline{e_0} = \frac{1}{2500}\sum_s e_{0_s}$ is the mean cohort life expectancy forecast as averaged over the simulation runs.

The second part of the variance decomposition, the expected cohort lifespan variance, or within-lifetable variance, is calculated as the mean of the simulated lifespan variances:

$$\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X|\boldsymbol{\mu})] = \frac{1}{2500}\sum_s \sigma_s^2. \tag{8}$$

Empirical forecast error quantification For scenario (B), a future with mortality shocks, we employ the Mean Square Error (MSE). The MSE is an empirical measure of $\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X|\boldsymbol{\mu})]$, the forecast variance component of the decomposition, based on actual forecast errors. This measure contributes to our analysis of sources of uncertainties in past lifespan forecasts. We use this empirical measure to analyse the share of forecast variance under a scenario that allows for the possibility of future extreme mortality or trend changes in the mortality. The MSE is based on the empirical forecast error that compares the observed cohort life expectancy $e_{0_{c,r}}^{\text{observed}}$ as published by the HMD (cohorts born from 1901 through 1932) with the mean Lee-Carter forecasts of cohort life expectancy $\overline{e_{0_{c,r}}}$. For each region r, we pool the error across all cohorts over the simulations s, resulting in the Mean Square Error:

$$\mathrm{MSE}_r = \frac{1}{33-1}\sum_c (e_{0_c}^{\text{observed}} - \overline{e_{0_c}})^2. \tag{9}$$

Due to the cross-validation procedure used for the empirical forecast error quantification, fewer data points are available for calculating the MSE than for the variance estimation from the simulated Lee-Carter forecasts. To increase the robustness of the measure, we therefore assume that the MSE is constant and does not vary across cohorts.

Classification by ages The methodological description above refers to the variance decomposition for cohort life expectancy at birth. In addition, we perform the variance decomposition for remaining cohort life expectancy at ages 40 and 80. We do so because the effect of changes in the death rates on the within-lifetable variance varies with age: it is positive for early ages (i.e., reducing infant mortality decreases lifespan inequality), but is negative for later ages and zero for the highest ages (Van Raalte and Caswell 2013). For the specification of Equations (7), (8) and (9) for different ages of remaining life expectancy, see the additional material. Further, we put emphasis on the results for remaining life expectancy at age 40

because it is invariant to changes in early life mortality and concerns an age at which information about the future lifetime is relevant for the individual when making, for example, economic decisions.

3 Results

3.1 Variance decomposition for remaining life expectancy at age 40

How much of the total individual lifespan variance is due to the variance in the mortality forecasts and how much is due to the variance in the lifetable lifespan dispersion? The decomposition of the total variance allows us to answer this question. Under the scenario of a linear future without mortality shocks (scenario A), with the exception of Iceland, the share of the forecast variance at age 40 is very low (see Figure 1): The highest values of approximately 2.5% are observed for cohorts of the 1860s in Sweden, Norway, France, and Denmark.

We can observe reductions in the share of forecast variance to overall lifespan variance over cohorts. For all countries, the lowest share is observed for the most recent cohorts with less than 0.2%, Iceland excluded. The decline over cohorts can be explained by the increasing stability of the mortality rates that allows for better forecasts. As a result, the contribution of forecast variance to the total lifespan variance decreases.

Iceland is the smallest country in our selection with roughly 390,000 inhabitants in 2025, compared to several million inhabitants in the other countries. In smaller populations, mortality trends are more volatile and thus harder to forecast. This is reflected in the much higher share of forecast variance for Iceland. Starting around a contribution of 15% for the cohorts of the 1860s, the trend decreases drastically after the cohort 1979 until the latest cohort of 1980 for which the share of forecast variance is 5.5%.

Figure 1 also includes the country-specific Mean Square Error (MSE) which is an empirical measure of the forecast variance and thus, takes the possibility for extreme mortality events into account (scenario B). The comparison of the average MSE across cohorts with the share of the forecast variance to the total lifespan variance from scenario (A) without shocks reveals different results depending on country. For Denmark, England and Wales, France, and Sweden, the MSE is very similar to the share of forecast variance. For Norway (MSE = 7%) and the Netherlands (MSE = 6%), the empirical measure is slightly higher than the share of forecast variance under the scenario with no mortality shocks. Regarding Iceland, the empirical forecast error is similar to the results for the earliest cohorts. With decreasing share of forecast variance for later cohorts, the gap between the two measures increases.

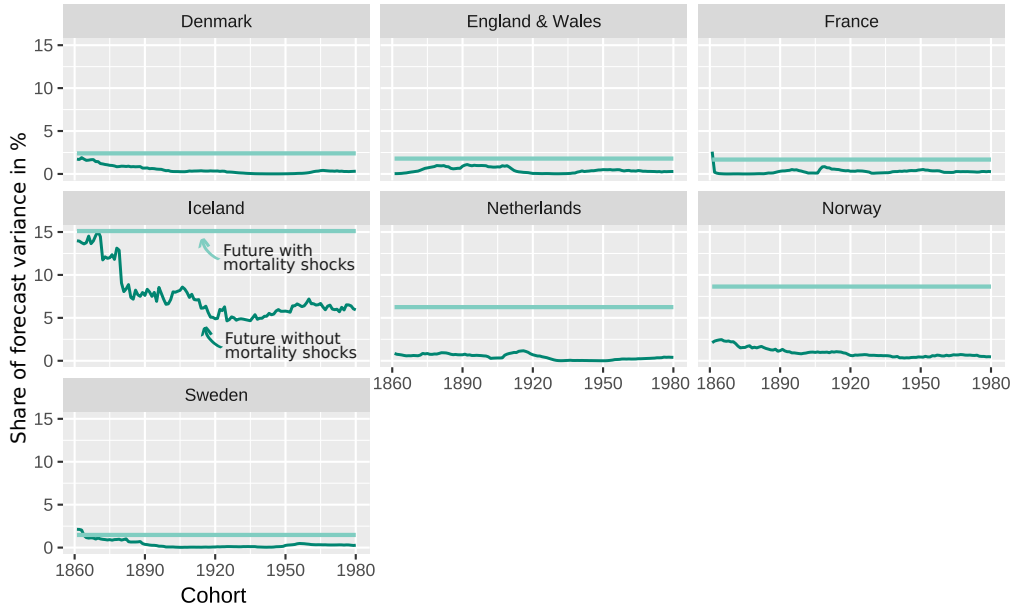
Figure 2 shows the total cohort lifespan variance at age 40 across cohorts (variance on left axis, standard deviation on right axis) split up into the two variance components from the variance decomposition under the linear mortality scenario. The changes in the total lifespan variance over cohorts are clearly driven by a reduction in the lifetable lifespan variance. For most of the countries, there is a slight decline in the total lifespan variance starting with the cohorts of the late 1800s, reaching a level below 12 years of standard deviation for cohort 1980. For the cohorts before 1900, the total lifespan variance is fluctuating between a standard deviation of 12 to 15 years, and 12 to 22 years for Iceland. With the exception of Iceland, for which declines in the forecast variance were responsible for stronger reductions in the total lifespan variance, the contribution of forecast variance to the change in total lifespan variance is small.

3.2 Variance decomposition for life expectancy at birth

We also performed the variance decomposition for male lifespan variance at birth (see Figures 3 and 4). Compared to the results for age 40, male lifespan variance at birth has a similar low share of forecast variance to total lifespan variance. The empirical error measure is slightly higher than the forecast variance component under scenario A without mortality shocks. Iceland has a more volatile pattern of the forecast component at a higher level, compared to the other countries, similar to the findings for age 40. Further, the empirical forecast uncertainty for France is much higher at 30%.

The total lifespan variance at age zero (see Figure 4) of all countries is on a higher level and declines across cohorts starting with the earliest cohorts of the nineteen hundreds. This can be attributed to consistent improvements in early life mortality. The change in total lifespan variance is mainly driven by reductions in the lifetable lifespan variance.

Figure 1: Share of forecast variance on total cohort lifespan variance at age 40 under a scenario without future mortality shocks (dark green) and under a scenario with mortality shocks (light green) among male birth cohorts of seven European countries.



Note: The estimate of the share of forecast variance under the scenario with mortality shocks is constant across cohorts. It is based on pooling the forecast errors across all cohorts and simulations.

3.3 Age profile of variance decomposition

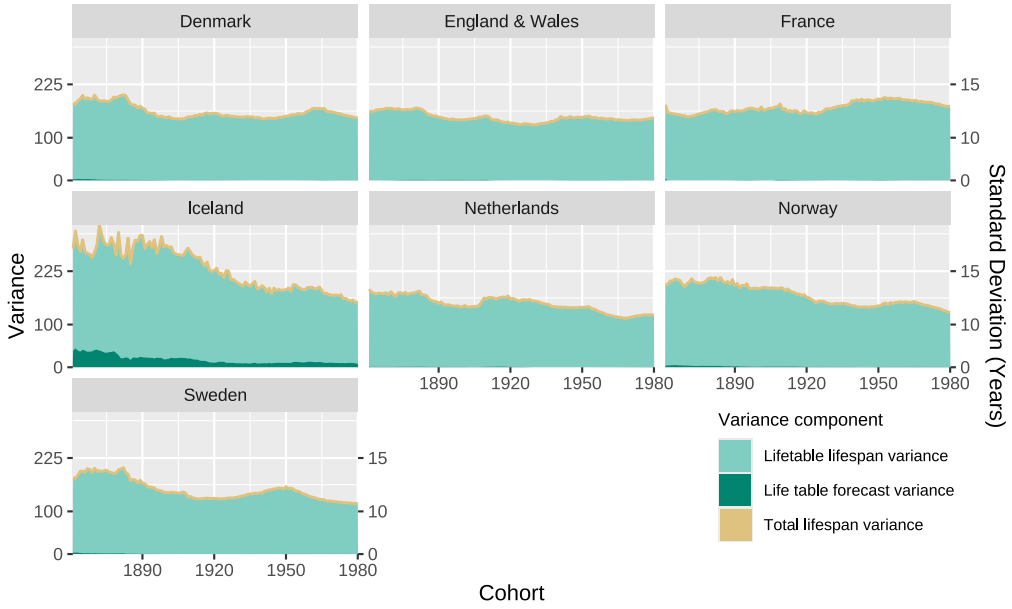
Figure 5 shows the share of forecast variance for three different cohorts for life expectancy at birth, age 40 and age 80 for a future without mortality shocks. In general, the importance of the forecast variance becomes even less important the shorter the remaining lifespan is. As a cohort ages, less of their lifetable needs to be completed to forecast the remaining life expectancy. We have seen that the total cohort lifespan variance is at a lower level for remaining life expectancy at age 40 compared to life expectancy at birth. In addition, the share of forecast variance on the total variance decreases with increasing age for which the remaining life expectancy is forecast.

4 Discussion

Under the expectation of a future with no extreme volatility in mortality, due to, for example, wars or pandemics, the results of the variance decomposition show that the overall contribution of forecast variance, or between-lifetable variance, to the total lifespan variance is low. Thus, if one does not expect crises that cause mortality shocks, the forecast variance is negligible for the overall lifespan variance. When the possibility for extreme mortality is included, as shown by the empirical measure of forecast variance, uncertain future mortality plays a larger, but in most cases, still minor role.

Lifespan variability differs by age (Singh and Kim 2021). While lifespan variability at birth has decreased due to declines in infant mortality, lifespan variability at older ages has increased because of improvements in adult survival (Engelman et al. 2014). Our results show that with increasing age of a cohort, the forecast variance contributes less to total cohort lifespan variance, due to the shorter forecasts needed to complete

Figure 2: Variance components of total cohort lifespan variance at age 40 among male birth cohorts in seven European countries.

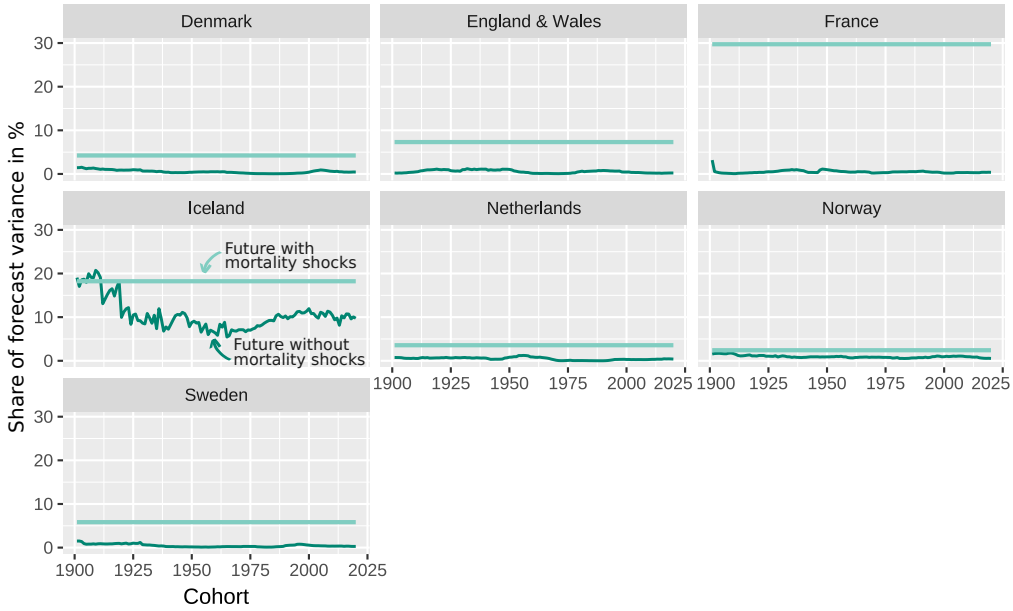


the cohort lifetables.

The small contribution of forecast uncertainty to lifespan uncertainty may come as a surprise. Yet, it is consistent with the approximate relative magnitudes of both quantities as regularly reported in lifespan variance calculations and forecast uncertainties. Caswell (2023) reports a lifetable lifespan variance for Swedish females in 2007 of around 100 – that is, a standard deviation of 10 years. Accordingly, forecast variance around life expectancy at birth would need to be 100 as well in order to contribute half to the total lifespan variance. This is equal to a standard deviation of forecast error of 10 years which results in 95% Normal prediction intervals with a width of 2*1.96*10=39.2 years. The World Population Prospects (United Nations, Department of Economic and Social Affairs, Population Division 2024) state a 95% prediction interval of 100.98-86.30=14.68 years for period life expectancy at birth of Swedish Females in 2100. While the cohort case differs from the period case, a forecast variance high enough to match the importance of life table variance in lifespan uncertainty would be much wider than commonly reported in the demographic literature.

The life expectancy trajectories of the seven countries analyzed are characterized by sustained mortality improvements, resulting in increasing life expectancy and declining life span disparity that have changed in lockstep over time. However, particularly countries in Central and Eastern Europe have exhibited unstable mortality patterns since the 1960s. In these countries, mortality was marked by patterns that let life expectancy and lifespan disparity change independently from each other (Aburto and van Raalte 2018). This was mainly due to divergent trends of mortality change in different age groups (e.g., mortality improvements at early and old ages alongside rising mortality at mid-ages). This raises the question of whether our results are transferable to such a mortality context. Unstable mortality trajectories with, for example, stalling life expectancy and diverging mortality development for different age groups would overall increase the total lifespan uncertainty an individual has about their age at death. Both the lifetable lifespan variance and the forecast variance would be increased, compared to a mortality regime with stable improvements. Especially when a previously stable mortality period is used as the basis for the forecasts, the forecast variance would

Figure 3: Share of forecast variance on total cohort lifespan variance at birth under a scenarios without future mortality shocks (dark green) and under a scenario with mortality shocks (light green) among male birth cohorts of seven European countries.



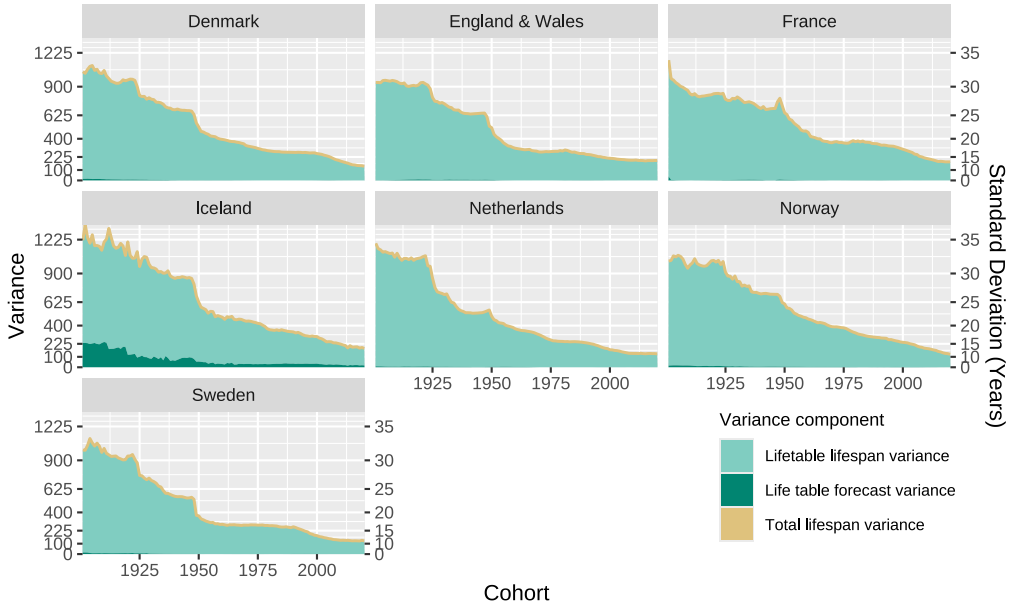
Note: The estimate of the share of forecast variance under the scenario with mortality shocks is constant across cohorts. It is based on pooling the forecast errors across all cohorts and simulations.

be higher. However, the mortality experience of the cohorts 1901 through 1932 which informs our forecast error quantification was also marked by non-linearities and disruptions. Still, we have not identified a case where the forecast variance dominates the total lifespan variance even after accounting for the possibility of wars and epidemics.

There are sources of variance that we could not account for in the decomposition of the total cohort lifespan variance. First, we assume complete homogeneity in the population – that is, everyone is exposed to the same mortality rates. However, lifespan inequality differs between several socioeconomic strata, such as education, ethnicity, and region (Permanyer et al. 2018; Sasson 2016; Brown et al. 2012). Further, the true uncertainty an individual has about their lifespan is also influenced by information that is highly heterogeneous and not accessible to us. For example, individual medical histories and lifestyle factors play a role in the overall uncertainty (or certainty) about one's age at death. However, Badolato et al. (2023) find that the amount of accuracy gained by including extensive information on the individual's life when predicting individual lifespan is not substantial. Caswell (2023) shows that even with extreme levels of heterogeneity in the population, only a few percent of variance in the length of life can be attributed to inequality between groups. Instead, it is the within-lifetable variance that explains the differences in the outcomes, which is in line with our findings from the variance decomposition. Similarly, Edwards (2013) comes to the conclusion that the forecast uncertainty from Lee-Carter forecasts is small compared to the lifetable uncertainty when looking at the standard deviation in length of life above age 10 for cohorts.

Another source of variance is the likelihood of extreme events, such as wars and pandemics. The likelihood of such events plays a role in the forecast variance component of the variance decomposition. While we take these outliers into account via the empirical forecast error, future research could explore the methodology of conformal prediction (Duerst and Schöley 2024; Angelopoulos et al. 2024; Fontana et al. 2023; Shafer and

Figure 4: Variance components of total cohort lifespan variance at birth among male birth cohorts in seven European countries.

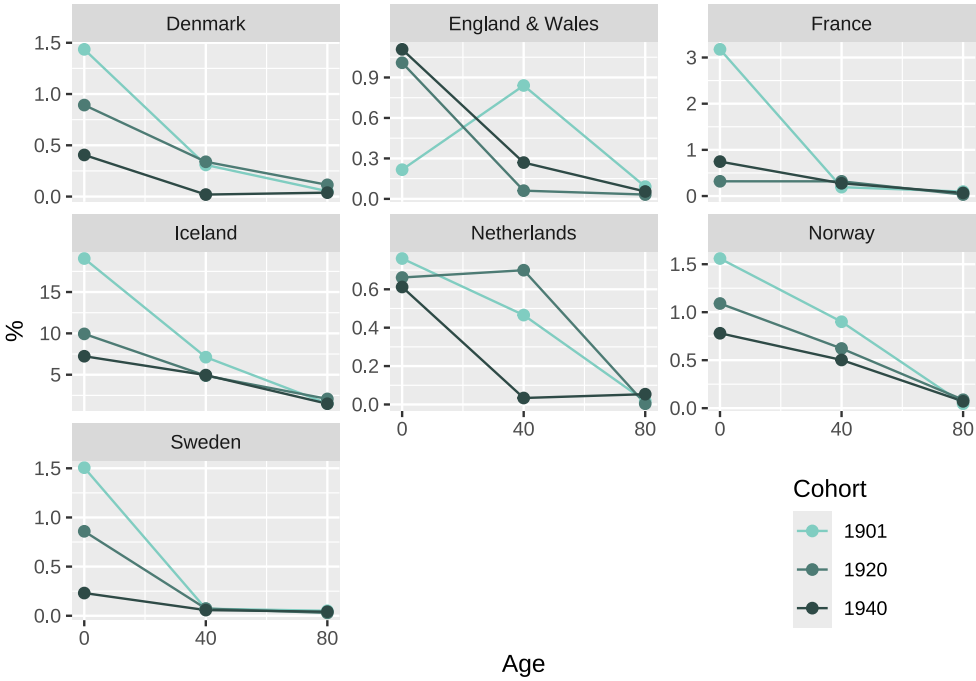


Vovk 2008) to assess this type of forecast variance using the accuracy of historical forecasts. The extreme value theory is another methodological branch worth investigating in this matter (Coles et al. 2001, for an introduction). Medford (2017) argues that their research on forecasting best practice life expectancy using extreme value theory could provide valuable insights on the likelihood of life expectancy crises. Similarly, a stochastic model of extreme value mortality may be used to obtain more realistic simulations of the forecast time-varying mortality trend term k_t of the Lee-Carter model. Wang et al. (2011) recommend the use of heavy-tailed distributions for the residuals of the Lee-Carter model and the first difference of mortality indices. This approach accounts for the possibility of extreme outliers as witnessed during, for example, wars.

5 Conclusion

Forecast variance only contributes a minor share to overall lifespan variance, under the assumption that mortality develops without extreme and prolonged period shocks. Likewise, the decline in total lifespan variance observed over more than 100 years is mostly driven by a change in the lifetable lifespan variance, and not by the drop in forecast variance. When allowing for future mortality shocks, the forecast variance still has a low share to the overall lifespan variance, with the exception of Iceland and France. Thus, these findings generally support the interpretation of lifetable lifespan variability as an indicator of lifespan uncertainty. To strengthen the case that the lifetable variance is a good proxy for individual lifespan uncertainty, other components contributing to an individual's knowledge about their eventual lifespan should be analyzed.

Figure 5: Share of total cohort lifespan variance over age explained by variance across forecast lifetables among selected male birth cohorts.



Acknowledgments

R.D. gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions, which have greatly improved the quality of this paper.

Availability of data and materials

The research presented in this article is fully reproducible. We have made our codebase available to the scientific community. You can access the complete codebase, including scripts and documentation, on the website of Demographic Research alongside the publication. In addition, the data used for our analyses is sourced from openly accessible and publicly available datasets. Researchers interested in reproducing our work or conducting further investigations can freely download the data from the following link: https://www.mortality.org. The availability of open-access data ensures transparency and promotes collaboration within the scientific community.

References

Aburto, J. M. and van Raalte, A. (2018). Lifespan dispersion in times of life expectancy fluctuation: the case of central and eastern europe. *Demography*, 55(6):2071–2096.

Aburto, J. M., Villavicencio, F., Basellini, U., Kjærgaard, S., and Vaupel, J. W. (2020). Dynamics of life expectancy and life span equality. *Proceedings of the National Academy of Sciences*, 117(10):5250–5259.

Angelopoulos, A., Candes, E., and Tibshirani, R. J. (2024). Conformal pid control for time series prediction. *Advances in neural information processing systems*, 36.

Badolato, L., Decter-Frain, A., Irons, N. J., Miranda, M., Walk, E., Zhalieva, E., Alexander, M., Basellini, U., and Zagheni, E. (2023). The limits of predicting individual-level longevity. Technical report, Max-Planck-Inst. für demografische Forschung.

Booth, H., Hyndman, R. J., Tickle, L., and De Jong, P. (2006). Lee-carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic research*, 15:289–310.

Brown, D. C., Hayward, M. D., Montez, J. K., Hummer, R. A., Chiu, C.-T., and Hidajat, M. M. (2012). The significance of education for mortality compression in the united states. *Demography*, 49(3):819–840.

Caswell, H. (2023). The contributions of stochastic demography and social inequality to lifespan variability. *Demographic Research*, 49:309–354.

Colchero, F., Rau, R., Jones, O. R., Barthold, J. A., Conde, D. A., Lenart, A., Nemeth, L., Scheuerlein, A., Schoeley, J., Torres, C., et al. (2016). The emergence of longevous populations. *Proceedings of the National Academy of Sciences*, 113(48):E7681–E7690.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

Duerst, R. and Schöley, J. (2024). Empirical prediction intervals applied to short term mortality forecasts and excess deaths. *Population Health Metrics*, 22(1):1–15.

Edwards, R. D. (2013). The cost of uncertain life span. *Journal of population economics*, 26(4):1485–1522.

Edwards, R. D. and Tuljapurkar, S. (2005). Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Population and Development Review*, 31(4):645–674.

Engelman, M., Caswell, H., and Agree, E. M. (2014). Why do lifespan variability trends for the young and old diverge? a perturbation analysis. *Demographic Research*, 30:1367.

Fontana, M., Zeni, G., and Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23.

Ford, J., Steel, N., Aasheim, E., Devleesschauwer, B., Gallay, A., Morgan, D., Schmidt, J., Ziese, T., and Newton, J. (2019). Slowing improvements in life expectancy across european economic area countries. *European Journal of Public Health*, 29(Supplement_4):ckz185–160.

Haberman, S. and Renshaw, A. E. (2009). Measurement of Longevity Risk Using Bootstrapping for Lee–Carter and Generalised Linear Poisson Models of Mortality. *Methodology and Computing in Applied Probability*, 11(3):443–461.

Hum, R. J., Verguet, S., Cheng, Y.-L., McGahan, A. M., and Jha, P. (2015). Are global and regional improvements in life expectancy and in child, adult and senior survival slowing? *PLoS One*, 10(5):e0124479.

Hurd, M. D., Smith, J. P., and Zissimopoulos, J. M. (2004). The effects of subjective survival on retirement and social security claiming. *Journal of Applied Econometrics*, 19(6):761–775.

King, G., Tomz, M., and Wittenberg, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science*, 44(2):347.

Lee, R. and Miller, T. (2001). Evaluating the performance of the lee-carter method for forecasting mortality. *Demography*, 38(4):537–549.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Li, J. S.-H., Hardy, M. R., and Tan, K. S. (2009). Uncertainty in mortality forecasting: an extension to the classical lee-carter approach. *ASTIN Bulletin: The Journal of the IAA*, 39(1):137–164.

Max Planck Institute for Demographic Research, University of California, Berkeley, and French Institute for Demographic Studies (2023). Human mortality database.

Medford, A. (2017). Best-practice life expectancy: An extreme value appr. *Demographic Research*, 36:989–1014.

Murphy, M. (2021). Recent mortality in britain: a review of trends and explanations. *Age and ageing*, 50(3):676–683.

Nepomuceno, M. R., Cui, Q., Van Raalte, A., Aburto, J. M., and Canudas-Romo, V. (2022). The cross-sectional average inequality in lifespan (cal†): A lifespan variation measure that reflects the mortality histories of cohorts. *Demography*, 59(1):187–206.

Permanyer, I., Spijker, J., Blanes, A., and Renteria, E. (2018). Longevity and lifespan variation by educational attainment in spain: 1960–2015. *Demography*, 55(6):2045–2070.

Sasson, I. (2016). Trends in life expectancy and lifespan variation by educational attainment: United states, 1990–2010. *Demography*, 53(2):269–293.

Schöley, J., Aburto, J. M., Kashnitsky, I., Kniffka, M. S., Zhang, L., Jaadla, H., Dowd, J. B., and Kashyap, R. (2022). Life expectancy changes since covid-19. *Nature human behaviour*, 6(12):1649–1659.

Seligman, B., Greenberg, G., and Tuljapurkar, S. (2016). Equity and length of lifespan are not the same. *Proceedings of the National Academy of Sciences*, 113(30):8420–8423.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.

Singh, A. and Kim, Y. (2021). What has contributed to the large sex differentials in lifespan variation and life expectancy in south korea? *Journal of Biosocial Science*, 53(3):396–406.

United Nations, Department of Economic and Social Affairs, Population Division (2024). World population prospects 2024: Summary of results. Technical report, United Nations.

Van Raalte, A. A., Basellini, U., Camarda, C. G., Nepomuceno, M. R., and Myrskylä, M. (2023). The Dangers of Drawing Cohort Profiles From Period Data: A Research Note. *Demography*, 60(6):1689–1698.

Van Raalte, A. A. and Caswell, H. (2013). Perturbation analysis of indices of lifespan variability. *Demography*, 50(5):1615–1640.

Van Raalte, A. A., Sasson, I., and Martikainen, P. (2018). The case for monitoring life-span inequality. *Science*, 362(6418):1002–1004.

Vaupel, J. W., Zhang, Z., and van Raalte, A. A. (2011). Life expectancy and disparity: an international comparison of life table data. *BMJ open*, 1(1):e000128.

Wang, C.-W., Huang, H.-C., and Liu, I.-C. (2011). A quantitative comparison of the lee-carter model under different types of non-gaussian innovations. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36(4):675–696.

Wilson, B., Drefahl, S., Sasson, I., Henery, P. M., and Uggla, C. (2020). Regional trajectories in life expectancy and lifespan variation: Persistent inequality in two nordic welfare states. *Population, Space and Place*, 26(8):e2378.

Additional Material for:
The contribution of forecast uncertainty to lifespan uncertainty

Authors: Ricarda Duerst[1,2], Jonas Schöley[1]

Affiliations:
[1] Max Planck Institute for Demographic Research; Rostock, Germany
[2] Universits of Helsinki; Finland

Conversion of period-age mortality forecasts into cohort-age forecasts

Figure 1 is a visual representation of the conversion of the period-age forecasts into cohort-age forecasts. On the example of the jump-off year 1901, we extract the forecast mortality rates from the cohort-diagonals for those cohorts being born, for those turning 40, and for those turning 80 in 1901.

Cohort variance decomposition: Specification for different ages

We perform the variance decomposition for remaining cohort life expectancy at ages $x = \{0, 40, 80\}$. We denote by X the random age at death of a cohort member and with $\boldsymbol{\mu}$ the random vector of age-specific cohort death rates parameterizing the distribution of X. We employ the stochastic cohort lifetable forecasts $\boldsymbol{\mu}_s$ to decompose the total individual uncertainty about the remaining lifespan. For brevity, we write $e_{x_s} = \mathrm{E}(X - x|X > x, \boldsymbol{\mu}_s)$ for the remaining life expectancy and $\sigma^2_{x_s} = \mathrm{Var}(X - x|X > x, \boldsymbol{\mu}_s)$ for the remaining lifespan variance at age x of a cohort member under mortality rates $\boldsymbol{\mu}_s$. For scenario A), a future without mortality shocks, we calculate the between-lifetable variance, that is the forecast variance of the remaining cohort life expectancies, as

$$\mathrm{Var}_{\boldsymbol{\mu}}[\mathrm{E}(X - x|X > x, \boldsymbol{\mu})] = \frac{1}{500 - 1} \sum_s (e_{x_s} - \overline{e_x})^2, \tag{1}$$

where $\overline{e_x} = \frac{1}{500} \sum_s e_{x_s}$ is the mean remaining cohort life expectancy forecast as averaged over the simulation runs.

The second part of the variance decomposition, the expected cohort lifespan variance, or within-lifetable variance, is calculated as the mean of the simulated lifespan variances:

$$\mathrm{E}_{\boldsymbol{\mu}}[\mathrm{Var}(X - x|X > x, \boldsymbol{\mu})] = \frac{1}{500} \sum_s \sigma^2_{x_s}. \tag{2}$$

For scenario B), a future with mortality shocks, we employ the Mean Square Error (MSE). The MSE is based on the empirical forecast error that compares the observed remaining cohort life expectancy $e^{\text{observed}}_{x_{c,r}}$ at ages x as published by the HMD (cohorts born from 1901 through 1932) with the mean Lee-Carter forecasts of remaining cohort life expectancy $\overline{e_{x_{c,r}}}$. For each region r we pool the error across all cohorts over the simulations s, resulting in the Mean Square Error:

$$\mathrm{MSE}_r = \frac{1}{33 - 1} \sum_c (e^{\text{observed}}_{x_c} - \overline{e_{x_c}})^2. \tag{3}$$

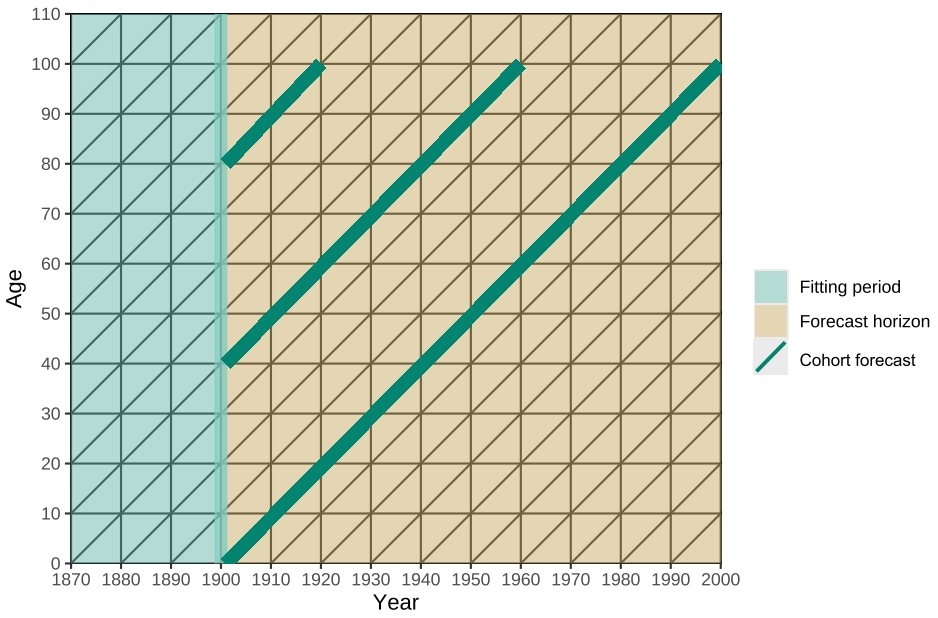


Figure 1: Visual representation of the extraction of cohort-diagonals from period mortality forecasts for the jump-off year 1901.

Influence of parameter variance on forecast variance

The contribution of parameter uncertainty to the share of forecast variance on total lifespan variance is modest. While the inclusion of parameter uncertainty in all cases increases the forecast variance, the increase does not change the interpretation of the results. Parameter uncertainty is most relevant for the small population of Iceland, where its inclusion changes the forecast contribution to lifespan variance from around 6 to 10% by 2020. We also assess the sensitivity of our results to the inclusion of parameter uncertainty for the non-penalized Lee-Carter and find that without penalties the covariance matrix has a higher chance of being ill-conditioned, making sampling impossible for some cohorts. Where parameter sampling was feasible, the impact on our results was small as well.

See Figure 3 for an example illustrating the estimated parameter uncertainty. In the penalized Lee-Carter model the variance of the estimates is concentrated in the k_t parameters. The uncertainty about the k_t (and to a lesser degree a_x and b_x) then propagates through to the drift and standard deviation of the k_t random walk.

Importance of model specification

To test the sensitivity of our results to alternative forecast model specifications we fit the Lee-Carter model without penalties as well as the Renshaw-Haberman (Haberman and Renshaw 2009) model which extends the Lee-Carter model by including an additional cohort term. Cohort life expectancy forecasts were substantially more erratic and generally higher for the Renshaw-Haberman model compared to the (Penalized)-Lee-Carter specification. The failure of the Renshaw-Haberman model to converge at all for Iceland indicates the increased demands for data to identify the added cohort effects, a demand that was not met by our 30 year fitting period. The penalized and non-penalized versions of the Lee-Carter model are mostly consistent in their forecasts with the penalized specification being less prone to erratic changes. As expected, the

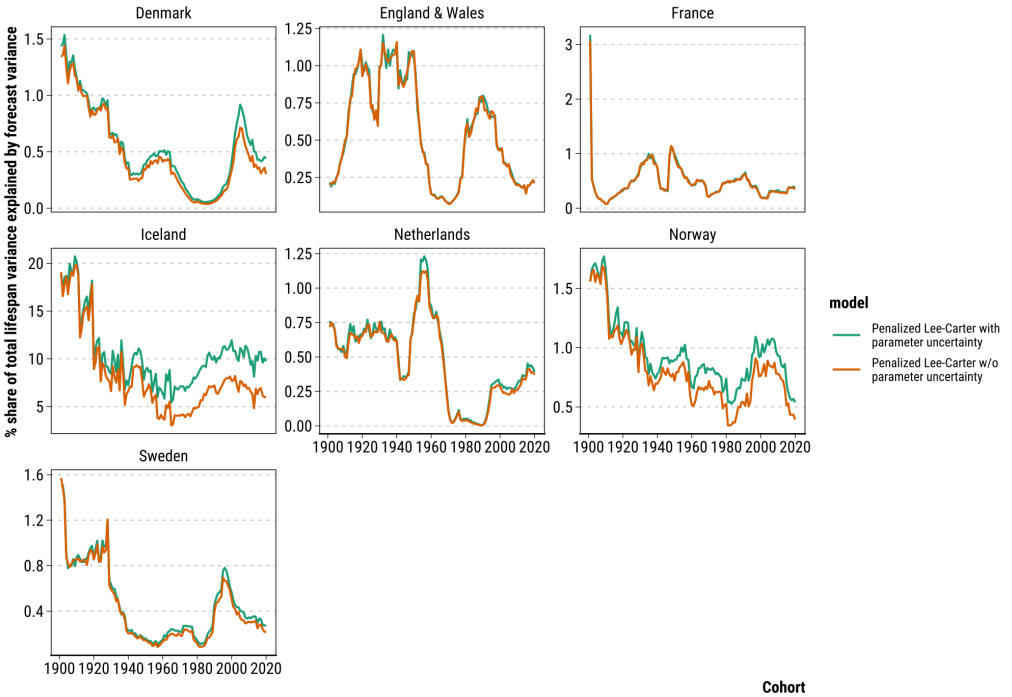


Figure 2: Forecast variance contribution to total lifespan variance with and without inclusion of forecast parameter uncertainty.

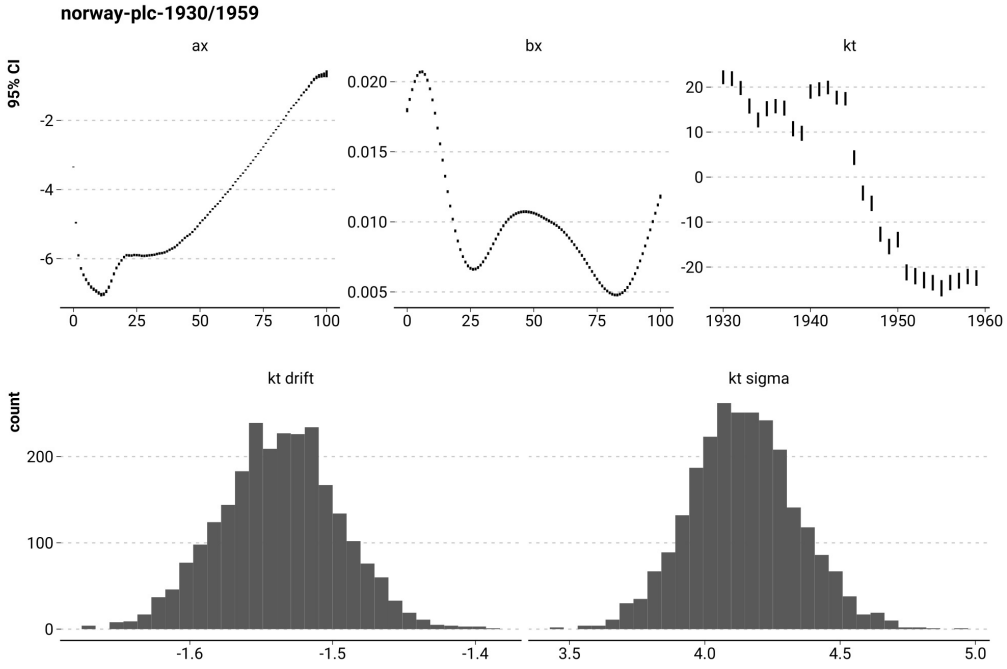


Figure 3: Distribution of penalized Lee-Carter model parameters for Norway fitting years 1930 through 1959.

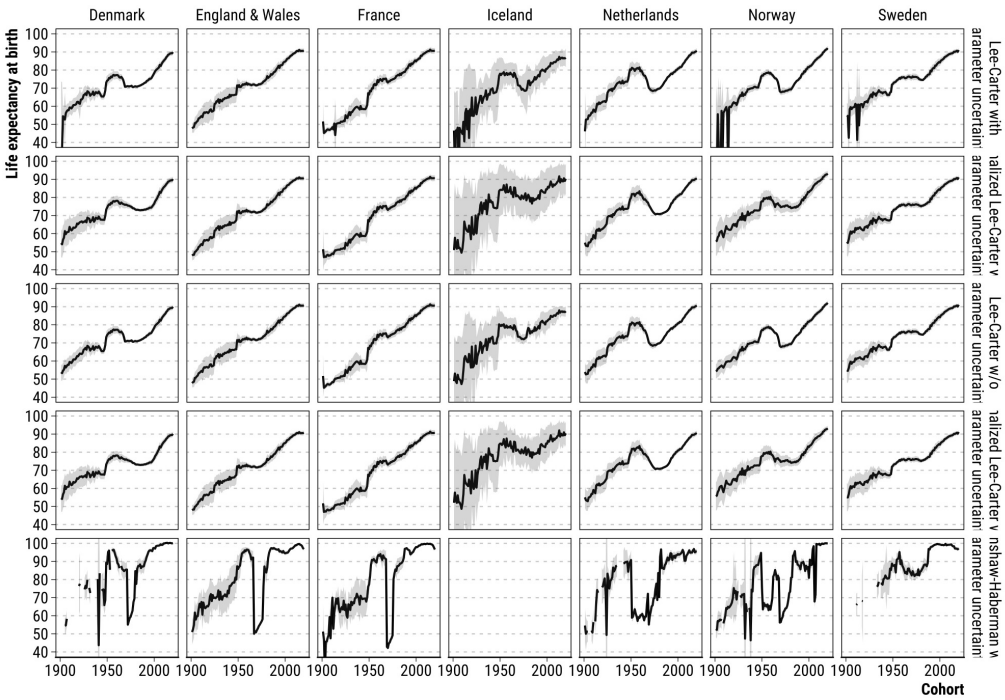
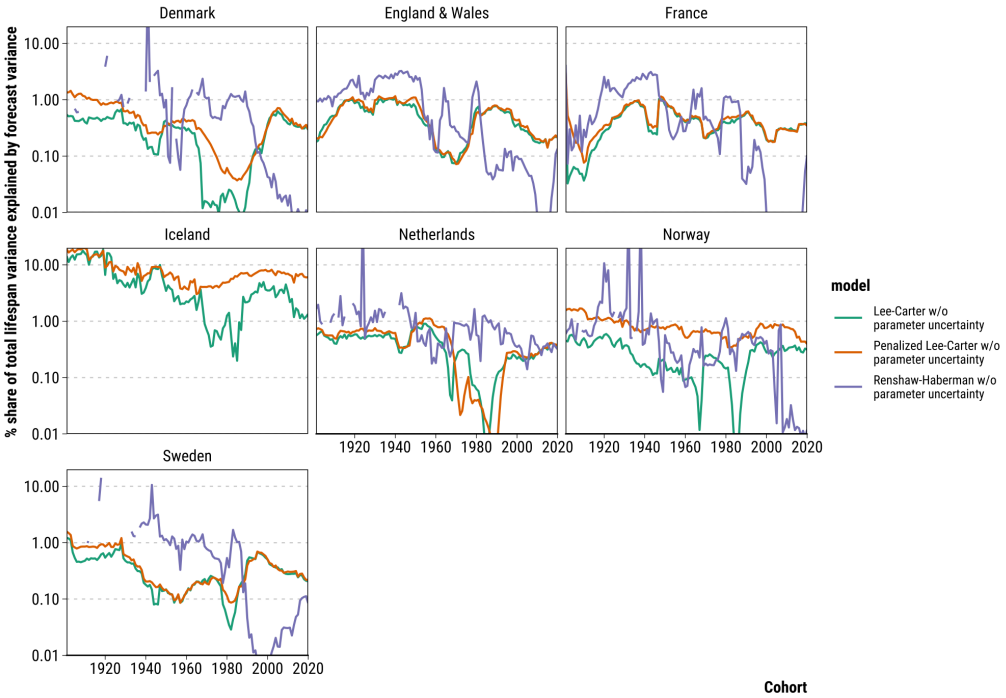


Figure 4: Cohort life expectancy forecasts 1901–2020 by model type. Grey areas indicate 95% prediction intervals.

prediction intervals are slightly wider when parameter uncertainty is reflected in the forecast.

The main outcome of our analysis, the contribution of forecast variance to total lifespan variance, is largely consistent between the non-penalized and the penalized Lee-Carter specifications with the penalized version being less prone to sudden changes in trend over cohorts (see Figure 5). The Renshaw-Haberman model deviates from the Lee-Carter derived estimates with respect to the pattern over cohorts, but not in the overall magnitude of the forecast variance contribution – even for the badly specified Renshaw-Haberman model the forecast contribution to lifespan variance remains low, around 1-2%, with the exception of Iceland.

Figure 5: Forecast variance contribution to total lifespan variance by model type.



References

Haberman, S. and Renshaw, A. E. (2009). Measurement of Longevity Risk Using Bootstrapping for Lee–Carter and Generalised Linear Poisson Models of Mortality. *Methodology and Computing in Applied Probability*, 11(3):443–461.

2

Empirical prediction intervals applied to short term mortality forecasts and excess deaths

Authors: Ricarda Duerst[1,2], Jonas Schöley[1]

Affiliations:
[1] Max Planck Institute for Demographic Research; Rostock, Germany
[2] University of Helsinki; Finland

Abstract

Background
In the winter of 2022/2023, excess death estimates for Germany indicated a 10% elevation, which has led to questions about the significance of this increase in mortality. Given the inherent errors in demographic forecasting, the reliability of estimating a 10% deviation is questionable. This research addresses this issue by analyzing the error distribution in forecasts of weekly deaths. By deriving empirical prediction intervals, we provide a more accurate probabilistic study of weekly expected and excess deaths compared to the use of conventional parametric intervals.

Methods
Using weekly death data from the Short-term Mortality Database (STMF) for 23 countries, we propose empirical prediction intervals based on the distribution of past out-of-sample forecasting errors for the study of weekly expected and excess deaths. Instead of relying on the suitability of parametric assumptions or the magnitude of errors over the fitting period, empirical prediction intervals reflect the intuitive notion that a forecast is only as precise as similar forecasts in the past turned out to be. We compare the probabilistic calibration of empirical skew-normal prediction intervals with conventional parametric prediction intervals from a negative binomial GAM in an out-of-sample setting. Further, we use the empirical prediction intervals to quantify the probability of detecting 10% excess deaths in a given week, given pre-pandemic mortality trends.

Results
The cross-country analysis shows that the empirical skew-normal prediction intervals are overall better calibrated than the conventional parametric prediction intervals. Further, the choice of prediction interval significantly affects the severity of an excess death estimate. The empirical prediction intervals reveal that the likelihood of exceeding a 10% threshold of excess deaths varies by season. Across the 23 countries studied, finding at least 10% weekly excess deaths in a single week during summer or winter is not very unusual under non-pandemic conditions. These results contrast sharply with those derived using a standard negative binomial GAM.

Conclusion
Our results highlight the importance of well-calibrated prediction intervals that account for the naturally occurring seasonal uncertainty in mortality forecasting. Empirical prediction intervals provide a better performing solution for estimating forecast uncertainty in the analyses of excess deaths compared to conventional parametric intervals.

Key words
Excess deaths; COVID-19; cross-validation; robustness; empirical prediction intervals

1 Introduction

In the winter 2022/2023 excess death estimates for Germany were hovering around 10%. Back then, the authors have been contacted by journalists inquiring about the significance of the elevated mortality. Given that statistical estimation always comes with errors attached, can one even reliably estimate a 10% deviation from the norm? In this paper we aim to answer this question via the careful analysis of the error distribution in forecasts of weekly deaths and the seasonality of fluctuations in weekly death counts. Derived from the distribution of errors we propose empirical prediction intervals for the probabilistic study of weekly expected and excess deaths and demonstrate the superior coverage and generality of these intervals compared with conventional parametric intervals. We employ these empirical prediction intervals to quantify, for a range of countries, the probability of observing at least 10% excess deaths in a given week, given the continuation of mortality trends observed prior to the COVID-19 pandemic. Using these p-values we can assess, on a per-country basis, how unusual a 10% mortality increase over the expectation is and whether it should be cause for concern.

Mortality forecasts on a sub-annual timescale have gained relevance as the basis for excess deaths calculations during the COVID-19 pandemic (e.g., Kontis et al. 2020; Karlinsky and Kobak 2021; Aburto et al. 2021). Framed as a forecasting problem, one aims to predict the weekly deaths which would have happened without COVID-19 by forecasting deaths over the pandemic period based on pre-pandemic trends. Those forecast "expected deaths" are associated with an error which can be expressed as a "prediction interval" within which the true expected deaths are to be found with a given probability. It is well known that prediction intervals around forecast values tend to be too narrow (e.g., Chatfield 1993, 1995). In the context of COVID-19 excess death modeling, this phenomenon may lead to wrong conclusions regarding the impact of the pandemic on population mortality, by giving an overly optimistic picture on how precise one can actually forecasts counterfactual weekly expected deaths deaths more than three years since the onset COVID-19. Precisely, given the time passed since the beginning of the pandemic, we expect the uncertainty of forecast expected deaths to have increased. Thus, one might (or should) ask whether we can still, in fall of 2022, reliably detect an, e.g., 10% increase in excess deaths or whether it disappears into the uncertainty of the expected deaths forecasts.

To answer this question, we need reliable measures of forecast uncertainty. It is common practice in demographic forecasting to use parametric prediction intervals, derived from the model structure, such as the random walk over the mortality index of a Lee-Carter model (Lee and Carter 1992), or the Poisson/Negative-Binomial variation around weekly death counts (Aburto et al. 2021; Msemburi et al. 2023). However, of the various sources of error that can contribute to the overall forecast uncertainty (see Keilman 1990, for an overview), these parametric prediction intervals do not reflect errors from the out-of-sample generalization of the model or inaccurate model specification. This is particularly problematic as the out-of-sample generalization error is the most common type of forecast error (Keilman 1990). Therefore, the parametric intervals only work if the model is correctly specified and the data generation process does not change over time. As a result, parametric intervals tend to be too narrow.

Empirical prediction intervals, instead of relying on correct parametric specifications, estimate the distribution of error around a forecast value from actual past, out-of-sample forecasting errors. The idea is simple: The error distribution for future forecasts will be similar to the error distribution of past forecasts. Thus, one seeks to estimate the full distribution of forecast error indexed over all desired forecasting strata and time points. In order to do so, a statistical model is fitted to known out-of-sample forecasting errors. This error distribution can then be used to construct empirical prediction intervals around the central forecast.

Research on empirical prediction intervals has been ongoing for at least half a century (Williams and Goodman 1971). The 1980s saw active demographic research on empirical intervals and the authors generally found that the intervals based on past errors had better coverage than model-based intervals (e.g., Stoto 1983; Cohen 1986; Smith and Sincich 1988). Subsequent research used the used the analysis of forecasting errors to calibrate prediction intervals for demographic forecasts. A summary of methodological advances (e.g., Keilman 1990, 2001; Lee and Scholtes 2014) and demonstrations of the application of the method can also be found in more recent demographic literature (e.g., Keilman et al. 2002; Keilman and Pham 2004b,a; Rayer et al. 2009).

In the machine learning literature, there is growing interest in "distribution free uncertainty quantification", given that popular techniques do not provide an intrinsic estimate of uncertainty. The community

has build a theory on empirical prediction intervals under the term "conformal prediction". Here too, the idea is to look at the distribution of historical error measures to assess how likely any given future deviation from the central prediction would be (Shafer and Vovk 2008).

In this paper, we demonstrate the construction of empirical prediction intervals for short-term mortality forecasts. We validate these intervals against parametric approaches and interpret the results in the context of COVID-19 excess death estimation across 23 European countries. Further, we compare the calibration of the empirical prediction intervals with a conventional model of parametric prediction intervals. The data on weekly deaths for these countries are sourced from the Short-term Mortality Fluctuations Database (STMF, Jdanov et al. 2021). Finally, we determine whether we can reliably detect a 10% increase in excess deaths in the fall of 2022, given the empirical uncertainty around the expected deaths forecasts.

2 Data and Methods

2.1 Data

We sourced data on weekly deaths from the Short-Term Mortality Fluctuations Database (STMF, Németh et al. 2021). The STMF is part of the Human Mortality Database (HMD, Max Planck Institute for Demographic Research, University of California, Berkeley, and French Institute for Demographic Studies 2023) and was established in response to the COVID-19 pandemic and the "increasing importance of short-term or seasonal mortality fluctuations that are driven by temporary hazards such as influenza epidemics, temperature extremes, as well as man-made or natural disasters" (Jdanov et al. 2021). The STMF provides open-access data on mortality by week, sex, and aggregated age group for 38 countries that follow the HMD's criteria of high data quality. Information on the data quality, sources, and completeness for each individual country can be found in the metadata file (Short-Term Mortality Fluctuations data series 2024). In general, STMF data for all countries is characterized by high quality, timely registration and completeness. The mortality counts are neither adjusted for e.g., under-reporting, nor are they smoothed (Jdanov et al. 2021). We use weekly death counts for 23 of the available countries. Because the data availability across time varies by country, we selected countries whose first observed data entry is for the year 2000 or before. These long time series are needed to robustly estimate the empirical error distribution, because the data is cut into several cross-validation series, which will be described in detail in this section.

We split the data for each country into three different types of data series (see Figure 1). We use these data series for different parts of our analysis. The first five data series are the calibration series. Using these, we calibrate the empirical prediction intervals. The following two series are used for the validation of the derived empirical prediction intervals. The forecast period of the last data series spans over the time of the COVID-19 pandemic and is the application series. Using the application series, we answer our research question on detecting an increase in excess deaths given the uncertainty of the forecasts of expected deaths.

Starting from the last observed week, we split the data into eight partially overlapping series of 365 weeks (7 years). Each series is divided into a training period (blue) of 261 weeks (5 years), and a forecast period (pink) of 104 weeks (2 years) (see Figure 1 for Germany and Table 2 for the starting and end dates of the data series).

2.2 Steps of Analysis

The methodology of our analysis is structured into several steps. With the first four steps, we obtain and calibrate the empirical prediction intervals, and parametric prediction intervals for comparison. For this, we use the calibration data series. During step five, we assess the calibration of the prediction intervals using the validation data series. Following these steps, we are able to answer our research question regarding the application of empirical prediction intervals to the COVID-19 pandemic.

I Modeling and Forecasting Expected Deaths First, we model and forecast the expected deaths. These are the deaths that would have occurred if a specific event, for example, the COVID-19 pandemic, did not happen. In line with the WHO approach (Weinberger et al. 2020), we forecast expected deaths $\hat{y}^E$

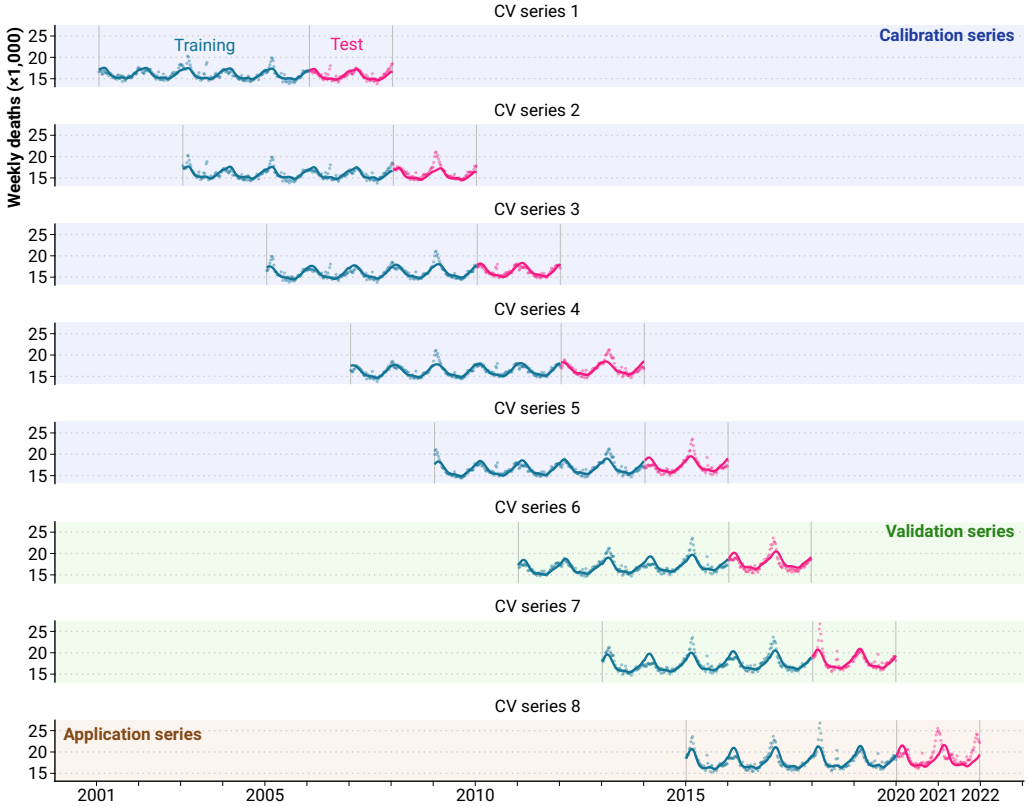


Figure 1: Data-splitting setup. 22 years of weekly death counts have been split into 8 overlapping data series. Each series features 5 years of training data and 2 years of test data. The empirical forecasting error is estimated from the test set of the calibration series, validated on the test set of the validation series and applied to the test set of the application series.

at time $T + h$ using an overdispersed count regression,

$$\hat{y}^E_{T+h} = \exp\left(\alpha + \beta_h h + s_w\left(w\left[h\right]\right)\right), \tag{1}$$

where T is the last time-point in the training data and h is the number of weeks into the forecasting horizon, $\beta_h h$ captures the long term trend of deaths, and $s_w\left(w\left[h\right]\right)$ is a cyclical penalized spline over the week of the year to reflect the seasonality of death counts. The model is fitted to the training periods of the data series and predictions are made over the testing periods. Note that empirical prediction intervals can be derived for any other model for excess deaths. For simplicity of exposition, we did not include further refinements like population offsets, temperature effects, or auto-correlated errors into the model. We assume the weekly deaths under the expected scenario to be distributed as

$$Y^E_{T+h} \sim \text{Neg.Binomial}(\hat{y}^E_{T+h}, \phi). \tag{2}$$

II Deriving the Forecast Error In a second step, we quantify the forecasting error at time $T + h$ as the logged ratio between observed and expected deaths over the test period:

$$u_{T+h} = \log \frac{y^O_{T+h}}{\hat{y}^E_{T+h}}. \tag{3}$$

We choose the log-ratio as the error scoring function because it reduces the scale dependence and asymmetry of the distribution of errors, in turn making it easier to model the observed error distribution.

III Modeling the Forecast Error Third, we model the time-varying distribution of the observed forecasting error over the forecasting horizon, U_{T+h}, via a skew-normal distribution with a location parameter μ, and time-varying scaling and skewness parameters σ_h and υ_h:

$$U_{T+h} \sim \text{SkewNormal}(\mu, \sigma_h, \upsilon_h). \tag{4}$$

As the observed log-errors are positively skewed, the skew-normal distribution allows for more accurate modeling of the error distribution over time, compared to a normal distribution.

We model the scaling parameter as

$$\sigma_h = \exp\left(\alpha_\sigma + s_\sigma(w[h])\right) \tag{5}$$

where $s_\sigma(w[h])$ is a smooth function of calendar week, capturing the annual seasonality of the error dispersion. In the context of expected deaths modeling, this seasonality is pronounced due to the challenges in predicting the severity of flu-waves during winter. While we would expect the error variance to increase with increasing forecasting horizon as well, we found that effect to be negligible over the duration of 100 weeks and did not include it here. The skewness υ_h is modeled equivalently via,

$$\upsilon_h = \alpha_\upsilon + s_\upsilon(w[h]), \tag{6}$$

and the mean error is captured in the constant

$$\mu = \beta_{\mu_0}. \tag{7}$$

IV Deriving Prediction Intervals Once estimates for σ_h, μ_h and υ_h are found, in the fourth step, we derive the distribution of expected deaths from the quantiles of the error distribution. The p quantile of the distribution of expected/forecasted deaths Y^E_{T+h} can be derived from the corresponding quantile of the error distribution via

$$Q_{Y^E_{T+h}}(p) = \exp(\hat{F}^{-1}_{U_{T+h}}(p)) \cdot \hat{y}^E_{T+h}, \tag{8}$$

where $\hat{F}^{-1}_{U_{T+h}}$ is the estimated distribution function of the forecasting error. Intuitively, if we estimat that 95% of the errors in our forecast for a given week do not exceed 1.5 times the central forecast, then the corresponding 95% prediction interval around the expected deaths is 1.5 times the central forecast for that week, likewise for other quantiles.

For comparative validation, we also derive prediction intervals from a negative-binomial generalized additive model (GAM). The negative-binomial GAM prediction intervals are commonly used for count data (see Olive et al. 2021, for examples). Additionally, we calculate the raw quantiles of the forecast error distribution as a second empirical model of the forecast error.

V Assessing Calibration of Empirical Prediction Intervals In the final step, we assess the calibration of the empirical and parametric prediction intervals using two different calibration metrics: the coverage, and the mean interval score.

The coverage measures the fraction of observations that are within the bounds of the prediction interval. Ideally, nominal coverage and observed coverage are the same, for example, a 95% prediction interval should, over the course of many forecasts, cover 95% of realized values. For a collection of prediction intervals (l_i, u_i) and associated observations y_i the coverage can be calculated as

$$\text{Coverage} = \frac{\sum_i^N \mathbf{1}(l_i < y_i < u_i)}{N} \tag{9}$$

where l_i and u_i are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles, $i = 1, \ldots, N$ is an index over all forecast data points, and $\mathbf{1}(\cdot)$ is the indicator function.

The mean interval score as proposed by Gneiting and Raftery (2007) is an example of a *proper scoring rule* and as such has optimal properties for assessing the quality of distributional and, specifically, interval forecasts. In the literature on COVID related prediction models the interval score has seen intensive use (Brooks et al. 2020; Bracher et al. 2021a,b; Sherratt et al. 2023), and is one of the evaluation measures of the COVID-19 Forecasting Hub Project (Cramer et al. 2021).

The mean interval score balances the two requirements of coverage and sharpness, rewarding narrower prediction intervals but penalizing deviations from the nominal coverage. Among two alternative prediction intervals with equal coverage, the interval which on average is narrower will yield the lower, that is, better, mean interval score. This balancing property is very valuable in the evaluation of prediction interval around forecasts of weekly deaths, because the variance of the error varies with the season. Thus, it is possible to have prediction intervals which have perfect coverage over the whole year but are too narrow in winter and too wide in spring. The mean interval score instead rewards prediction intervals that are narrow where they can be, and wide where they must be. The score is defined as follows:

$$S_{\alpha,i}(l_i, u_i, y_i) = \begin{cases} (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i) & \text{for} \quad y_i < l_i \\ (u_i - l_i) + \frac{2}{\alpha}(y_i - u_i) & \text{for} \quad y_i > u_i \end{cases} . \tag{10}$$

We then average the interval scores over all forecasts in the validation series across all countries. Following a suggestion by Bosse et al. (2023), we calculate the interval scores over log-transformed observations and prediction intervals as otherwise the mean interval score would be dominated by countries with large population numbers.

3 Results

3.1 Method demonstration for Germany

In the following, we will showcase the application of the empirical prediction interval methodology for Germany. The results of the cross-country analyses will be presented in the following sections.

Following steps I to III of our methodology, Figure 2 shows the distribution of the weekly observed forecast errors in the data series used for calibration, along with modelled forecast error distributions resulting from three different models. Panel (a) shows the observed forecast errors contrasted against the 95% error interval derived from the negative binomial variation, a common parametric specification in excess death modelling (Aburto et al. 2021; Msemburi et al. 2023). Panels (b) and (c) show two types of post-hoc (empirical) estimates of the forecast error distribution: the skew-normal error model (b) and the raw quantiles of the empirical error distribution (c).

As expected, the observed errors (pink) are positively skewed with a strong seasonal pattern. Both the skewness and the seasonality in the distribution of forecasting errors, with elevated errors in winter (weeks 0-12, 49-66, and 95-104) and, to a lesser extend, in summer (weeks 20-30 and 75-85), are due to the irregularities in the annual occurrence of flu-epidemics and heat-waves. This seasonal shift in the variance of forecasting error is well reflected by the skew-normal model (b) and the empirical quantiles (c). In contrast, the often used negative binomial specification with fixed overdispersion implies a constant forecasting error (a). Figure 5 in the Appendix shows the forecast error distribution modelled with the skew-normal distribution for all 23 countries.

As specified in step IV, we derive empirical prediction intervals around the central forecasts of expected deaths from the estimated distribution of forecasting error. For the negative binomial model, we simply report the parametric prediction intervals. Figure 3 contrasts the three approaches in the context of the German COVID-19 pandemic, showing observed and expected weekly deaths since January 2020 along with 95% prediction intervals. The difference between the empirical and parametric prediction intervals is especially notable during winter, when excess deaths seem far more unusual under negative binominal intervals than

under the seasonally widened empirical intervals. The prediction intervals derived from the raw quantiles of the error distribution are not as smooth as the skew-normal modeled errors. Figure 6 shows the empirical skew-normal prediction intervals along with observed and forecast weekly deaths over the application data series for all 23 countries.

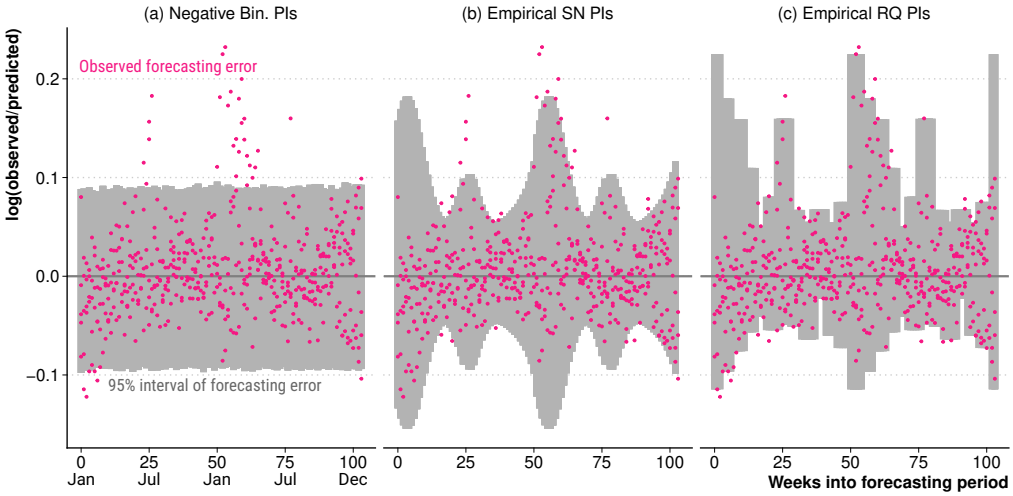


Figure 2: Distribution of the observed forecast errors in the calibration data series (pink, test periods from 2006 to 2016), with modelled forecast error distributions, on the example of Germany.

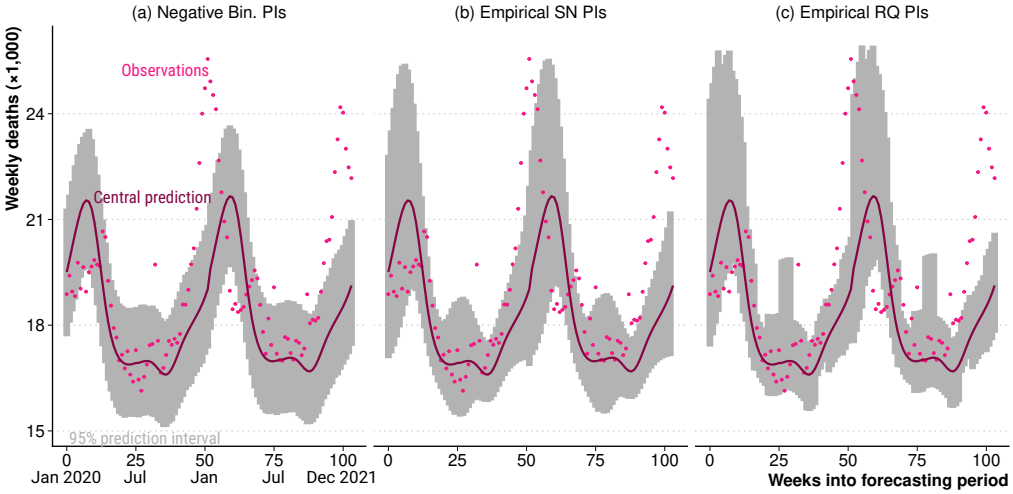


Figure 3: Negative binomial prediction intervals (a) and empirical prediction intervals (b) applied to the application (COVID-19) data series (test periods from 2020 to 2021), on the example of Germany.

3.2 Cross-country evaluation of forecast calibration

Table 1 shows calibration metrics for the three different prediction interval types as calculated from the validation series. We report the coverage and the mean interval score aggregated across all 23 countries, annually and by season, for a nominal coverage of 95%.

All three types of prediction intervals tend to be too narrow, although close to the nominal coverage. On an annual basis, the negative binomial 95% prediction intervals (PIs) perform best with an actual coverage of 93%, closely followed by the empirical skew-normal PIs with 91%. However, looking at the performance over different seasons, the empirical skew-normal PIs achieve better actual coverage in winter and autumn than the parametric PIs, as the latter are too wide in autumn (99% actual coverage) and too narrow in winter (85% actual coverage). These peculiarities are hidden when only the calibration for the whole year is considered.

These findings are also supported by the mean interval scores. The empirical skew-normal PIs are better calibrated than the negative binomial PIs overall, and in all seasons except summer. This indicates an overall better probabilistic calibration of the empirical skew-normal PIs compared to the conventional negative binomial PIs.

The prediction interval derived from the raw quantiles of the forecast error distribution perform comparatively poorly which may be due to over-fitting. Thus, further results in the paper will be shown without the inclusion of the raw-quantile method.

Table 1: Calibration metrics for nominal 95% prediction intervals by season and type of interval. The metrics have been calculated on the test periods of the validation data (2016-2018), i.e. on data not seen either during training or calibration, and aggregated over 23 countries.

	Coverage				
	Annual	Dec–Feb	Mar-May	Jun–Aug	Sep–Nov
Negative Bin. PIs	**0.93 (1)**	0.85 (3)	**0.91 (1)**	**0.95 (1)**	0.99 (3)
Empirical SN PIs	0.91 (2)	**0.89 (1)**	0.89 (2)	0.91 (2)	**0.94 (1)**
Empirical RQ PIs	0.86 (3)	0.86 (2)	0.82 (3)	0.86 (3)	0.91 (2)
	Mean Interval Score				
Negative Bin. PIs	0.365 (2)	0.553 (3)	0.376 (2)	**0.287 (1)**	0.248 (2)
Empirical SN PIs	**0.346 (1)**	**0.467 (1)**	**0.369 (1)**	0.310 (2)	**0.241 (1)**
Empirical RQ PIs	0.398 (3)	0.534 (2)	0.437 (3)	0.363 (3)	0.260 (3)

Note: bold font highlights best/same performance among types of prediction intervals
Data Source: Short-term Mortality Fluctuations (STMF) data series, own calculations

3.3 Probability of 10% excess death under pre-pandemic mortality

The choice of prediction interval is crucial for evaluating the significance of a given excess death estimate. How unusual is a finding of 10% weekly excess deaths in a given country and season, given non-pandemic mortality conditions? Figure 4 answers this question for each of the 23 chosen countries under two different prediction intervals: the seasonally adjusted empirical intervals based on past forecast errors and the commonly used negative binomial intervals with time-constant overdispersion. Under the hypothesis that past, non-pandemic mortality trends continue, the graph shows the probability (p-value) for excess deaths in a given week to exceed 10%.

If a standard negative binomial GAM was chosen to derive the prediction intervals, a researcher would hold the belief that exceeding 10% excess deaths would be just as likely in spring as in winter (Figure 4, blue line). The empirical prediction intervals (pink line) shows that this is not the case and that the likelihood of exceeding the 10% threshold varies by season. A case in point is France, with a peak 20% probability of seeing at least 10% excess deaths for a winter week and a p-value approaching zero during spring. In all countries, there are periods when the p-value exceeds the level of 0.05, yet, there are differences in the

seasonal patterns. For most of the countries, the probability of observing 10% excess or more is elevated during the winter and also for the summer weeks (e.g. Austria, Hungary, Portugal, Belgium, Poland, Spain). In contrast, in Norway, Sweden, and Finland, we only see a winter and no summer elevation of the p-value. Further, for Croatia and Bulgaria, the p-value is higher in summer than in winter. Therefore, across all of the 23 countries, we find periods when 10% weekly excess deaths are not unusual under normal mortality conditions. This is especially true for smaller populations (e.g., Scotland, Latvia, Luxembourg) where sampling fluctuations in death counts regularly produce excess death estimates above 10%, even under normal mortality conditions. In a complementary finding, for Austria, France, Germany, Poland, Spain, and Sweden, the empirical prediction intervals predict a 10% excess p-value close to zero for parts of spring and summer, indicating that 10% excess deaths are not unusual in winter, but unusual in spring or summer.

The question about how unusual a given level of weekly excess deaths is, can be inverted: Under pre-pandemic mortality, what level of weekly excess deaths is needed so that they fall outside, for example, a 95% prediction interval and thus, be significant? To answer this question, we evaluated the expected deaths distribution at the 95% prediction interval and derived the implied excess deaths by country (Figure 7 in the Appendix). If negative binomial GAM prediction intervals are applied, for most countries, a level of 8-12% of excess deaths would be needed so that they fall outside of the 95% prediction interval. This level of excess deaths remains constant throughout the year. However, the use of the empirical prediction intervals reveals the seasonality of the uncertainty and that a level of excess deaths of up to 16% (and even more for smaller regions such as Scotland and Luxembourg) is needed to keep the excess deaths from being hidden by the uncertainty in the expected deaths forecasts.

4 Discussion

Forecasts of weekly deaths have grown in importance since the onset of the COVID-19 pandemic and the need for estimating excess deaths. These forecasts of expected deaths have an error that is probabilistically expressed as a prediction interval. Usually, these prediction intervals are model based parametric derivations. However, they tend to be too narrow, giving a too optimistic outlook on the precision with which one can estimate excess deaths during an ongoing pandemic. In this paper, we propose the use of empirical prediction intervals for the study of weekly expected and excess deaths and show the superior calibration of these intervals compared to parametric intervals. Further, we demonstrate the application of empirical prediction intervals to the COVID-19 pandemic, answering the question whether a 10% increase in excess deaths can be detected or whether it disappears into the uncertainty of the expected deaths forecasts.

Empirical prediction intervals are based on modeling the distribution of the forecast errors from past out-of-sample forecast errors. We use a skew-normal distribution with a time varying seasonality parameter to model the forecast errors for weekly deaths, accounting for the positive skewness and seasonality of errors. For the validation analysis, we compare the empirical prediction intervals with parametric prediction intervals obtained from a negative binomial GAM, using two measures of calibration: the coverage and the mean interval score. We obtain weekly mortality data from the Short-Term Mortality Database (STMF, Jdanov et al. 2021) for 23 countries. We showcase the calculation of empirical prediction intervals for these countries, perform the validation analyses, and apply the different intervals to excess deaths in the COVID-19 pandemic.

We showed that our model for the empirical prediction intervals is able to capture the seasonality in weekly deaths and the positive skewness due to the varying interval width over the year. In contrast, the negative binomial prediction intervals have a constant width throughout the year, resulting in excessively wide intervals in fall and excessively narrow intervals in winter. Further, our validation analysis demonstrated that the empirical prediction intervals are better calibrated than the negative binomial prediction intervals. This is reflected in the mean interval score that is, on average, better for the empirical prediction intervals. Both types of intervals exhibit good coverage that is, on average, slightly short of the nominal 95%.

Regarding the application question, using negative binomial prediction intervals would give the result that a 10% increase in excess deaths would be detectable in fall of 2022. However, the empirical prediction intervals show that this is not the case during the whole year. Regarding the example of Germany, in winter a single week with 10% increase in excess deaths would disappear into the uncertainty of the forecasts of expected deaths. We found similar results for the other countries in our sample: there are periods throughout

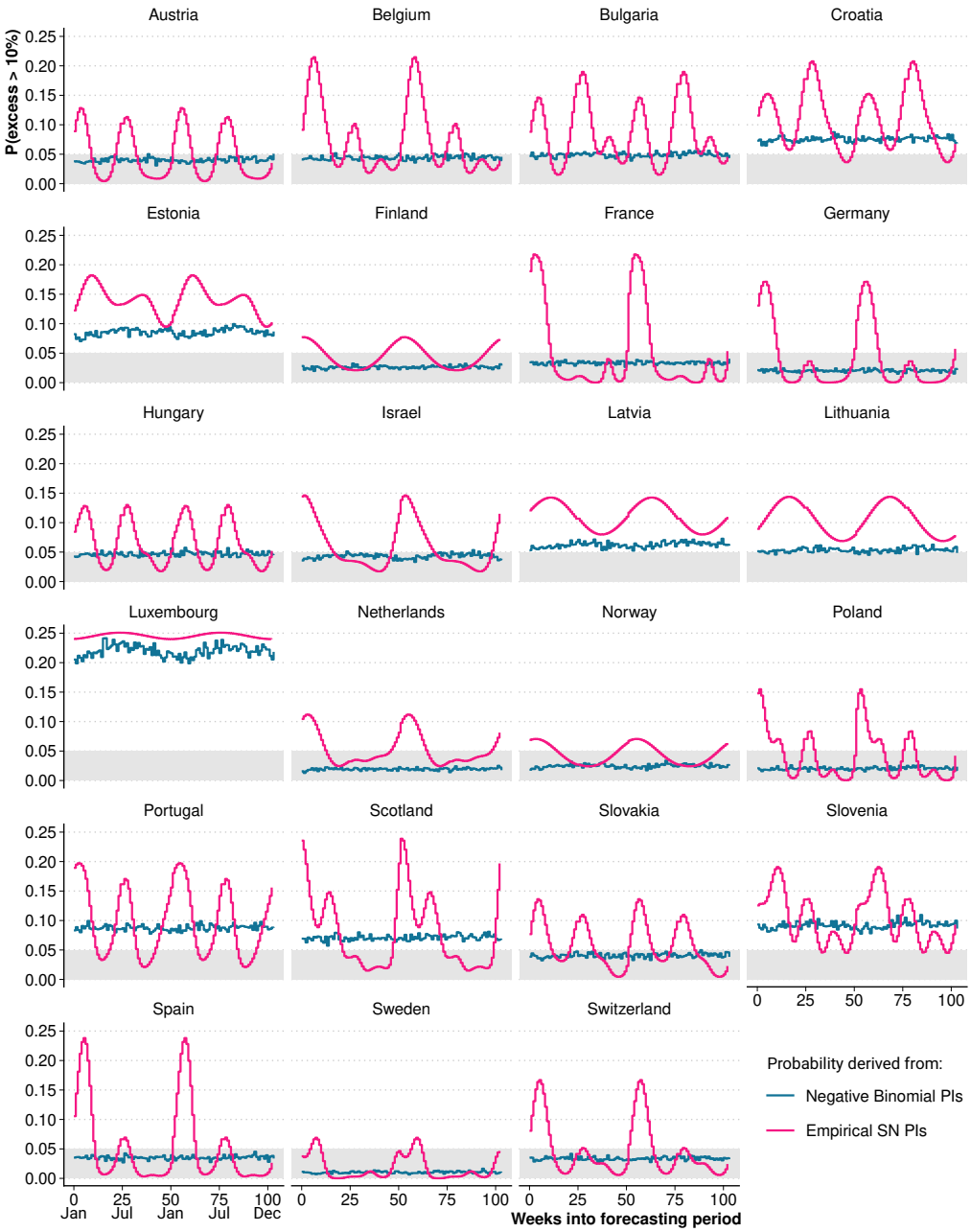


Figure 4: P-value for 10% excess deaths derived from empirical prediction intervals and negative binomial prediction intervals for 23 European countries.

the year in which a 10% increase in weekly excess deaths can not be reliably predicted for every country. These periods differ between the countries.

The difference in forecast error seasonality between countries may be connected to differences in the prevalent climate. Future research could further investigate this relationship between the climate of a country of interest and the seasonal patterns of uncertainty in mortality forecasting. Both the general effects of the climate and its variability from year to year are likely to shape the uncertainty in the expected deaths forecasts. A region's climate influences population health outcomes, for example, via heat-related stress. However, heatwaves are difficult to predict and usually not included in mortality forecast models, contributing to increased uncertainty in mortality forecasts for the summer months. In colder or more moderate climates, the uncertainty stems from the unpredictability of the flu season severity, which can vary significantly from year to year, increasing uncertainty during the winter months. Regarding the study of excess deaths due to COVID-19, our results highlight the importance of well calibrated prediction intervals that also address the naturally occurring seasonal uncertainty that comes with mortality forecasting.

Weeks are the common time frame for tracking the mortality during the COVID-19 pandemic, because, unlike days, they are less sensitive to reporting delays that occur on weekends and holidays, while maintaining the timeliness of the mortality monitoring. If the time frame of the analyzed death counts was changed from one week to one month or even one year, the error distribution around the expected deaths would change as well. Determining the exact level of excess deaths needed in order to be detectable for longer time periods, is a topic for future research.

The application of empirical prediction intervals to any measure, in our case excess deaths, is based on the underlying assumption that the forecasting errors observed in the past resemble the forecasting errors in the future. One might ask whether this assumption is valid. However, as Alho et al. argue, "[...] if one does not believe that they will be, it is necessary to provide arguments as to why the future is expected to be different from the past" (Alho et al. 2008, p. 42). The counterfactual assumption that future forecasting errors will be fundamentally different from past errors is the more complicated one. Our assumption is the more conservative option, which is necessary to allow us to use empirical prediction intervals as a measure of forecast uncertainty.

Our findings are broadly generalizable in the sense that the proposed methodology is easily implemented, as long as there are sufficiently long data series to perform the data splitting into calibration, validation, and application data sets. Further, we are confident that the empirical prediction intervals will outperform other types of prediction intervals for the analysis of excess deaths due to their ability to capture the seasonality in the uncertainty of the expected deaths forecasts, regardless of the country under study. However, due to the different seasonal patterns found for different countries, our findings regarding the likelihood of detecting excess deaths in a given week of the year are specific to the countries of our analysis.

Regarding applications of the proposed methodology, a real-time implementation of empirical prediction intervals for monitoring excess deaths would offer a chance to judge the "unusualness" of registered death counts. On the one hand, this would help to monitor excess deaths during a pandemic, compared to pre-pandemic mortality, and would help researchers, decision makers, and the public to assess the severity of the pandemic and to act accordingly. Empirical prediction intervals can be used, especially later in a pandemic, when counterfactual analyses become less sensible because too much time has passed. On the other hand, a real-time implementation could serve as an early indicator of public health crises due to pandemics or other environmental causes. Further, the methodology of empirical prediction intervals can increase understanding when communicating research results on excess deaths to the public. This is due to the intuitive notion that future forecast errors reflect past errors. Obviously, data on an ongoing pandemic is of great public interest. Therefore, it is important that researchers communicate their findings in a way that is easily understood not only by other researchers, but also by the public. In the case of excess deaths and the uncertainty surrounding them, we would recommend the use of confidence bands to visually convey uncertainty, rather than relying on p-values, for example. Further, the comparison with past observed fluctuations in death counts could help to contextualize and assess the severeness of current excess deaths.

Empirical prediction intervals are generally applicable. Not only can they capture different patterns of seasonality in deaths (e.g., high winter uncertainty and low summer uncertainty and vice versa). They can be applied to all models of excess deaths and other methods of forecasting, for example, forecasts using expert opinions, as they only derive from observable data and not from unobservable model parameters. As long as the forecasts can be validated over a long time, empirical prediction intervals can be used. Therefore, they

are not limited to mortality forecasting, either. Using adapted specifications to model the forecast error, they can be applied to, for example, fertility forecasts. Thus, research concerned with demographic forecasting in general can profit from the use of empirical prediction intervals. We invite researchers to use empirical prediction intervals for their forecasting problems and showcase the possibilities of their application.

Availability of data and materials

The research presented in this article is fully reproducible. We have made our codebase available to the scientific community. You can access the complete codebase, including scripts and documentation, at the following link: `https://github.com/jschoeley/epunc`. In addition, the data used for our analyses is sourced from openly accessible and publicly available datasets. Researchers interested in reproducing our work or conducting further investigations can freely download the data from the following link: `https://www.mortality.org/Data/STMF`. The availability of open-access data ensures transparency and promotes collaboration within the scientific community.

Competing interests

JS is one of the editors of the Collection "COVID-19 and Impact on Mortality and Population Health".

Author contributions

RD and JS performed the statistical analyses, prepared figures and tables, and wrote the main manuscript text. JS designed and supervised the project, and edited the final manuscript. All authors reviewed the manuscript.

Acknowledgements

RD gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS). The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions, which have greatly improved the quality of this paper.

Appendix

Table 2: Structure of data used for analyses.

Data Series	Category	Start Date of Training Period	Start Date of Testing Period	End Date of Testing Period
1	Calibration	22.01.2001	23.01.2006	14.01.2008
2	Calibration	20.01.2003	21.01.2008	11.01.2010
3	Calibration	17.01.2005	18.01.2010	09.01.2012
4	Calibration	15.01.2007	16.01.2012	06.01.2014
5	Calibration	12.01.2009	13.01.2014	04.01.2016
6	Validation	10.01.2011	11.01.2016	01.01.2018
7	Validation	07.01.2013	08.01.2018	30.12.2019
8	Application	05.01.2015	06.01.2020	17.12.2021

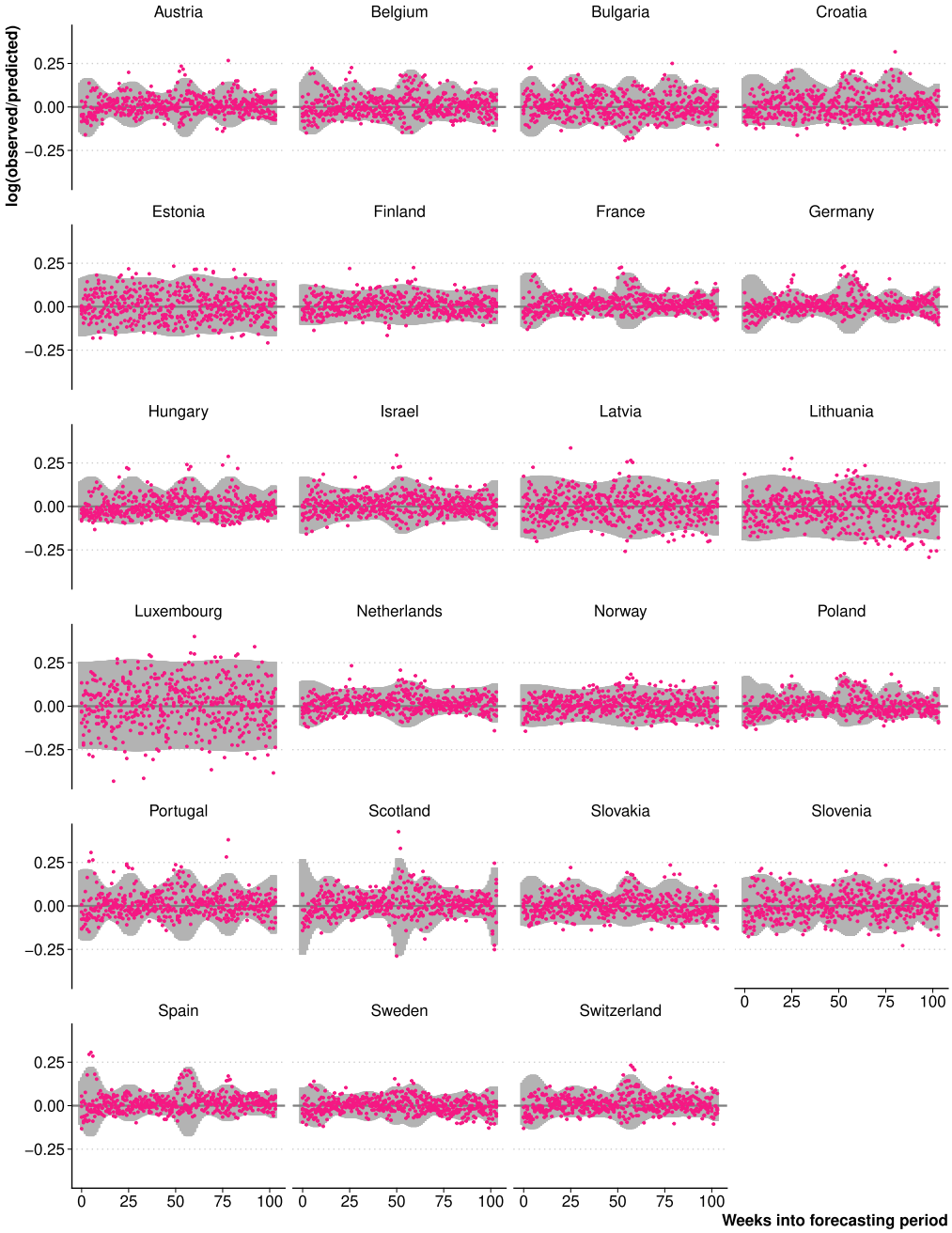


Figure 5: Distribution of the forecast error for weekly death counts as estimated from the calibration data series (test periods from 2006 to 2016) across 23 European countries.

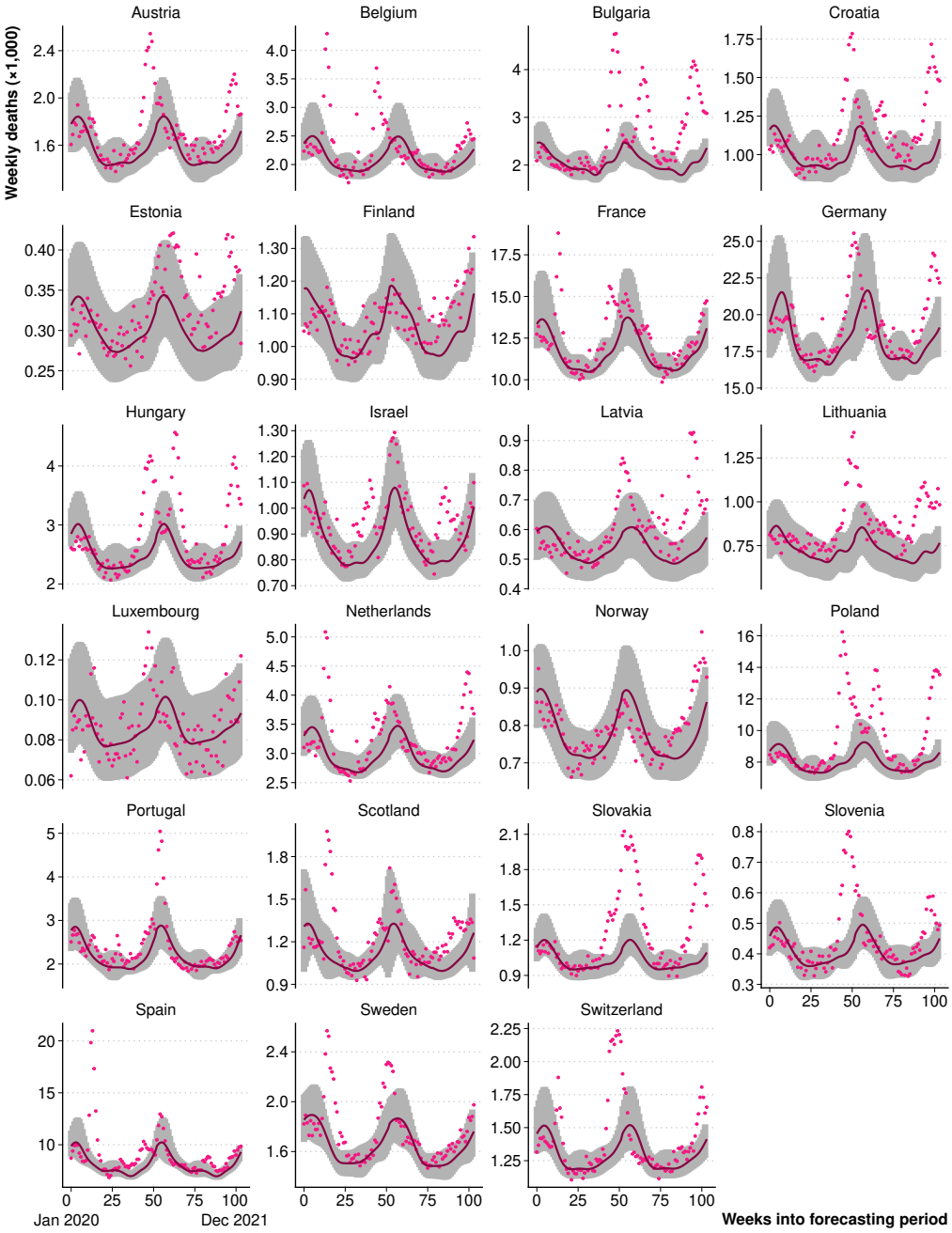


Figure 6: Empirical prediction intervals (95%) applied to the COVID-19 application data series (test periods from 2020 to 2021) across 23 European countries.

Figure 7: Percentage of excess deaths at the 95% prediction interval of expected deaths derived from the calibration data series (test periods from 2006 to 2016) for 23 European Countries.

References

Aburto, J. M., Kashyap, R., Schöley, J., Angus, C., Ermisch, J., Mills, M. C., and Dowd, J. B. (2021). Estimating the burden of the covid-19 pandemic on mortality, life expectancy and lifespan inequality in england and wales: a population-level analysis. *J Epidemiol Community Health*, 75(8):735–740.

Alho, J. M., Cruijsen, H., and Keilman, N. (2008). *Uncertain demographics and fiscal sustainability*, chapter Empirically based sprecification of forecast uncertainty, pages 34–54. Cambridge University Press, Cambridge.

Bosse, N. I., Abbott, S., Cori, A., van Leeuwen, E., Bracher, J., and Funk, S. (2023). Scoring epidemiological forecasts on transformed scales. *PLoS Computational Biology*, 19(8):e1011393.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021a). Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618.

Bracher, J., Wolffram, D., Deuschel, J., Görgen, K., Ketterer, J. L., Ullrich, A., Abbott, S., Barbarossa, M. V., Bertsimas, D., Bhatia, S., et al. (2021b). A pre-registered short-term forecasting study of covid-19 in germany and poland during the second wave. *Nature communications*, 12(1):5173.

Brooks, L. C., Ray, E. L., Bien, J., Bracher, J., Rumack, A., Tibshirani, R. J., and Reich, N. G. (2020). Comparing ensemble approaches for short-term probabilistic covid-19 forecasts in the us. *International Institute of Forecasters*.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2):121–135.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(3):419–444.

Cohen, J. E. (1986). Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. *Demography*, 23(1):105–126. Publisher: Duke University Press.

Cramer, E. Y., Huang, Y., Wang, Y., Ray, E. L., Cornell, M., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., House, K., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Niemi, J., Stark, A., Shah, A., Wattanachit, N., Zorn, M. W., Reich, N. G., and Consortium, U. C.-. F. H. (2021). The united states covid-19 forecast hub dataset. *medRxiv*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. Publisher: Informa UK Limited.

Jdanov, D. A., Galarza, A. A., Shkolnikov, V. M., Jasilionis, D., Németh, L., Leon, D. A., Boe, C., and Barbieri, M. (2021). The short-term mortality fluctuation data series, monitoring mortality shocks across time and space. *Scientific Data*, 8(1). Publisher: Springer Science and Business Media LLC.

Karlinsky, A. and Kobak, D. (2021). Tracking excess mortality across countries during the covid-19 pandemic with the world mortality dataset. *elife*, 10:e69336.

Keilman, N. (2001). Uncertain population forecasts. *Nature*, 412(6846):490–491.

Keilman, N. and Pham, D. Q. (2004a). Empirical errors and predicted errors in fertility, mortality and migration forecasts in the european economic area. *Dicussion Papers*, 386.

Keilman, N. and Pham, D. Q. (2004b). Time series based errors and empirical errors in fertility forecasts in the nordic countries. *International Statistical Review*, 72(1):5–18.

Keilman, N., Pham, D. Q., and Hetland, A. (2002). Why population forecasts should be probabilistic-illustrated by the case of norway. *Demographic research*, 6:409–454.

Keilman, N. W. (1990). *Uncertainty in national population forecasting: Issues, Backgrounds, Analyses, Recommendations*. Swets & Zeitlinger, Amsterdam.

Kontis, V., Bennett, J. E., Rashid, T., Parks, R. M., Pearson-Stuttard, J., Guillot, M., Asaria, P., Zhou, B., Battaglini, M., Corsetti, G., et al. (2020). Magnitude, demographics and dynamics of the effect of the first wave of the covid-19 pandemic on all-cause mortality in 21 industrialized countries. *Nature medicine*, 26(12):1919–1928.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Lee, Y. S. and Scholtes, S. (2014). Empirical prediction intervals revisited. *International Journal of Forecasting*, 30(2):217–234.

Max Planck Institute for Demographic Research, University of California, Berkeley, and French Institute for Demographic Studies (2023). Human mortality database.

Msemburi, W., Karlinsky, A., Knutson, V., Aleshin-Guendel, S., Chatterji, S., and Wakefield, J. (2023). The who estimates of excess mortality associated with the covid-19 pandemic. *Nature*, 613(7942):130–137.

Németh, L., Jdanov, D. A., and Shkolnikov, V. M. (2021). An open-sourced, web-based application to analyze weekly excess mortality based on the Short-term Mortality Fluctuations data series. *PLOS ONE*, 16(2):e0246663. Publisher: Public Library of Science (PLoS) tex.owner: jon.

Olive, D. J., Rathnayake, R. C., and Haile, M. G. (2021). Prediction intervals for glms, gams, and some survival regression models. *Communications in Statistics-Theory and Methods*, 51(22):8012–8026.

Rayer, S., Smith, S. K., and Tayman, J. (2009). Empirical prediction intervals for county population forecasts. *Population Research and Policy Review*, 28(6):773–793. Publisher: Springer Science and Business Media LLC.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421. Number of pages: 51 Publisher: JMLR.org tex.issue_date: 6/1/2008.

Sherratt, K., Gruson, H., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolffram, D., Abbott, S., Ullrich, A., et al. (2023). Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *Elife*, 12:e81916.

Short-Term Mortality Fluctuations data series (2024). Metadata. `https://www.mortality.org/File/GetDocument/Public/STMF/DOC/STMFmetadata.pdf`.

Smith, S. K. and Sincich, T. (1988). Stability over time in the distribution of population forecast errors. *Demography*, 25(3):461–474. Publisher: Duke University Press.

Stoto, M. A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78(381):13–20. Publisher: Informa UK Limited.

Weinberger, D. M., Chen, J., Cohen, T., Crawford, F. W., Mostashari, F., Olson, D., Pitzer, V. E., Reich, N. G., Russi, M., Simonsen, L., et al. (2020). Estimation of excess deaths associated with the covid-19 pandemic in the united states, march to may 2020. *JAMA internal medicine*, 180(10):1336–1344.

Williams, W. H. and Goodman, M. L. (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, 66(336):752–754. Publisher: Informa UK Limited.

Calibrating Probabilistic Forecast Paths on Past Forecast Errors: An Application to the Finnish Total Fertility Rate

Authors: Ricarda Duerst[1,2], Jonas Schöley[1], Julia Hellstrand[2,1], Mikko Myrskylä[1, 2, 3]

Affiliations:
[1] Max Planck Institute for Demographic Research; Rostock, Germany
[2] University of Helsinki; Finland
[3] Max Planck - University of Helsinki Center for Social Inequalities in Population Health; Rostock, Germany and Helsinki, Finland

Abstract

In probabilistic forecasting, the key to reasonable results is good calibration of the forecast uncertainty. Analysing forecast errors offers an empirical solution for the calibration of such forecasts. We propose a novel quantile-mapping approach whereby we map non-calibrated forecast paths to a calibrated distribution derived from historical out-of-sample forecast errors. We present probabilistic forecasts of the Finnish Total Fertility Rate (TFR) from 2024 to 2070 calibrated on the historically observed distribution of forecast errors. We source the data from the Human Fertility Database. The forecasts come from two scenario-based models. The postponement time series model assumes that fertility postponement will gradually decline and eventually stop. The second model is a naive freeze-rates approach to forecasting fertility. The validation shows that our TFR forecasts calibrated on historical data outperform the non-calibrated TFR forecasts in the Continuous Ranked Probability Score and interval score metrics. The results demonstrate the efficacy of empirical error quantification and quantile-mapping in calibrating probabilistic demographic forecasts.

Key words
Conformal prediction; empirical prediction intervals; fertility forecasting; Finland; quantile-mapping

1 Introduction

Demographic forecasting involves predicting future population trends based on factors such as birth rates, death rates, migration patterns, and aging. These forecasts are essential for governments, businesses, and policymakers to plan for future resource needs, including infrastructure, healthcare, and social services. However, demographic forecasting is subject to uncertainties due to unpredictable changes in behavior, policy, and environmental factors. Forecast uncertainty arises from variability in data inputs, model assumptions, and external events such as pandemics or economic crises. Managing this uncertainty and expressing it in the forecasts is essential to providing a comprehensive view of future demographic outcomes.

In the 1960s and early 1970s, Finland, and other Nordic countries, have experienced a sharp decline in fertility. After a period of recovery in the 1980s, followed by relatively stable period fertility in the 1990s and 2000s, Finland's Total Fertility Rate (TFR) fell again in the 2010s and reached a historical low of 1.26 in 2023 (Statistics Finland 2024a). Starting in the 1970s, fertility postponement, the delay of childbearing to older ages, has depressed fertility levels. However, the strong decrease in TFR in the recent decade is not only due to further acceleration of fertility postponement. Instead, the decrease in Finnish period fertility likely is a quantum effect (Hellstrand et al. 2020). This was not foreseen by demographic theories, making the future of Finnish fertility particularly interesting yet highly challenging to predict.

In light of yet another fluctuation in the trend of Finnish fertility, it is important to capture and convey the uncertainty that comes with forecasting Finland's fertility. Probabilistic forecasts, in contrast to deterministic point-forecasts, offer a solution to this problem. This type of forecast assigns a probability to each possible outcome and therefore allows one to distinguish between likely and extreme scenarios and to plan accordingly (see e.g. Lee 1998; Keilman 2018). However, probabilistic forecasts are only of use if they are well calibrated. Good calibration means that the forecasted probabilities of an outcome are consistent with the observed relative frequencies of that outcome, or put differently, forecasted rare events occur rarely and forecasted typical events occur regularly. Further, because of the highly fluctuating nature of the Finnish period fertility in the past, forecast models that extrapolate observed trends into the future are unlikely to provide reasonable results, as they are unable to predict trend changes. Therefore, we propose the use of scenario-based approaches that can take into account the changing nature of the fertility development.

We present probabilistic forecasts of Finland's TFR from 2024 to 2070 from two scenario-based forecast models. The first model, based on Nisén et al. (2020), is the Postponement Time Series Model (PPS). The PPS model operates under the assumption that the trend of delaying childbirth to later ages, known as fertility postponement, will continue but at a decreasing pace, until it eventually stops. Hence, the increase in the average age of childbirth slows down and stabilizes. As childbirth is postponed, the TFR temporarily decreases because children are born later during the life course. Therefore, the fertility is lower, compared to a scenario in which childbirth is not delayed. When fertility postponement slows down, fertility increases. To mitigate the delay's impact on period fertility measures, we use the tempo-adjusted TFR. This adjusted rate is an estimate of what the TFR would be if the timing of childbearing did not change (Bongaarts and Feeney 1998). The PPS model assumes that the TFR and the tempo-adjusted TFR will converge due to the assumed slowing and stopping of fertility postponement by 2050.

We use a second model as a naïve baseline to compare the results from the PPS model to. The freeze-rates model operates under the assumption that the recent decrease in TFR is not a result of delayed childbirth, meaning that fertility has declined without postponement of births by the population. While this hypothesis is theoretically feasible, considering demographic data, it seems improbable. The second assumption is that this fertility decline will come to a halt, and age-specific fertility rates will stay at the levels observed in the last recorded year, 2023. This freeze-rates approach is commonly used as a fertility scenario in population projections published by statistical offices and in other research (e.g., Institut national de la statistique et des études économiques 2008; Statistics Sweden 2024; United Nations, Department of Economic and Social Affairs, Population Division 2024; Statistics Finland 2024b; Zeman 2024; Nisén et al. 2020; Wilke 2018).

To ensure a good calibration of the probabilistic forecasts, we use an empirical approach which originated from the analysis of errors of demographic forecasts (Williams and Goodman 1971; Stoto 1983; Cohen 1986; Smith and Sincich 1988; Alho and Spencer 2005; Alho et al. 2008). Notably, Keilman and Pham (2004) constructed empirical prediction intervals around forecasts of Nordic fertility, deriving the intervals from the standard deviation of errors of historical forecasts published by statistical agencies. After comparing the width of the prediction intervals with those from time series models, the authors concluded that empirical

forecast errors provide useful information when constructing prediction intervals for TFR forecasts. However, they noted that their empirical errors might not be normally distributed and "one has to be cautious" when using them.

In the field of machine learning, the methodology is known under the term conformal prediction (CP, Shafer and Vovk 2008; Fontana et al. 2023) and is used to construct probabilistic forecasts calibrated on out-of-sample errors. Since its introduction in Gammerman et al. (1998) until today, there have been strong methodological advances in this field, resulting in a variety of CP methods for different applications including, in more recent publications, time series forecasting (Fontana et al. 2023; Angelopoulos et al. 2024). From this variety of methods, we focus on the split-conformal prediction or inductive conformal prediction, in which prediction intervals are estimated from a hold-out data set (Papadopoulos et al. 2007), and demonstrate it in the context of scenario-based forecasts parameterized by expert opinion. In climate modeling, calibration of predictions using historical data is widely done under the term quantile-mapping as a form of bias-correction of forecast distributions (Cannon 2018; Qian and Chang 2021).

The existing methodology provides probabilistic forecasts calibrated on historical data in the form of prediction intervals, meaning lower and upper bounds in which the forecast outcome will lie with a given probability, e.g. 95%. However, so far, research has not provided solutions for forecast outcomes in the form of calibrated time series trajectories that correspond to these bounds. Simulated trajectories of demographic outcomes are needed as input for down-stream modeling, e.g. for the modeling of determinants of social security systems. Therefore, we propose a flexible methodology to obtain forecasts of demographic measures in the form of simulated trajectories that have been calibrated on historical data. We provide forecasts of the Finnish TFR from 2024 to 2070 from the PPS model and the naïve baseline model that are designed to be transparent, probabilistic, well calibrated, and easy to integrate into further probabilistic modeling downstream. We build on the prevailing research, by proposing to combine existing approaches of empirical forecast calibration. First, we use the idea of learning a smoothed, time-varying distribution of historical forecast errors, as described in the "scorecaster" approach by Angelopoulos et al. (2024). Second, we combine this approach with the technique of quantile-mapping (Cannon 2018) in order to calibrate stochastic forecasting paths to a distribution of empirical forecast errors, rather than just providing calibrated upper and lower bounds of prediction intervals around a point-forecast.

2 Methods and data

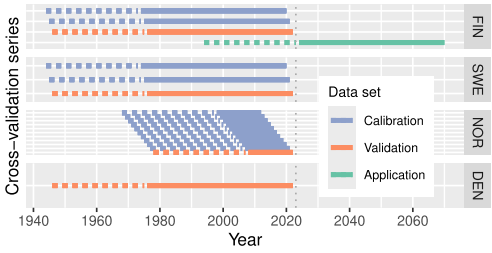
Figure 1 outlines our methodological approach. Figure 1.1: Although our country of interest is Finland and the final forecast output is the Finnish TFR from 2024 through 2070, we use data from Finland, Sweden, Norway and Denmark to calibrate this forecast. Therefore, we need to split the available data from these four Nordic countries into cross-validation series (cv-series) that slide along the observation period in one-year steps. Each of these cv-series is cut into a training period (dotted line) and a testing period (solid line). There are three different types of cv-series: The application data set (green) holds a single series for the Finnish forecast into the future, the calibration data set holds cv-series from Finland, Sweden and Norway that are used to learn the distribution of forecast errors of the PPS model (blue), and the validation data set holds cv-series used to validate the forecast approach on (orange). We have included data series from the other Nordic countries besides Finland in the calibration and validation data sets to increase the amount of data used for the calibration and avoid overfitting during the validation analysis. The similarity of the fertility regimes of the four Nordic countries justifies the use of these additional data series.

Figure 1.2: The next step after data splitting is the estimation of the time-varying distribution of forecast errors. We apply the PPS model to the training periods of the calibration data series and "forecast" the TFR for the duration of the testing period. By calculating the forecast errors we are able to derive a forecast error distribution over the forecast years that include the errors of all cv-series of the calibration data set. We then estimate this empirical error distribution using a skew-normal model.

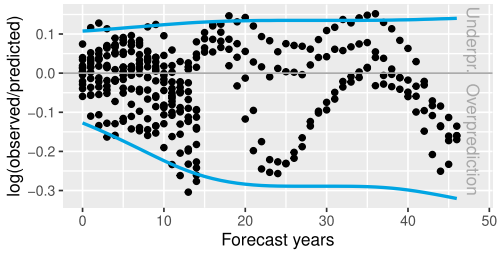
Figure 1.3: After applying the PPS model to the training periods of the validation data, we calibrate the resulting forecasts using the empirical forecast error distribution estimated in step 2. This is done using quantile-mapping of the forecast paths.

Figure 1.4: We compare the forecast performance of the PPS model before and after this calibration using several calibration scores. In addition, we compare the forecast performance to the naïve baseline

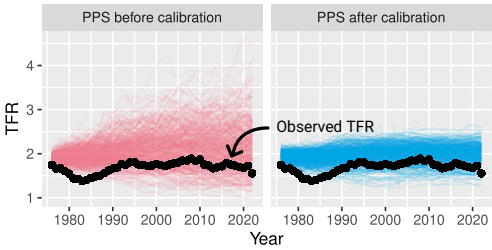
1 Split data into sliding window cross-validation sets
Split data into calibration, validation and application data sets. Divide each cross-validation series (cv-series) into a training period (dotted) and a testing period (solid).



2 Estimate time-varying distribution of forecast errors
Apply the postponenment scenario model (PPS) to the training period of all calibration data series to calculate the **forecast errors** for the testing period. Pool the forecast errors across all cv-series and model the empirical forecast error distribution.



3 Make calibrated forecasts on validation series
Apply the PPS model to the training period of all validation data series. Then calibrate the resulting forecast paths using the model of the empirical forecast error distribution. The graphs show the forecasts for the Danish validation series.

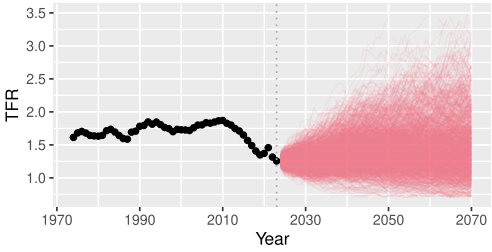


4 Calculate calibration scores
Calculate calibration scores for the testing period of all validation data series and compare the forecast performance of the PPS model before and after the calibration with the empirical forecast error distribution.

Model	Coverage	Mean Interval Score	Average width
PPS before calibration	**0.915**	1.728	1.340
PPS after calibration	0.901	**1.087**	**0.710**
naive freeze-rates model	**0.915**	1.590	1.240

The table shows error scores for the 95% quantile, pooled across all validation series. Best scores highlighted in **bold**.

5 Generate forecasts for Finland up to 2070
Apply the PPS model to the training period of the application data series to produce probabilistic forecast paths for **Finland** up to 2070.



6 Calibrate Finnish forecasts using error distribution
Calibrate the PPS forecast for the Finnish application series using the empirical forecast error distribution to produce the final forecast of the Finnish TFR up to 2070.

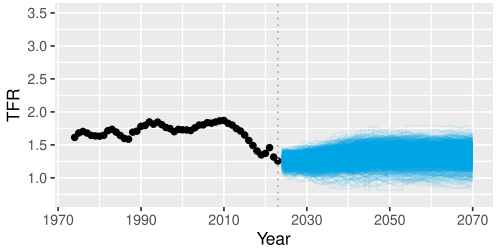


Figure 1: Graphical abstract of the main methodological steps of the analysis.

model.

Figure 1.5-6: We use the application data series and produce forecasts of Finnish TFR up to 2070 by first applying the PPS model to the training data of the application series (step 5) and then calibrating the

resulting forecast paths using the empirical error distribution (step 6).

In the remainder of this paper, we first introduce the PPS model and describe the details of the methodological steps. After we present the results of Finland's probabilistic TFR forecasts, we validate our forecast models and summarize the quality of the probabilistic forecasts with several calibration metrics.

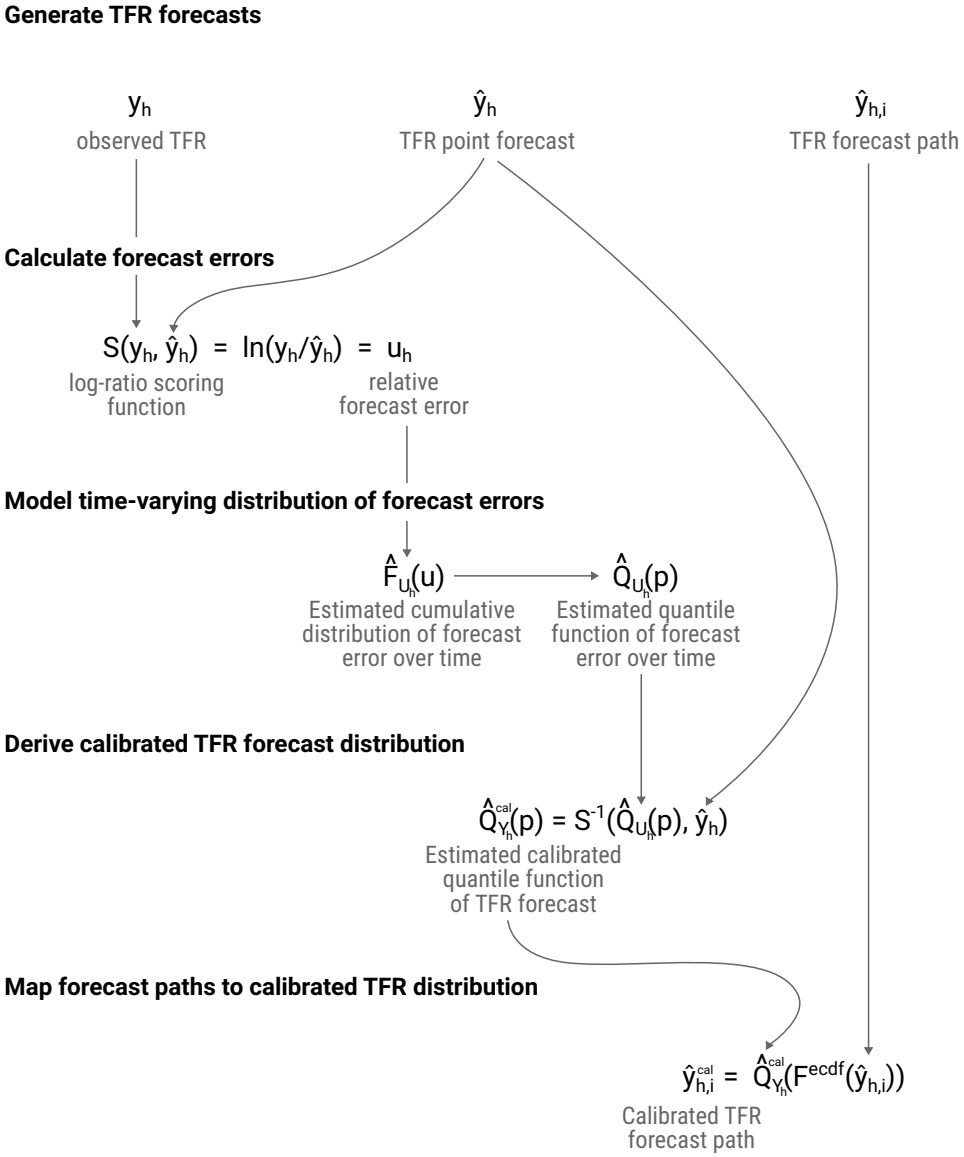


Figure 2: Illustration of the quantile-mapping calibration approach employed in this paper.

We source the time series of the fertility data from the Human Fertility Database (Human Fertility

Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) 2025) for the years 1944 to 2023 for Finland and Sweden, years 1968 to 2023 for Norway, and years 1946 to 2023 for Denmark.

Figure 2 presents a flowchart outlining the main steps of our approach for calibrating the probabilistic forecast paths using quantile-mapping. It also introduces the notation used in the remainder of the Methods section. For a detailed description of the methodological steps mentioned in the flowchart, see Sections 2.3 to 2.7.

2.1 Forecast Models

We use two different models to forecast fertility:

(a) *Postponement Time Series Model (PPS):* This scenario is based on the demographically meaningful assumption that fertility postponement in the Finnish population (the delay of childbirth to older ages) will continue, but will slow down and eventually stop. Specifically, we assume that the fertility postponement comes to an halt in 2050 and the mean age at childbirth approaches 33, which is approximately the current highest value reported in the Human Fertility Database (32.9 years in the Republic of Korea in 2020). This assumption is implemented in our forecast model by calculating the tempo-adjusted TFR for 2023 and forcing the TFR and the tempo-adjusted TFR to converge by 2050. After the convergence is completed, the TFR stays fixed until the end of the forecast period. The convergence of TFR and tempo-adjusted TFR to account for biological and social limitations was introduced in a model implemented in Nisén et al. (2020). In Appendix B, we describe the PPS model with its assumptions and parametrization in more detail. Using the PPS model, we first forecast 5,000 simulated time series of age-specific fertility rates (ASFRs, see Appendix B). Then, these ASFR forecast paths are aggregated into forecast paths for the TFR as the sum of the forecast ASFRs multiplied with the width of the age groups.

(b) *Naïve freeze-rates model:* In this model, the average ASFRs of the future are assumed to be at the level of the ASFRs at the jump-off year (last observed year). The probabilistic forecast paths are sampled from a random walk model with errors correlated across age (see Appendix E). Like in the PPS model, the forecast ASFRs are then aggregated into the TFR. This model serves as a naïve baseline model.

2.2 Split data

The time series of observed annual fertility values y_t for each country is partitioned into three data sets: calibration data $\mathcal{D}_{\text{cal}}$, validation data $\mathcal{D}_{\text{val}}$, and application data $\mathcal{D}_{\text{app}}$. See Figure 1.1 for a visual representation of the data splitting and the resulting data sets. The application data $\mathcal{D}_{\text{app}}$ holds a single Finnish series y_{Train} of 30 years. This data is used to forecast the Finnish fertility for 47 years up to 2070. We use $\mathcal{D}_{\text{cal}}$ to estimate the time-dependent distribution of the forecast errors around y_t which later is used for the calibration of the forecast paths. The calibration data from Finland and Sweden each have two cross-validation series with 30 years of training data y_{Train} and 47 years of testing data y_{Test}. Due to data constraints, we split the Norwegian calibration data into 11 smaller cross-validation series, each with 30 years of training data and 15 years of testing data. The hold-out set $\mathcal{D}_{\text{val}}$ is used to validate the properties of the different forecast models. The validation data contains the latest observed data series from Finland, Sweden, Norway and Denmark. We do not use data from Denmark for the calibration to increase the amount of left-out data to validate on. The validation series of the other countries are partially overlapping with series from the calibration data. Thus, excluding Denmark from the calibration data set reduces over-fitting. See Appendix A for the years covered by each of the data sets.

2.3 Generate uncalibrated TFR forecasts

Using the PPS model and the naïve model, we produce probabilistic forecast paths of TFR $\hat{y}_{h,i}$ over forecasting horizon $h \in 1, 2, \ldots, H$ and simulation draws $i \in 1, 2, \ldots, N$, where $N = 5,000$. The TFR forecasts are produced for the testing periods of each cross-validation series in $\mathcal{D}_{\text{cal}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{app}}$. These simulated forecast paths are the source for our uncalibrated prediction intervals. We define the point-forecast $\hat{y}_h$ as the median of the forecast paths.

2.4 Calculate forecast errors

We calculate the forecast error u_h for each point-forecast over the forecast horizon of the calibration data set $\mathcal{D}_{\text{cal}}$. To do so, we define a scoring function,

$$u_h = S(y_h, \hat{y}_h) = \ln(y_h/\hat{y}_h), \tag{1}$$

measuring the deviance between the observed TFR value y_h, and the point-forecast $\hat{y}_h$. The log-ratio scoring function, being a measure of relative error, is suitable for positive values which vary in scale, like the TFR (Alho et al. 2008). We pool the forecast errors from Finland, Sweden and Norway to achieve more robust estimates of the forecast error distribution by decreasing autocorrelation (Alho et al. 2008).

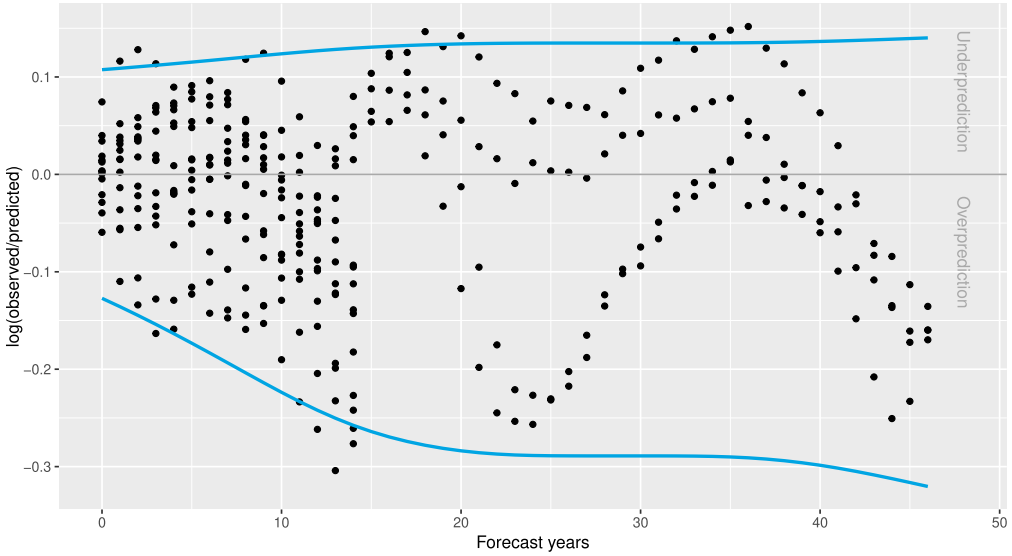


Figure 3: Forecast errors from the calibration data series with 95% quantiles of the modeled error distribution (blue).

2.5 Model time-varying distribution of forecast errors

Due to data limitations in the training data resulting in fewer long-term forecasts, it is necessary to model the distribution of the forecast errors rather than using the empirical distribution function. Figure 3, a scatter plot of the calculated forecast errors with the modeled error distribution at the 0.025 and 0.975 quantiles, illustrates the need for modeling.

Given the observed forecast errors u_h, we estimate the cumulative error distribution $\hat{F}_{U_h}(u) = \mathrm{P}(U_h \leq u)$ over the forecast horizon of $\mathcal{D}_{\text{cal}}$. We employ a *Time-varying Skew-normal model (SN)* for the distribution of the error,

$$U_h \sim \text{SkewNormal}(\mu, \sigma_h, \tau). \tag{2}$$

Using the AIC, we identified the Skew-Normal distribution as the most suitable error distribution among a set of alternatives, see Table 5. The μ parameter captures any systematic bias in the forecasting procedure, such as the tendency to overpredict future TFR. We model the dispersion parameter via a monotonically increasing P-spline over forecasting horizon: $\sigma_h = \exp(g_\beta(h))$. The monotonicity constraint ensures that the variance of the forecast error increases with time. Finally, the skewness parameter τ captures any asymmetry in the magnitude of over- and underprediction. The estimated distribution and quantile functions of forecast

errors, $\hat{F}_{U_h}$ and $\hat{Q}_{U_h}$, are analytically given by the Skew-Normal distribution. The distributional regression on the errors is performed via tha GAMLSS R package (Stasinopoulos and Rigby 2008). Angelopoulos et al. (2024) propose a similar approach and call the model of the forecast error distribution a scorecaster.

2.6 Derive calibrated TFR forecast distribution

Given the estimated distribution of forecast errors $\hat{F}_{U_h}$, we construct a calibrated TFR forecast distribution $\hat{F}^{\text{cal}}_{Y_h}$. Intuitively, if a given relative forecast error is exceeded 10% of the time, then the 90% quantile of the corresponding forecast distribution is equal to the central forecast times the 90% quantile of the relative forecast error, likewise for other p-quantiles of the error distribution. Formally, the p-quantile of the calibrated forecast distribution $\hat{F}^{\text{cal}}_{Y_h}$ is derived from the corresponding quantile of the error distribution $\hat{F}_{U_h}$ via

$$Q^{\text{cal}}_{Y_h}(p) = S^{-1}(\hat{Q}_{U_h}(p), \hat{y}_h), \tag{3}$$

where $S^{-1}(u_h, \hat{y}_h) = \hat{y}_h \exp(u_h)$ is the inverse of the log ratio scoring-function.

2.7 Map forecast paths to calibrated TFR distribution

We calibrate the forecasted TFR paths $\hat{y}_{h,i}$ such that their distribution at forecasting step h reflects the modeled historical distribution of forecast errors. In order to calibrate the forecast paths to the calibrated TFR distribution we map the p-quantile of any single forecast path at time h to the corresponding p-quantile of the calibrated TFR distribution. We sort the 5,000 forecasts at time h in increasing order, thus, the forecast with rank 1,000 will be mapped to the 20% quantile of the calibrated distribution, rank 2,500 will be mapped to the median, rank 4,000 will be mapped to the 80% quantile etc. For a visual representation of the forecast paths before and after the quantile-mapping, see Figure 8.

Formally, we calibrate the forecast fertility paths such that the marginal distribution of the calibrated paths is equal to the calibrated TFR distribution $\hat{F}^{\text{cal}}_{Y_h}$. This calibration is achieved by deriving calibrated paths as

$$\hat{y}^{\text{cal}}_{h,i} = \hat{Q}^{\text{cal}}_{Y_h}(F^{\text{ecdf}}_{Y_h}(\hat{y}_{h,i})), \tag{4}$$

where $F^{\text{ecdf}}_{Y_h}$ is the empirical cumulative distribution function over the non-calibrated forecast paths and $\hat{Q}^{\text{cal}}_{Y_h}$ is the quantile function of the calibrated TFR forecast distribution at h. To prevent the calibrated paths from taking infinite values, we adjust the empirical cumulative distribution function for the maximum values of the forecast paths (see Appendix D), so that calibrated paths remain finite.

2.8 Validate prediction intervals

The validation of the forecast paths is done on the validation data set $\mathcal{D}_{\text{val}}$ over the years of the testing data. Five metrics are evaluated and compared for the PPS forecast paths that have been calibrated on the empirical error distribution (PPS cal.), the un-calibrated forecast paths from the PPS model (PPS non-cal.) and the paths from the naïve model. We evaluate the forecast metrics over the full range of the forecast years (47 years) and over sub-sections of the forecast years (1 to 5, 6 to 15, 16 to 25, and 26 to 47 years). In the following, we index the observations in the validation series with subscript $j \in 1, \ldots, k$.

1. *Coverage:* We compare the nominal 95% coverage with the actual coverage of the corresponding quantiles of the forecast paths, i.e. the prediction intervals. The actual coverage is the fraction of observed TFR values that are inside the bounds of the declared interval. We calculate the coverage over the observations in the validation series as

$$\text{Cov} = \frac{\sum_{j=1}^{k} 1\{l_j \leq y_j \leq u_j\}}{k}, \tag{5}$$

where l_j and u_j are the 2.5% and 97.5% quantiles of the forecast distribution at observation j.

2. *Mean Interval Width:* We define the mean interval width as the mean difference between the 97.5% and 2.5% quantiles of the forecast distribution over the validation data:

$$\overline{W} = \frac{1}{k}\sum_{j=1}^{k} u_j - l_j. \tag{6}$$

3. *Mean Interval Score (MIS):* The interval score takes both the coverage and the width of the prediction intervals into account. Given the same actual coverage, a prediction interval that is on average wider is penalized by the interval score (Gneiting and Raftery 2007). The interval score is defined as follows:

$$\text{IS}_{\alpha,j}(l_j, u_j, y_j) = \begin{cases} (u_j - l_j) + \frac{2}{\alpha}(l_j - y_j) & \text{for} \quad y_j < l_j \\ (u_j - l_j) + \frac{2}{\alpha}(y_j - u_j) & \text{for} \quad y_j > u_j \end{cases} \tag{7}$$

where l_j and u_j are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles at validation point j, and $\frac{2}{\alpha}(l_j - y_j)$ and $\frac{2}{\alpha}(y_j - u_j)$ are the penalty terms for observations that fall below or above the bounds, respectively. The penalty is proportional to the $1 - \alpha$ level, set by us at 95%. We average the interval score across all observations over the forecast horizon to be able to compare the different forecast models (Bracher et al. 2021), resulting in the *Mean Interval Score* (MIS):

$$\text{MIS}_\alpha = \frac{\sum_{j=1}^{k} \text{IS}_{\alpha,j}}{k}. \tag{8}$$

4. Further, we use the *Continuous Ranked Probability Score* (CRPS), as implemented in Jordan et al. (2019), to evaluate the fit of the entire distribution of the simulated forecast paths over the forecast lengths. Given that the $N = 5,000$ draws from the forecast distribution at validation point j are sorted in increasing order such that $\hat{y}_{j,(i)}$ denotes the ith lowest path, the CRPS is calculated as

$$\text{CRPS}_j = \frac{2}{N^2}\sum_{i=1}^{N} (\hat{y}_{j,(i)} - y_j)(N1\{y_j < \hat{y}_{j,(i)}\} - i + \frac{1}{2}). \tag{9}$$

We then average over all validation points j,

$$\text{CRPS} = \frac{\sum_j \text{CRPS}_j}{k}. \tag{10}$$

5. Finally, we calculate the *Mean Absolute Percentage Error* (MAPE) as a validation measure of the point-forecasts via

$$\text{MAPE} = \frac{1}{k}\sum_{j=1}^{k} \left| \frac{y_j - \hat{y}_j}{y_j} \right| \times 100. \tag{11}$$

3 Results

In the following, first, we present the forecast results of the PPS and the naïve model, followed by the calibration of the PPS forecasts with the empirical error distribution. Second, we present the results from the validation analysis.

3.1 Forecast Results

Figure 4 contrasts the calibrated and non-calibrated forecast paths of the PPS model. Panel A shows the TFR paths from the PPS model without calibration on the historical error distribution. Due to the assumption of the PPS model that the fertility postponement will continue but slow down and eventually stop, the median of the forecast paths rises in the first 15 forecast years and then levels off to reach a TFR of 1.41 in 2070. The 95% prediction interval starts narrow around the first forecast and increases with

increasing forecast length, ranging from 0.73 to 2.69. Panel B shows the TFR trajectories derived from the PPS model calibrated on the historical forecast error distribution. In contrast to the non-calibrated forecasts, the 95% prediction interval is notably narrower, ranging from 1.02 to 1.62. Further, the median forecast is lower at a TFR of 1.35 in 2070. The changed shape and median of this forecast distribution is the result of calibrating the forecast paths to represent the historical forecast error distribution that starts wider, widens more slowly, and is skewed towards overestimation. We illustrated the different distributions of the forecast paths by adding the density plots for the years 2024 and 2070 (in grey colour). Overall, the calibrated PPS forecast predicts that Finland's TFR is likely to be higher in 2070 than last observed. The probability that the TFR will fall below the level of 2023 (1.26) is 29%.

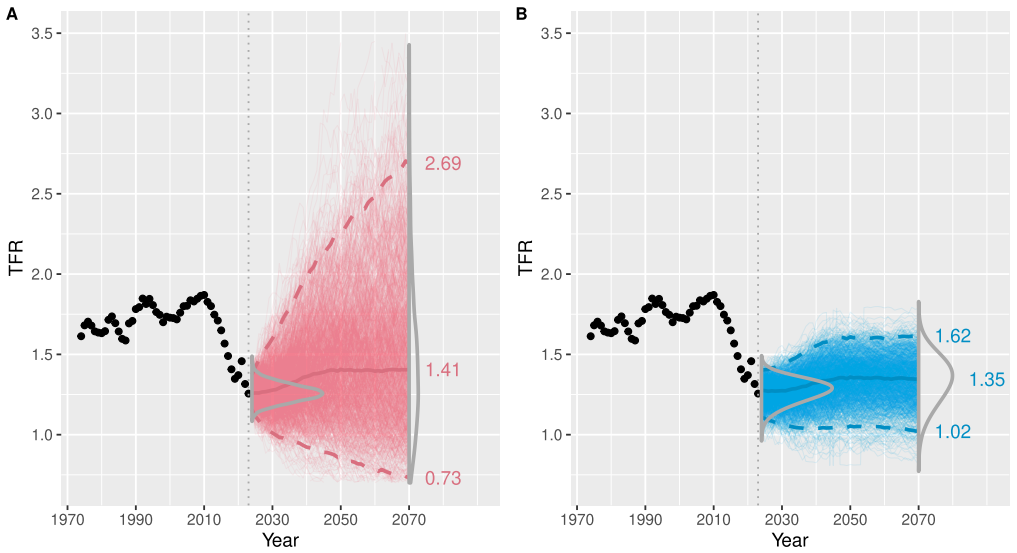


Figure 4: Observed TFR of Finland and 250 forecast paths from the PPS model (A) and after calibration on the historical error distribution (B), with median and 95% quantiles, and density in 2024 and 2070.

3.2 Forecast Validation

The validation analysis shows a higher performance of the calibrated forecast paths for the validation data, compared to the non-calibrated PPS and naïve forecasts. Table 1 summarizes different forecast metrics for the full forecast horizon and shorter forecast periods.

The calibrated PPS forecasts have the smallest CRPS for each forecast horizon except the period 16-25 years. In terms of MIS, the calibrated PPS forecasts consistently outperform the other models. Over the full forecasting period, the calibrated PPS model achieves similar coverage as the non-calibrated and the naive model, but with substantially narrower interval width. However, the coverage varies with forecasting period. Over the first five forecast years, although the average width of the prediction intervals across models is similar, the coverage of the calibrated PPS paths is substantially better than that of the naive model and the non-calibrated PPS model. At 6 to 15 years into the forecast, all three models suffer from under-coverage. In contrast, at forecast interval 16 to 25, all three models exhibit overly conservative prediction intervals. Those overly wide prediction intervals persist into the furthest forecast period for the non-calibrated PPS model and the naive model. The calibrated PPS model, however, achieves a perfect 95% coverage at 26 to 47 forecast years. While the calibrated PPS paths are the best distributional forecast, MAPE identifies the naive model as the best point-forecast.

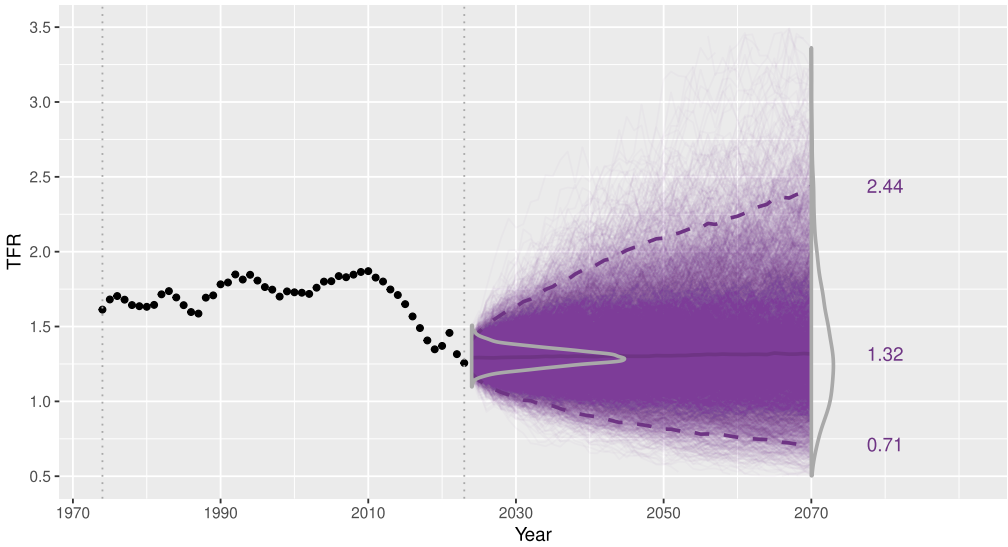


Figure 5: Observed TFR of Finland and 250 forecast paths from the naïve model, with median and 95% quantiles, and density in 2024 and 2070.

Table 1: Evaluation of model calibration.

Forecast Period	Model	Cov.	W	MIS	CRPS	MAPE
full	PPS (cal.)	90%	*0.71*	*1.09*	*0.13*	11.3%
	PPS (non-cal.)	*92%*	1.34	1.73	0.17	13.7%
	Naïve	*92%*	1.24	1.59	*0.13*	*10.1%*
1-5 Years	PPS (cal.)	*80%*	0.45	*0.86*	*0.11*	9.9%
	PPS (non-cal.)	67%	*0.44*	1.28	*0.11*	*9.1%*
	Naïve	67%	0.44	1.36	0.12	9.4%
6-15 Years	PPS (cal.)	73%	*0.60*	*2.06*	*0.16*	13.8%
	PPS (non-cal.)	*77%*	0.86	2.26	0.17	14.6%
	Naïve	*77%*	0.85	2.02	*0.16*	*13.7%*
16-25 Years	PPS (cal.)	*100%*	*0.75*	*0.75*	0.12	10.6%
	PPS (non-cal.)	*100%*	1.30	1.30	0.16	12.9%
	Naïve	*100%*	1.19	1.19	*0.11*	*9.15%*
26-47 Years	PPS (cal.)	*95%*	*0.80*	*0.78*	*0.13*	10.9%
	PPS (non-cal.)	100%	1.78	1.78	0.18	14.6%
	Naïve	100%	1.62	1.62	*0.13*	*9.03%*

Note: Evaluation for the full forecast length of 47 years and different forecast lengths using the Coverage (Cov.), the Mean Interval Score (MIS) and the average width (W) of the prediction intervals for a nominal coverage of 95%, the Continuous Ranked Probability Score (CRPS) of the full forecast distribution, and the Mean Absolute Percentage Error (MAPE) of the median forecasts.

4 Discussion

Fertility rates are prone to sudden shifts in trend. This makes them harder to forecast than, e.g., death rates. We do address this issue by only asking a modest question: What forecasting errors are to be expected under the continuation of the Nordic fertility regime which was prevalent the last 30 years? We introduced a novel approach to calibrating probabilistic forecasts by combining a scorecaster to estimate an empirical forecast error distribution and the quantile-mapping of simulated forecast paths. By incorporating empirical forecast errors into the forecast uncertainty, we provide a data-driven estimate of the forecast uncertainty. Using a model of the empirical forecast error distribution allows for greater generalizability through smoothing and interpolation. The chosen skew-normal model does so using only four parameters. Moreover, the calibration of forecast paths with quantile-mapping, rather than providing uncertainty intervals, offers more flexibility for downstream modeling. Our results demonstrated strong performance in the validation analyses. These findings reflect the strength and versatility of the techniques and their combined use.

We presented probabilistic forecasts of the Finnish Total Fertility Rate (TFR) from 2024 to 2070. We use two different models to forecast the TFR of Finland. The first model is a scenario-based approach that assumes that the fertility postponement, i.e. the delay of childbearing to older ages, that started in Finland in the 1970s, will slow down and eventually stop. The second scenario-based model serves as a naïve baseline and assumes that the most recently observed age-specific fertility rates will remain constant. We then calibrate 5,000 paths of future TFR values so that their distribution matches the distribution of the modeled historical out-of-sample errors. These TFR paths can be easily integrated into further analyses, such as population projection models or economic and health planning models.

The paths of the PPS model which have been calibrated on the historical error distribution predict the Finnish TFR to increase until 2050 and then level off. The central forecast reaches a value of 1.35 in 2070 (95% PI [1.02, 1.62]). The validation analysis shows that the uncertainty around this calibrated forecast is better calibrated in terms of the Mean Interval Score and the Continuous Ranked Probability Score, compared to the non-calibrated paths from the PPS model (median 1.41, 95% PI [0.73, 2.69] in 2070) and the naïve baseline model (median 1.32, 95% PI [0.71, 2.44] in 2070).

Similar to our results, the latest edition of the UN World Population Prospects (UNWPP 2024) projects Finland's TFR to slightly increase in their median variant and to level off at 1.51 children by 2100, reaching a value of 1.47 by 2070. The 95% prediction interval around this median ranges from 0.89 to 2.0 children in 2070, which is wider than the 95% prediction intervals of our results. The UN produces probabilistic projections with country-specific assumptions based on the country's past experience (see United Nations, Department of Economic and Social Affairs, Population Division 2024, for detailed methodology).

Calibrating time series forecasts using historical data is data intensive. This need for long available time series is the main limitation of the proposed methodology. In addition, the scenario nature of the PPS model restricted the applicability of the model to periods where fertility change is strongly affected by tempo effects and the main assumption of continuing but slowing down fertility postponement holds. Therefore, we extended the training data set for Finland by including data from Sweden and Norway, which have similar childbearing patterns in terms of timing and level of fertility. We also allowed the cross-validation series to overlap to increase the number of series. However, the overlapping of the cross-validation series carries the risk of over-fitting the model of the forecast error distribution. To mitigate this, we included data from Denmark that were only used in the validation analyses and not in the calibration of the forecasts.

Our approach of using the empirical forecast error distribution to calibrate future forecasts is considered a split-conformal or inductive conformal prediction approach in the vast conformal prediction (CP) literature (see Fontana et al. 2023, for a review of approaches). The main characteristic of this approach is to employ a calibration data set so that a "general rule" can be derived from which the predictions are made. In our case, this rule is the model of the empirical error distribution. In contrast, full-conformal prediction, also called transductive conformal, does not employ a general rule and jumps directly from the observation to the prediction. However, this approach tends to be computationally inefficient and is not suitable for long data series. As a response, split-conformal approaches have been developed (Papadopoulos et al. 2007). In our case, in addition to higher computational efficiency, using the split-conformal approach allows for more generalizability in the face of data scarcity.

Weighted-conformal prediction uses sample weights to treat more recent observations as more relevant (Tibshirani et al. 2019; Barber et al. 2023). Further, Lei et al. (2018) propose an extension to their split-

conformal approach resulting in a locally-weighted conformal prediction to produce prediction bands with locally varying width. Compared to the locally-weighted CP, we achieve a similar goal with our approach, because the forecast paths' distribution gets wider with increasing forecast length and is not symmetrically distributed around the central forecast. This is due to the use of the model of the empirical forecast errors to calibrate the forecast paths. The difference is that our outcome is a calibrated set of forecast paths, whereas Lei et al. (2018) produce prediction intervals.

Keilman and Pham (2004) introduce the use of published historical forecasts from statistical agencies and other official sources to calculate forecast errors and thus derive empirical prediction intervals. In this way, the uncertainty of expert forecasts in the past informs the uncertainty of the forecasts today, regardless of the methodology used to produce the historical forecasts. This type of empirical prediction intervals can be a valuable tool for scenario-based forecast models, where the lack of applicability to historical data limits the data availability for the training data.

The underlying assumption that allows us to use empirical forecast errors to derive measures of forecast uncertainty is that future forecast errors will resemble past errors. One might ask, however, why this should be the case. Alho et al. (2008) argue that "[...] if one does not believe that they will be, it is necessary to provide arguments as to why the future is expected to be different from the past". This raises the question of whether the PPS model is just as valid now as it would have been in the past. Current Finnish fertility trends, prominently the postponement of fertility, have been observed since the late 1960s. We believe that applying our scenario-based PPS model to current and past periods is a valid choice. We see no reason why the historic forecast errors of the PPS model should not be used to inform the uncertainty of the current PPS forecasts. Demographers have for a long time been aware of the distorting impact of changes in fertility timing on period fertility (Hajnal 1947; Bongaarts and Feeney 1998). While not explicitly predicting the end of fertility postponement, they illustrated how fertility rises due to the slowing down of fertility postponement or catching up of postponed births. This has later been referred to as "the third phase of fertility recuperation" (Sobotka 2017). Further, Finnish research of the past expected that the observed fertility postponement would not go on forever: "At present it seems reasonable to assume in long-range studies that fertility will stabilize at the level prevailing at the end of the 1980s" (Auvinen 1989, p. 54). This belief is also reflected in the "high"-scenario of the 1984 population projection of Statistics Finland (Hämäläinen and Honkanen 1984).

Although the validation analyses have shown that the forecasts calibrated on the empirical error data perform better than the non-calibrated ones, the calibration scores are not perfect. The main problem is the under-coverage of the prediction intervals for forecasts up to 15 years ahead due to overprediction of the TFR. Although we have taken this forecast bias into account when modeling the forecast error distribution, the problem persists to some extent as revealed in the validation. A possible reason for these results is the lack of data, which leads to less robust validation results, as we only have four data series available for the validation analysis.

The results of this study show how forecast error analysis and quantile-mapping serve to improve the quality of probabilistic demographic forecasts. We would like to emphasize that this is a flexible methodology that can be applied to all kinds of demographic (or non-demographic) measures, regardless of the type of forecast model or outcome. In this study, we have applied it to a scenario-based model for forecasting Finland's Total Fertility Rate. However, any type of forecast model, such as extrapolation models, can be calibrated using the presented methodology. The critical component for successful application is a long time series of historical data to which the chosen forecast model can be applied. We encourage researchers to use probabilistic forecast methods, to be transparent about their assumptions, and to calibrate and validate their results using historical data. Time has shown that demographic forecasts made by researchers in the past, to the best of their knowledge, have turned out to be wrong. We see no reason why current forecasts should be any different. It is intuitive to us, therefore, to use the knowledge of past forecast errors to our advantage and let it inform our forecasts of today.

Conflicts of interest

The authors declare that they have no competing interests.

Funding

R.D. was supported by the Finnish Centre for Pensions (ETK2023031).

J.H. was supported by the Strategic Research Council (SRC) of the Academy of Finland, FLUX consortium (Family Formation in Flux—Causes, Consequences, and Possible Futures), decision numbers 364374 and 364375, and by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 101019329).

M.M. was supported by the Strategic Research Council (SRC), FLUX consortium, decision numbers 364374 and 364375; by the National Institute on Aging (R01AG075208); by grants to the Max Planck – University of Helsinki Center from the Max Planck Society (Decision number 5714240218), Jane and Aatos Erkko Foundation, Faculty of Social Sciences at the University of Helsinki, and Cities of Helsinki, Vantaa and Espoo; and the European Union (ERC Synergy, BIOSFER, 101071773). Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Data availability

The research presented in this article is fully reproducible. We have made our codebase available to the scientific community. You can access the complete codebase, including scripts and documentation, at the following link: `https://github.com/RicardaDuerst/empPIs_Nordic_Fertility_Forecasts`. In addition, the data used for our analyses is sourced from openly accessible and publicly available datasets. Researchers interested in reproducing our work or conducting further investigations can freely download the data from the following link: `https://www.humanfertility.org`. The availability of open-access data ensures transparency and promotes collaboration within the scientific community.

Author contributions statement

R.D.: Conceptualization, Methodology, Formal analysis, Software, Visualization, Writing - Original Draft. J.S.: Conceptualization, Methodology, Software, Supervision, Writing - Review & Editing. J.H.: Methodology, Software, Writing - Review & Editing. M.M.: Conceptualization, Methodology, Supervision, Writing - Review & Editing, Funding acquisition.

Acknowledgments

The authors would like to thank Nico Keilman, Rob Hyndman and Marie-Pier Bergeron-Boucher for discussions on empirical uncertainty quantification. R.D. gratefully acknowledges the resources provided by the International Max Planck Research School for Population, Health and Data Science (IMPRS-PHDS).

References

Alho, J. M., Cruijsen, H., and Keilman, N. (2008). Empirically based specification of forecast uncertainty. In Alho, J. M., Hougaard Jensen, S. E., and Lassila, J., editors, *Uncertain demographics and fiscal sustainability*, pages 34–54. Cambridge University Press, Cambridge.

Alho, J. M. and Spencer, B. D. (2005). Uncertainty in demographic forecasts: Concepts, issues, and evidence. In *Statistical demography and forecasting*, pages 226–268. Springer.

Angelopoulos, A., Candes, E., and Tibshirani, R. J. (2024). Conformal pid control for time series prediction. In *Advances in neural information processing systems 36*.

Auvinen, R. (1989). Finland's low fertility and the desired recovery. *Finnish Yearbook of Population Research*, 27(1989):53–59.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

Bongaarts, J. and Feeney, G. (1998). On the quantum and tempo of fertility. *Population and development review*, pages 271–291.

Bracher, J., Ray, E. L., Gneiting, T., and Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618.

Cannon, A. J. (2018). Multivariate quantile mapping bias correction: an n-dimensional probability density function transform for climate model simulations of multiple variables. *Climate dynamics*, 50(1):31–49.

Cohen, J. E. (1986). Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. *Demography*, 23(1):105–126.

Fontana, M., Zeni, G., and Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23.

Gammerman, A., Vovk, V., and Vapnik, V. (1998). Learning by transduction. arXiv:1301.7375.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldstein, J. R. (2006). How late can first births be postponed? some illustrative population-level calculations. *Vienna Yearbook of Population Research*, pages 153–165.

Hajnal, J. (1947). The analysis of birth statistics in the light of the recent international recovery of the birth-rate. *Population Studies*, 1(2):137–164.

Hellstrand, J., Nisén, J., and Myrskylä, M. (2020). All-time low period fertility in finland: Demographic drivers, tempo effects, and cohort implications. *Population Studies*, 74(3):315–329.

Human Fertility Database (HFD). Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria) (2025). Available at: https://humanfertility.org/.

Hämäläinen, H. and Honkanen, O. (1984). Väestöennusteet lääneittäin kuntamuodon mukaan 1980 - 2000 [Population projections by region and municipality type 1980–2000]. Statistics Finland, https://www.doria.fi/bitstream/handle/10024/185495/xmui$_9$1$_1$984$_d$*ig.pdf?sequence* = 1.

Institut national de la statistique et des études économiques (2008). Demographic projections: main mechanisms and feedback from the french experience. https://www.insee.fr/fr/metadonnees/source/fichier/projections_population_mecanismes.pdf.

Jordan, A., Krüger, F., and Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90:1–37.

Keilman, N. (2018). Probabilistic demographic forecasts. *Vienna Yearbook of Population Research*, 16:25–36.

Keilman, N. and Pham, D. Q. (2004). Time series based errors and empirical errors in fertility forecasts in the nordic countries. *International Statistical Review*, 72(1):5–18.

Lee, R. D. (1998). Probabilistic approaches to population forecasting. *Population and Development Review*, 24:156–190.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Nisén, J., Hellstrand, J. I. S., Martikainen, P., and Myrskylä, M. (2020). Hedelmällisyys ja siihen vaikuttavat tekijät suomessa lähivuosikymmeninä [Fertility and factors affecting it in the coming decades in finland]. *Yhteiskuntapolitiikka*, 85(4):358–369.

Papadopoulos, H., Vovk, V., and Gammerman, A. (2007). Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395.

Qian, W. and Chang, H. H. (2021). Projecting health impacts of future temperature: a comparison of quantile-mapping bias-correction methods. *International journal of environmental research and public health*, 18(4):1992.

Rothman, K. J., Wise, L. A., Sørensen, H. T., Riis, A. H., Mikkelsen, E. M., and Hatch, E. E. (2013). Volitional determinants and age-related decline in fecundability: a general population prospective cohort study in denmark. *Fertility and sterility*, 99(7):1958–1964.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3):371–421.

Smith, S. K. and Sincich, T. (1988). Stability over time in the distribution of population forecast errors. *Demography*, 25(3):461–474.

Sobotka, T. (2017). Post-transitional fertility: childbearing postponement and the shift to low and unstable fertility levels. *Vienna Institute of Demography Working Papers*, 2017(1).

Stasinopoulos, D. M. and Rigby, R. A. (2008). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23:1–46.

Statistics Finland (2024a). Official statistics of Finland (osf): Births [online publication]. https://stat.fi/en/publication/cln3ad0zibipb0cutxy95o145.

Statistics Finland (2024b). Population projection: documentation of statistics. https://stat.fi/en/statistics/documentation/vaenn.

Statistics Sweden (2024). Alternative population projections for sweden 2024–2070, demographic reports 2024:4.

Stoto, M. A. (1983). The accuracy of population projections. *Journal of the American Statistical Association*, 78(381):13–20.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in neural information processing systems 32*.

United Nations, Department of Economic and Social Affairs, Population Division (2024). World popoulation prospects 2024: Methodology of the united nations population estimates and projections.

Wilke, C. (2018). *German pension Reform: On Road towards a Sustainable Multi-Pillar System*. Peter Lang International Academic Publishers.

Williams, W. H. and Goodman, M. L. (1971). A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, 66(336):752–754.

Zeman, K. (2024). The fertility assumptions for the population projection of the czech republic of the czech statistical office from 2023. *Demografie*, 66(4):280–293.

Appendices

A Cross-validation series of split data sets

Table 2: Calibration data D_{cal}

Country	Series Nr.	Training Period Start	Training Period End	Testing Period Start	Testing Period End
Finland	1	1944	1973	1974	2020
	2	1945	1974	1975	2021
Sweden	1	1944	1973	1974	2020
	2	1945	1974	1975	2021
Norway	1	1968	1997	1998	2012
	2	1969	1998	1999	2013
	3	1970	1999	2000	2014
	4	1971	2000	2001	2015
	5	1972	2001	2002	2016
	6	1973	2002	2003	2017
	7	1974	2003	2004	2018
	8	1975	2004	2005	2019
	9	1976	2005	2006	2020
	10	1977	2006	2007	2021

Table 3: Validation data D_{val}

Country	Training Period Start	Training Period End	Testing Period Start	Testing Period End
Finland	1946	1975	1976	2022
Sweden	1946	1975	1976	2022
Norway	1978	2007	2008	2022
Denmark	1946	1975	1976	2022

Table 4: Application data D_{app}

Country	Training Period Start	Training Period End	Forecast Period Start	Forecast Period End
Finland	1994	2023	2024	2070

B The Postponement Time Series Model

The Postponement Time Series Model (PPS) is a scenario-based forecast model parameterized by expert opinion. It was developed to produce probabilistic forecasts of the Finnish Total Fertility Rate (TFR) using age-specific fertility rates (ASFRs) for 5-year age groups up to 2070. This scenario is based on the demographically meaningful assumption that the Finnish fertility postponement will continue, but will slow down and eventually stop. Due to biological factors, fertility postponement cannot continue forever, as the ability to become pregnant declines with age, and does so especially fast for women in their mid to late 30s (Rothman et al. 2013). According to Goldstein (2006), the mean age at first birth could plausibly rise to around 33 years before reaching biological and social limits. Further, Sobotka (2017) suggested that fertility postponement would continue only for another two to three decades. Based on these upper limits, we make

the following assumptions about fertility postponement in our model: First, the forecast TFR converges with a target TFR by a pre-defined target year. Second, the mean age at childbirth (MAB) approaches a target MAB by the same target year.

The target TFR is defined as the tempo-adjusted TFR of the jump-off year (last observed year before the forecast starts). To determine the target value for the MAB we assume that the historically observed mean annual increase of the Finnish MAB (0.1 years per year) decreases gradually until it stops by the time the target year is reached. For the Finnish application of the PPS model, the target year is defined as 2050 (the 28th forecast year), the target MAB is 32.88 and the target TFR is 1.34.

After defining the target values for TFR and MAB, we employ an optimization procedure to find a set of target ASFRs that match these target summary statistics under the constraint that the bell-shaped age-profile of fertility should be preserved.

To produce the central forecast of ASFRs, the time series from the jump-off ASFRs towards the target ASFRs is modeled using a logistic curve with a growth rate of 0.3 and the steepest growth at the mid-point between the jump-off year and the target year. This ensures a smooth transition from the jump-off year to the target year. After the target year is reached, the ASFRs remain fixed until the end of the forecast period, resulting in a total forecast length of 47 years.

After constructing the central forecast, we use an age-correlated random walk with drift model (RWD) to produce probabilistic forecasts. We simulate individual forecast paths based on the covariance of the ASFRs of the 30 years prior to the forecast period. Using the PPS model, we first forecast 5,000 individually simulated time series of age-specific fertility rates (ASFRs). Finally, the ASFR forecast paths are transformed into forecast paths for the TFR, which is defined as the sum over ASFRs multiplied with the width of the age groups: $\text{TFR}_t = 5 * \sum_{x=15}^{45} \hat{y}_{x,t}$.

C Selection of scorecaster model

We tested four different specifications for modeling the time-varying distribution of the PPS forecast error (scorecaster model): 1. A Normal distribution with fixed mean and time-varying dispersion (PPS-Normal); 2. a Student's t distribution with fixed mean, fixed tail parameter, and time-varying dispersion (PPS-Student-t); 3. a Skew-Normal distribution with fixed mean, fixed skewness, and time-varying dispersion (PPS-Skew-Normal); 4. a Skew-Student's t / Azzalini distribution with fixed mean, fixed tail parameter, fixed skewness and time-varying dispersion (PPS-Skew-Student-t). We specified the time-varying dispersion as a P-spline monotonically increasing over the forecasting horizon.

Table 5 shows the effective degrees of freedom, the log-likelihood, the deviance and the AIC over the calibration set of forecast errors for all four scorecaster models.

Table 5: Model fit statistics of four alternative specifications for the time-varying distribution of forecast errors

	DF	**LL**	**Dev**	**AIC**
PPS-Normal	3.3	655.3	-1310.5	-1303.9
PPS-Student-t	4.4	671.7	-1343.3	-1334.4
PPS-Skew-Normal	4.4	687.1	-1374.1	-1365.3
PPS-Skew-Student-t	5.4	687.1	-1374.1	-1363.3

Note: Statistics derived from the calibration data. All four models were specified with a constant mean, a time-varying dispersion implemented via monotonic P-splines, and fixed skewness (Skew) and df (Student-t) parameters.

D Adjustment of the empirical cumulative distribution function

To prevent the calibrated paths from taking infinite values, we adjust the empirical cumulative distribution function $F^{\text{ecdf}}_{\hat{Y}_h}$ for the maximum values of the forecast paths, $\hat{y}_{h,max}$. Instead of $F^{\text{ecdf}}_{\hat{Y}_h}(\hat{y}_{h,max})) = 1$, which

results in $\hat{Q}^{\text{cal}}_{Y_h}(F^{\text{ecdf}}_{Y_h}(\hat{y}_{h,max})) = \infty$, we set $F^{\text{ecdf}}_{Y_h}(y_{h,max})) = 1 - 0.5\frac{1}{N}$. This adjustment places the empirical cumulative distribution function's value halfway between 1 and the value for the second-highest $\hat{y}^{\text{cal}}_{h,i}$, ensuring that the calibrated paths remain finite.

E ASFR forecasts

Figure 6 and 7 show an intermediate outcome of the PPS and the naïve model in form of the forecast paths of the age-specific fertility rates (ASFR) of Finland together with the observed ASFR, by 5-year age groups. The observed ASFR values, represented by black dots, cover the period from 1944 to 2023, followed by 250 of the 5,000 forecast paths up to the year 2070. As described in Section 2, the ASFR paths are then used to calculate the forecast TFR paths, shown in Figure 5 for the naïve model, and in Figure 4 (panel a) for the PPS model.

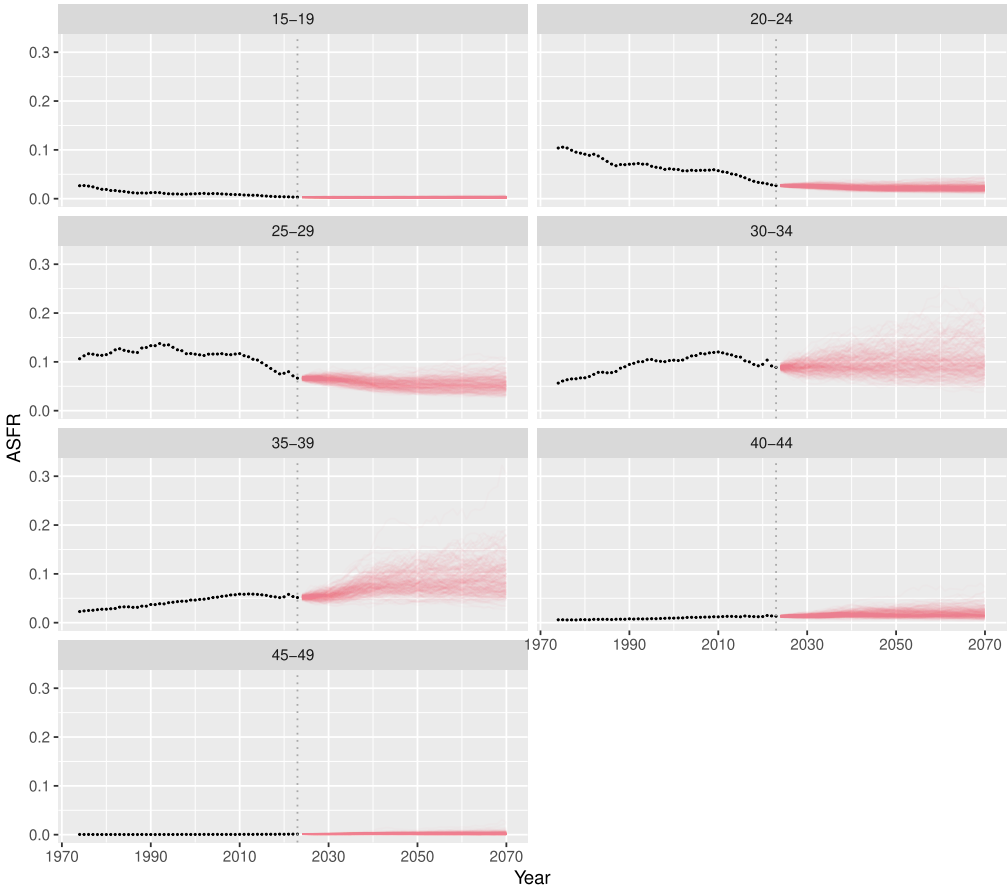


Figure 6: Observed ASFR of Finland and 250 forecast paths from the PPS model by 5-year age groups for 2024 through 2070.

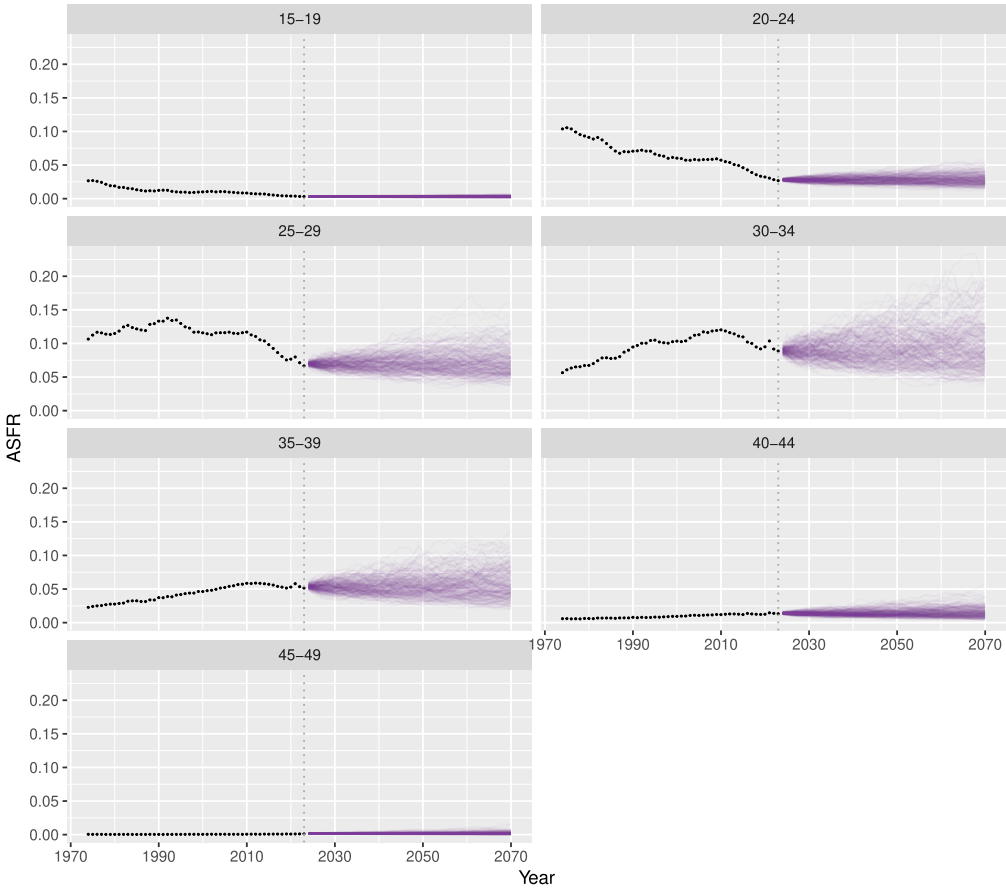


Figure 7: Observed ASFR of Finland and 250 forecast paths from the naïve model by 5-year age groups for 2024 through 2070.

F Quantile-mapping

Figure 8 depicts 500 forecast paths from the PPS model (applied to the validation data series), before and after quantile-mapping of the forecast paths to the calibrated TFR distribution. The same forecast path has been highlighted in both panels to visualize how quantile-mapping transforms an individual forecast path.

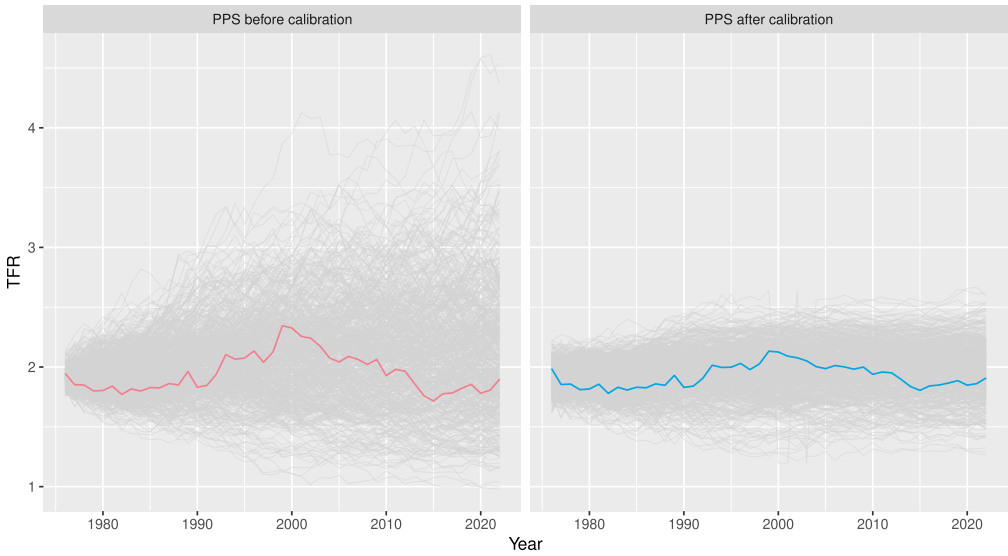


Figure 8: Forecast paths from the PPS model (applied to the validation data series), before and after quantile-mapping.