Improving the robustness, accuracy, and utility of chemistry-climate model ensembles

Matt Amos, MPhys
Lancaster Environment Centre
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

February, 2022

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. Excerpts of this thesis have been published in journals, as indicated within. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography. A rough estimate of the word count is: 40000.
Matt Amos

Improving the robustness, accuracy, and utility of chemistry-climate model ensembles

Matt Amos, MPhys.
Lancaster Environment Centre, Lancaster University
A thesis submitted for the degree of *Doctor of Philosophy*. February, 2022

Abstract

Ensembles of chemistry-climate models (CCMs) are fundamental for the exploration of the chemistry-climate system. A particular focus of chemistry-climate modelling is stratospheric ozone, whose concentrations have been decreased by anthropogenic releases of ozone depleting substances. In conjunction with observational data, CCM ensembles have been relied upon to simulate historic effects of ozone depletion and to project future ozone recovery.

However, many widely used ensemble analysis methods are simplistic and are based upon incorrect assumptions about the design of the ensemble. Multi-model means used to construct future ozone projections do not account for variable model performance or similarity and therefore give biased and inaccurate projections. Similarly, simplistic linear regression methods used to infill historic ozone records underestimate interannual variability and are inaccurate in regions of sparse data coverage. Moreover, given advances in machine learning and data science and their increased use in environmental science, it is timely to apply more advanced tools to CCM ensembles.

To address this methodological deficit, this thesis presents a set of novel tools to improve the predictions and projections from CCM ensembles of stratospheric ozone. A process-based weighted mean is developed which accounts for model performance and similarity in CCM ensembles. This improvement over pre-existing methods was used to generate accurate ozone hole recovery projections. This thesis also developed a Bayesian neural network (BNN) which fuses together CCMs with observational data to produce accurate and uncertainty-aware predictions. The BNN framework was used to produce historic continuous datasets of total ozone column and vertically resolved ozone, and represents a significant improvement in methods used to ensemble models.

Though designed for CCM ensembles these flexible tools have the potential to be applied to other environmental modelling disciplines to improve the accuracy of projections, better understand uncertainty and to make better use of historic observations.

Acknowledgements

I wish to recognise and appreciate the input of numerous people from across my academic, social, and family networks. For those named or otherwise acknowledged, please know I am immeasurably grateful the advice, support and friendship which enabled my PhD.

First and foremost, I owe an enormous and sincere thanks to Dr Paul Young, whose expert mentoring, encouragement, provision of beer, and down to earth advice was instrumental throughout my PhD journey. His constructive and thoughtful input was vital through many ups and downs. I must also recognise his useful proficiency in locating every last comma splice and punctuation misuse. Similarly, I extend great gratitude to Dr Scott Hosking who was immensely encouraging of my research interests and provided a wonderfully supportive research environment. I could not have wished for two better mentors and supervisors, without whom my introduction into research would have been far less successful.

At Lancaster I am grateful for all the members of our research group and the ENVISION cohort who provided a rigorous sounding board for ideas and a sympathetic ear throughout the trials of the Covid pandemic. I would like to particularly thank, Dr Tabish Ansari for many insightful and engaging late afternoon discussions about the more philosophical side of science and the wider world. Additionally, I offer many thanks to Ushnish Sengupta for helping me traverse new scientific disciplines and for contributing to work within this thesis.

I have many people to thank for providing welcome distractions for when research regularly proved troublesome: two amazingly inspirational, supportive and vivacious pals Ben Cianchi and Jonathan Doyle who, through their wonderful friendship, have given me unbelievable support over the years, as well as many memorable adventures; the climbers and cavers in my life (Jade Bowling, Dexter Williamson, Martin Paley, Sam O'Rourke, John Hartley, and many others) with whom many splendid days were spent playing around on mountains and crags or exploring the subterranean depths of Yorkshire; and a strong network of wonderful humans who have been there for me countless times (Alex Cartwright, Katie Madine, Aoife Marrett, Jack Mckay).

To my greatest support, Alex Welsh, who offers her ceaseless affirmation and endless encouragement whatever the weather, I am immensely grateful. Thank you for always being there for me.

Lastly, to my family who always provide the stable and supportive foundations from which everything else is constructed upon, I offer my heartfelt thanks. Since I was young you instilled within me a strong curiosity which enabled my scientific passion and interest in the mechanisms that underpin the world and the people within it. I will forever be grateful for this.

For my sister, Hope – a source of unparalleled strength, determination, and courage. This thesis is as much yours as it is mine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis sits at the confluence between statistics, machine learning (ML) and atmospheric science. Access to increasingly sophisticated ML techniques has improved for atmospheric scientists, driven by a surge in off-the-shelf tools and codes, and cheaper access to accelerated computing required in ML applications (Karpatne et al., 2018; Virts et al., 2020). ML tools have good investigative and predictive powers and therefore have much to offer atmospheric science (Ford et al., 2016) which regularly requires analyses of large observational and model datasets (e.g., Taylor et al., 2012; Eyring et al., 2016b; Morgenstern et al., 2017). These datasets have the 4 Vs of big data – volume, variety, veracity and velocity – predicated by researchers' desires to simulate and observe the earth system at continually increasing resolution, speed and complexity (Guo et al., 2015). However, much atmospheric data, particularly ensembles of models, is still analysed and used with traditional methods, as communities have not yet widely adopted data science research tools to aid data exploration and boost predictive ability (Faghmous and Kumar, 2014). Given the complexity and nuances of atmospheric data, modern data science and ML cannot be applied blindly, instead requiring interdisciplinary efforts within environmental data science (Kanevski, 2009; Maganathan et al., 2020). Environmental data science, the merger of environmental and data sciences, explores how new methods and data-driven approaches from statistics and ML can bring greater insight into environmental data and problems, helping to tackle many broad and important global challenges (Blair et al., 2019).

This thesis presents a collection of sophisticated data science techniques to make

better use of climate model ensembles, with a particular focus on atmospheric composition. Atmospheric science relies on ever-increasing model complexity, both computationally and in the physical processes they represent (Collins et al., 2017; Morgenstern et al., 2017). Despite the increase in quantity and quality of model output, many of the techniques for analysing the data have not undergone the same increase in sophistication (IPCC, 2010; Faghmous and Kumar, 2014). The set of environmental data science tools presented in this thesis seeks to address this methodological deficit, through the development of novel and broadly applicable approaches focussed on ensembles of chemistry-climate models (CCMs), and by doing so, improve our understanding of the historic and future evolution of ozone in the stratosphere.

Within this introduction the three overarching themes of the thesis are introduced and summarised: ozone in the upper atmosphere, which forms the atmospheric theme to which the developed data science methods are applied; ML and data science and why these approaches can be beneficial in atmospheric science; and ensembles of physical models and how modern techniques can aid analysis of these large datasets. The introduction concludes with a short summary of the thesis contributions where the remaining thesis structure is also laid out. Subsequent chapters are self-contained pieces of work each containing separate introductions and literature reviews which expand upon the following literature review.

1.1 Stratospheric ozone

This thesis explores and develops data science tools focussing on ozone in the upper atmosphere, although the tools are widely applicable. Stratospheric ozone is chosen because it has been an active research area of continued interest for the last 40–50 years (Solomon, 1999; WMO, 2018), resulting in a large range of observational data (e.g., Tegtmeier et al., 2013) and CCM simulations necessary for data-driven approaches (e.g., Eyring et al., 2010, 2013; Morgenstern et al., 2017). The system is relatively well understood, prompted initially by the immediate need to understand ozone depletion, and later maintained and encouraged through the Montreal Protocol and subsequent amendments, alongside associated activities (SPARC/IO3C/GAW, 2018; WMO, 2018).

In this section a broad description of stratospheric ozone is presented: the production, destruction and transport of ozone and how that influences the distribution and variability of ozone; the anthropogenic influence on ozone in the form of ozone depleting substances and how this led to increased ozone depletion; and how ozone in the stratosphere is modelled and observed.

1.1.1 Ozone photochemistry

Stratospheric ozone is primarily formed by the photochemical reaction

$$O_2 + h\nu(\lambda < 243\,\text{nm}) \rightarrow 2O \tag{1.1}$$

followed by a subsequent reaction with a collision partner M, most likely either N_2 or O_2 due to their large abundances.

$$O + O_2 + M \rightarrow O_3 + M. \tag{1.2}$$

Ozone removal also happens photochemically

$$O_3 + h\nu(\lambda < 320\,\text{nm}) \rightarrow O + O_2 \tag{1.3}$$

$$O + O_3 \rightarrow 2O_2, \tag{1.4}$$

completing the cycle of photochemical production and destruction of ozone known as the Chapman cycle (Chapman, 1930). In the stratosphere, through reactions 1.2 and 1.3, O and O_3 quickly reach photochemical equilibrium.

The absorption of ultraviolet radiation (UV) in reaction 1.3 is the dominant cause of heating in the stratosphere, although the collision in reaction 1.2 is exothermic and also contributes. This heating creates a stratospheric temperature profile that increases with height, therefore determining the dynamical stability of the stratosphere (Murgatroyd and Singleton, 1961). In addition to providing stratospheric structure, the absorption of UV radiation by ozone and oxygen within the stratosphere is fundamental to the survival of life on Earth, protecting the DNA and chemical structures of organisms from UV's harmful effects (WMO, 2018). Without this protective layer against UV-C (100–280 nm) and UV-B (280–315 nm) radiation, organisms are vulnerable to

the carcinogenic effects of UV, photodamage and immunosuppression (Harm, 1980; Schwarz, 2005).

The Chapman cycle has a description of ozone production and loss which is able to partially describe the structure of ozone in the atmosphere. It generates an ozone layer because UV radiation, the source of oxygen photolysis, decreases down through the atmosphere whilst the oxygen density per unit volume increases. However, reaction 1.4 is slow and underestimates the removal of ozone. The Chapman cycle alone is unable to account for observed abundances.

Ozone is additionally destroyed through other catalytic cycles with odd hydrogen ($HO_x = H + OH + HO_2$) (Bates and Nicolet, 1950), nitrogen ($NO_x = NO + NO_2$) (Crutzen, 1970) and chlorine ($ClO_x = Cl + ClO$) radicals (Stolarski and Cicerone, 1974). Although ozone loss is affected by anthropogenic sources of stratospheric HO_x and NO_x (Portmann et al., 2012), for example through emissions of methane and from combustion, the main source of ozone depletion over the last half a century is due to anthropogenic emissions of halogenated compounds (WMO, 2018).

These catalytic reactions result in ozone loss but not the loss of the catalyst. They take the form

$$X + O_3 \rightarrow XO + O_2 \tag{1.5}$$

$$XO + O \rightarrow X + O_2, \tag{1.6}$$

where X is a catalyst, resulting in a net reaction of

$$O + O_3 \rightarrow 2O_2. \tag{1.7}$$

The above reactions require a single O atom and therefore dominate in the upper stratosphere (above 40km) where the abundance of O is highest because of the high intensity of incident solar radiation. Lower in the stratosphere, catalytic cycles that do not depend on the availability of O atoms are particularly important. There are three main types of ozone-specific catalytic cycles. Firstly, a reaction between X and

O_3 such as

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{1.8}$$

$$HO_2 + O_3 \rightarrow OH + 2O_2 \tag{1.9}$$

$$\text{Net reaction: } 2O_3 \rightarrow 3O_2, \tag{1.10}$$

a reaction which dominates below 25 km. Secondly, a cycle which involves two different X species, for example

$$Br + O_3 \rightarrow BrO + O_2 \tag{1.11}$$

$$OH + O_3 \rightarrow HO_2 + O_2 \tag{1.12}$$

$$HO_2 + BrO \rightarrow HOBr + O_2 \tag{1.13}$$

$$HOBr + h\nu \rightarrow OH + Br \tag{1.14}$$

$$\text{Net reaction: } 2O_3 + h\nu \rightarrow 3O_2, \tag{1.15}$$

in which O_2 is eliminated upon formation of the new compound (in this case HOBr). The third type is similar except that the O_2 is eliminated after the formation of a compound from the reaction of two different X species. A key example is the dimer cycle (Molina and Molina, 1987) which is the main cause of ozone destruction within springtime polar vortices and is therefore responsible for the formation of the ozone hole described in subsubsection 1.1.2.1.

$$2(Cl + O_3 \rightarrow ClO + O_2) \tag{1.16}$$

$$2ClO + M \rightarrow Cl_2O_2 + M \tag{1.17}$$

$$Cl_2O_2 + h\nu \rightarrow Cl + ClOO \tag{1.18}$$

$$ClOO + M \rightarrow Cl + O_2 + M \tag{1.19}$$

$$\text{Net reaction: } 2O_3 + h\nu \rightarrow 3O_2. \tag{1.20}$$

The rates of the above catalytic reactions are controlled in part by null cycles that temporarily remove catalysts and prevent their reactions with ozone through the formation of reservoir compounds such as N_2O_5, which removes NO_x during the absence of sunlight. $ClONO_2$, HNO_3, HCl and $BrONO_2$ are other important reservoir

species that lock up halogenated compounds reducing the rate of ozone destruction and are stable enough to be transported throughout the stratosphere.

1.1.2 Anthropogenic influence on stratospheric ozone

The initial concerns about the stability of stratospheric ozone came in response to nuclear weapons and high-altitude supersonic flight (Crutzen, 1971; Muller, 2009). However, over the last half a century the focus has shifted to anthropogenic emissions of halogenated compounds, initially in the form of chlorofluorocarbons (CFCs) such as CFC-12 (CF_2Cl_2), that have been widely used as unreactive refrigerants and propellants (Molina and Rowland, 1974; Rowland, 2009; WMO, 2018). Although relatively inert in the troposphere where they are emitted and accumulate, once transported into the more photochemically active stratosphere these ozone depleting substances (ODSs) are photolysed into highly reactive halogen gases (e.g., Cl, Br, ClO and BrO) which react to destroy ozone via any of the three catalytic cycles detailed in reactions 1.8 onward. As a result, a global decrease in total column ozone has been observed since the 1980s (Solomon, 1999; Rowland, 2009; WMO, 2018). Beyond the efficient removal of ozone by catalytic cycling, ODSs pose a long-lasting problem due to their stratospheric lifetimes of up to 100 years (Owens et al., 1982; Rigby et al., 2013).

1.1.2.1 Antarctic ozone depletion

Antarctic ozone depletion strongly affects the Antarctic climate (Perlwitz et al., 2008), more widely the surface climate in the southern hemisphere (Polvani et al., 2011; "Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change" 2011), and the stratosphere which experiences cooling (Randel and Wu, 1999b). The southern polar region where decline was first identified by Farman et al. (1985) experiences particularly strong depletion in springtime, resulting from a unique set of chemical and meteorological conditions (Solomon et al., 1986). This phenomenon known as the *ozone hole* in which total column ozone values can fall below 100 DU, begins in autumn when the polar vortex forms around Antarctica as stratospheric air cools and descends, isolating the polar air and creating a barrier to the ozone rich mid-latitudes (Schoeberl and Hartmann, 1991). As temperatures decrease, the vortex strengthens and polar stratospheric clouds form.

Polar stratospheric clouds (PSCs) and heterogeneous chemistry

When stratospheric temperatures fall below -78 °C PSCs can form within polar regions, influencing ozone depletion via heterogeneous chemistry (Solomon et al., 1986). The PSCs important in ozone depletion (Type 1) are nitric acid-containing PSCs rather than their rarer and colder Type 2 counterparts formed of ice particles. Reactions that take place on the PSCs prime the polar air for depletion during polar night by converting chlorine containing reservoir species (e.g., HCl, $ClNO_3$ and $ClONO_2$) into reactive forms of chlorine. During winter these reactions make ClO the most abundant chlorine species. When sunlight returns in spring, ClO drives rapid depletion through catalytic cycles (particularly the dimer cycle reactions 1.17–1.20) (Molina and Molina, 1987; Barrett et al., 1988) and creating the ozone hole.

PSCs have a secondary impact on ozone depletion known as the denitrification of the stratosphere (Salawitch et al., 1989). Nitric acid (HNO_3) containing PSCs descend through the stratosphere over winter and spring at a rate of about 1.5 km per day, removing a significant proportion of available HNO_3 from the ozone layer (Crutzen and Arnold, 1986; Toon et al., 1986). As HNO_3 is a source of NO_x, the reaction between NO_x and ClO to form $ClONO_2$ occurs less, resulting in an increased lifetime of ClO and therefore increased ozone destruction.

1.1.3 Controlling ozone depleting substances

As the UV protection afforded by stratospheric ozone is crucial to the survival of most life on Earth, international policy was swiftly enacted in an attempt to reverse depletion (McKenzie et al., 2011). The Montreal Protocol (MP) and its amendments seek to do this through monitoring and restricting the production of halogenated compounds, subsequently banning hydrochlorofluorocarbons and hydrofluorocarbons (the Kigali amendment in 2016) in addition to the initial CFC ban. The MP is widely considered one of the world's most successful environmental agreements (Brack, 2003; Gareau, 2015) limiting ozone depletion (Chipperfield et al., 2015) in addition to mitigating climate change (Velders et al., 2007; Young et al., 2021). However, whether recovery is significant for different atmospheric regions is still an active research question (Chipperfield et al., 2017; Ball et al., 2018; Weber et al., 2018; Ball et al., 2019), one confounded by the relatively short and incomplete observational records,

difficulty in disentangling trends from atmospheric variability (SPARC/IO3C/GAW, 2018; WMO, 2018), and new unregulated emissions of ODSs (Hossaini et al., 2017; Montzka et al., 2018). Additionally, as the impacts of ozone depletion are wide reaching, disrupting tropospheric circulation (Polvani et al., 2011), impacting surface climate (Perlwitz et al., 2008) and changing the Brewer-Dobson circulation (Abalos et al., 2019), continued study and monitoring of stratospheric ozone is of vital importance.

1.1.4 Ozone distribution and variability

Ozone concentrations in the stratosphere are governed by photochemical production, photochemical destruction through catalytic cycles and transport. The dominant large-scale dynamical system in the stratosphere is the Brewer-Dobson circulation (BDC) (Brewer, 1949; Dobson, 1956), caused by high latitude Rossby wave breaking (Dunkerton, 1978). Planetary-scale Rossby waves caused by topographically large features propagate vertically into the stratosphere, where they deposit westward momentum decelerating the eastward stratospheric winds and the wintertime polar jet stream (Andrews and Mcintyre, 1976; Boyd, 1976). To conserve angular momentum a small poleward flow is produced which drives a poleward circulation, in turn driving vertical transport in the tropics moving air from the troposphere to the stratosphere (Butchart, 2014).

The short photochemical lifetime of ozone in the upper stratosphere means that the BDC has less impact on ozone distributions than in the lower stratosphere, where the lifetime is longer (on the order of months; Sankey and Shepherd (2003)) and therefore transport dominates (Shepherd, 2008). Tropospheric air, that is more ozone deficient than the stratospheric air, is injected into the stratosphere by the BDC resulting in reduced tropical stratospheric ozone. This transport from troposphere to stratosphere also injects into the upper atmosphere water vapour and reactive gases, such as those species required for the catalytic chemistry in subsection 1.1.1. The tropical air, which receives the highest intensity of sunlight and therefore has the highest ozone production, is then transported by the BDC poleward. As it does so the now ozone-rich air descends over extratropical regions into the lower stratosphere, and as the ozone lifetime here is long, ozone accumulates (Weber et al., 2011).

Total ozone column is the measure of all ozone from surface to atmospheric

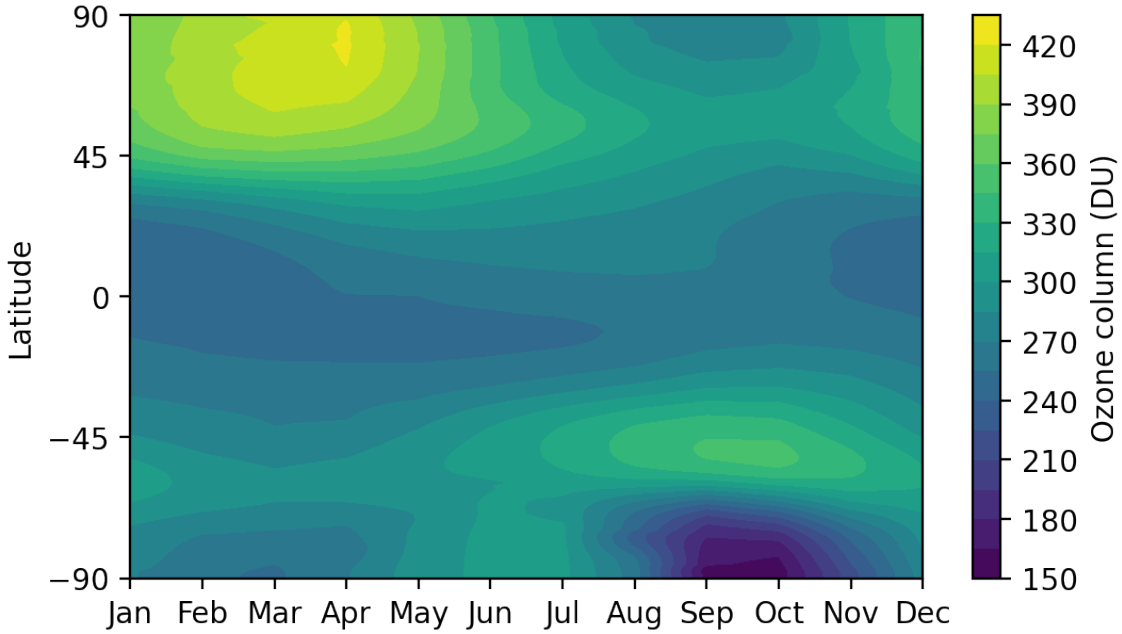

Figure 1.1: The mean total column ozone as a function of latitude and season for the years 1979–2016. Total column data is from the Bodeker Scientific patched (infilled) total column ozone dataset (Bodeker and Kremser, 2021).

boundary and is observed by both satellites (e.g., the Total Ozone Mapping Spectrometer (TOMS); Heath et al., 1975) and ground based instruments (e.g., Dobson spectrophotometers; Komhyr et al., 1989). A climatology of total ozone column is shown in Figure 1.1 calculated from the assimilated and infilled observational NIWA-BS dataset (Bodeker et al., 2021). Across the tropics for all seasons there is a clear minima caused by the upwelling of ozone deficient air into the stratosphere driven by the BDC. A local minimum centred over the southern pole in September/October, is the Antarctic ozone hole (discussed in subsubsection 1.1.2.1), a region of large ozone depletion driven by its unique combination of chemistry and dynamics. Springtime ozone depletion is also a feature over the northern pole, but it lacks the regularity and severity of its southern counterpart (Solomon et al., 2007).

Measuring ozone as a total column predominantly captures stratospheric ozone as this is where the majority (90 %) of atmospheric ozone exists. However, this measure provides no information about the vertical distribution of ozone, the importance of

which has already been highlighted by the dependence of ozone production and removal on altitude. Figure 1.2 shows the variation of ozone in latitude and pressure and highlights the high-density layer of ozone which resides at altitudes between 20 and 25 km, a region which contains much of the stratospheric ozone. This ozone layer is higher in the tropics due to the increased height of the tropopause and the transport of ozone deficient air into the lower stratosphere by the BDC and tropical convection. Even in its highest concentrations, ozone is still a trace gas reaching a maximum density of about 10 ppm.

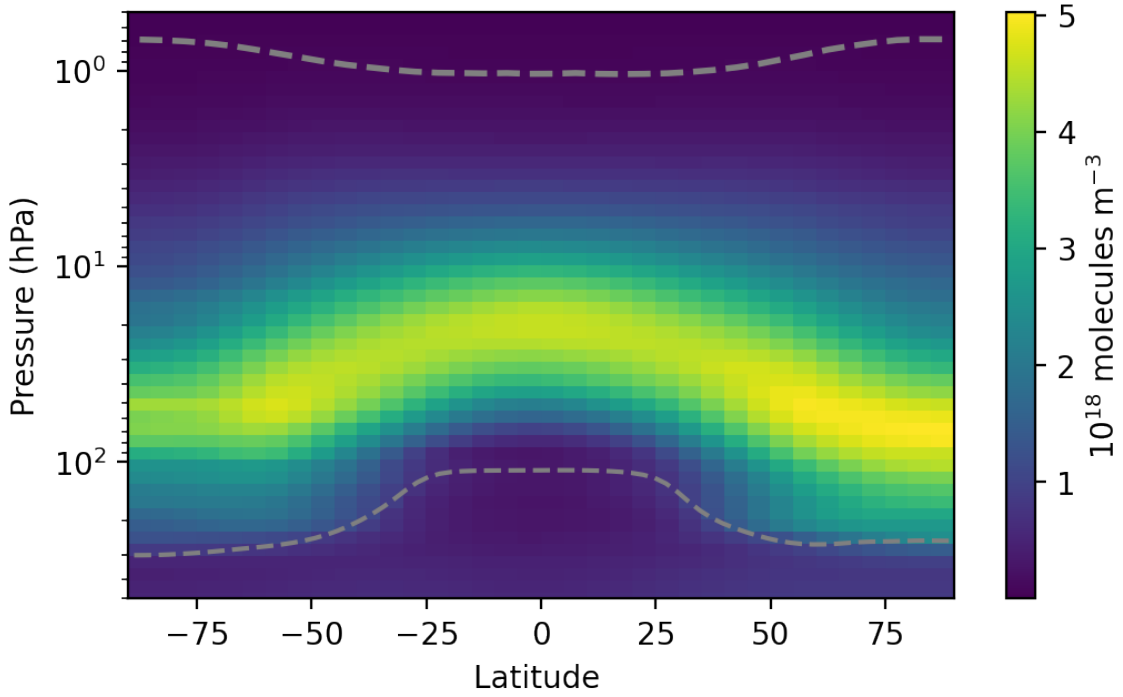

Figure 1.2: Average ozone density as a function of pressure and latitude for the period 1980–2016. Ozone data is from the Bodeker Scientific (Tier1.4) patched (infilled) vertically resolved ozone dataset detailed by Bodeker et al. (2013). Dashed grey lines show approximate locations of the tropopause (bottom) and stratopause (top) derived from NCEP reanalysis (Kalnay et al., 1996).

The distribution of ozone varies on several naturally driven timescales. The solar cycle ($T = 11$ years) partially controls photochemical ozone production through the modulation of incident UV radiation, displaying a 2–4 % variation in the solar ozone

response in the upper stratosphere and about a 2 % variation in the lower stratosphere (Dhomse et al., 2016; WMO, 2018). Both chemical and dynamical processes concerning ozone are affected by the quasi-biennial oscillation (QBO), a major source of tropical low frequency variability ($T = 28$ months). The circulatory changes produced by the QBO induce meridional flow, modulating tropical upwelling, transport between the tropics and extratropics and therefore abundances of ozone and other reactive gases (Baldwin et al., 2001). Similarly, the El Niño Southern Oscillation (ENSO) affects lower stratospheric tropical ozone abundance through its modulation of tropical upwelling (Bodeker et al., 1998; Randel et al., 2009), although partial correlation between ENSO and QBO confound the dissection of their individual impacts on ozone (Oman et al., 2013). Identifying the dependence of ozone on natural oscillations is important for accurately quantifying ozone trends which requires the removal of natural variability (e.g., Ball et al., 2018; Braesicke et al., 2018; SPARC/IO3C/GAW, 2018).

1.1.5 Ozone observation and modelling

To fully understand these impacts and monitor recovery and adherence to the MP, accurate and widespread observational datasets are required, in addition to a suite of numerical models capable of simulating stratospheric chemistry and dynamics (WMO, 2018). CCMs are widely used to project the evolution of stratospheric ozone (Dhomse et al., 2018) and to investigate the response of the system to different forcing and emissions (WMO, 2018). These are typically climate models that additionally simulate chemistry, including gas-phase and heterogeneous chemistry, dependence of reaction rates on temperature and aerosols, to measure the impact that changing chemical concentrations has on the atmospheric system (Morgenstern et al., 2010; Weber et al., 2021). Regular coordinated model intercomparisons are organised to compare and validate CCMs (Eyring et al., 2008; Eyring et al., 2010; Morgenstern et al., 2017) and to produce ensemble projections (Young et al., 2013; Dhomse et al., 2018). The ability of CCMs to recreate and simulate ozone depletion and the ozone hole is well documented (e.g., Struthers et al., 2009; CCMVal, 2010; Morgenstern et al., 2018).

Throughout this thesis, model runs from the Chemistry-Climate Model Initiative (CCMI) (Morgenstern et al., 2017) are used. This ensemble has a central focus on ozone modelling (Eyring et al., 2013) making it extremely suitable for considering future

ozone projections. Most of the CCMs within CCMI (a couple are chemical transport models which use offline meteorology) explicitly represent tropospheric chemistry in addition to stratospheric chemistry, although they vary in their treatment and grouping of halogen species relevant for simulating ozone depletion. CCMI simulations will be used in chapter 2 for creating weighted ozone depletion projections, and in chapters 3, 4 and 5 for improving historic predictions of ozone.

Complementary to modelling efforts are observational products, comprising of measurements recorded by ground-based instruments (e.g., Fioletov et al., 2008), satellites (e.g., Tegtmeier et al., 2013), ozonesondes (e.g., Witte et al., 2017) and other airborne in-situ instruments. These observations aid ozone monitoring efforts and have led to the discovery of illicit ODS production (Montzka et al., 2018; Rigby et al., 2019). Ground-based observations and those from nadir-viewing satellites provide measurements of the total ozone column, whereas in-situ instruments and satellite-borne limb sounders measure a vertically resolved distribution of ozone, both of which are considered in this thesis. As observations from a single instrument rarely have continuous coverage, measurements are often assimilated from multiple observational sources (e.g., Miller et al., 2002; Davis et al., 2016; Ball et al., 2017) providing long-term partially continuous records that can be used for trend analysis (Ball et al., 2018; SPARC/IO3C/GAW, 2018) or as inputs to offline models (Cionni et al., 2011). However, the methods used to infill gaps in observational ozone records are relatively simple statistical models, lacking physical and chemical understanding of the system resulting in low accuracy where data is sparse (Davis et al., 2016). Additionally, a complete estimate of statistical and observational uncertainty is rarely considered. Producing continuous ozone datasets from sparse observations assimilated with model simulations forms the basis of chapters 3 and 5 in which a new Bayesian methodology is designed to fuse together models and observations.

1.2 Why apply ML and data science to atmospheric science?

ML is an area of study in which computers learn without explicitly being programmed. Traditional numerical weather prediction predicts tomorrow's weather by knowing the

previous days' weather and the physical laws that govern the meteorological system. Numerical weather prediction was first suggested by Richardson (1922) who imagined grids of human computers (about 64000) simultaneously solving the governing laws of the dynamical atmospheric system, laying the foundations of atmospheric modelling. In contrast, an ML approach would take the weather from previous days and use these data to learn the rules and patterns of the evolving weather system, which in turn can be used to make future weather predictions (Haupt et al., 2018). ML is used extensively throughout industry and scientific disciplines, excelling at tasks such as identifying patterns (Anzai, 2012), learning complex relationships amongst big data (Ratner, 2017) and making accurate predictions (e.g., Kourou et al., 2015; Patel et al., 2015), all without requiring direct human input.

One particular strength of ML is finding patterns and relationships amongst big data, for which manual analysis would be time consuming or even unfeasible. Atmospheric science is awash with large observational datasets, such as the Tropospheric Ozone Assessment Report database (Schultz et al., 2017) containing over 10 billion entries, and relies heavily on large output from numerical model simulations of the earth system, the largest being on the order of 10 PB in size (Eyring et al., 2016b). Not only are datasets continually increasing in size they are also increasing in scope, ensemble size and model diversity, regularly measuring or simulating highly nonlinear systems of a complexity inscrutable by traditional analysis (Faghmous and Kumar, 2014). For analysing, classifying and predicting large, nonlinear, complex and uncertain data common in atmospheric science, ML approaches are highly applicable. As a result of increased access to sophisticated computing and open access ML software, coupled with a drive to improve on traditional methods, the use of ML in environmental science has exponentially increased in recent years with 6 % of publications within the American Geophysical Union containing the term "ML" in 2020 compared to 0.5 % in 2010 and 0.1 % in 2000.

In atmospheric science ML has been used to create and improve forecasts of atmospheric rivers (Chapman et al., 2019) and severe weather (McGovern et al., 2014), classify extreme weather events (Grazzini et al., 2020) and aerosols (Christopoulos et al., 2018), and identify signals of climate change (Barnes et al., 2019). Recent progress in causal discovery has led to new methods to evaluate climate models (Nowack

et al., 2020) and understand teleconnections in the Earth system (e.g., Kretschmer et al., 2017). Neural networks are particularly popular ML tools in environmental data science due to their ability to model complex nonlinear systems at a computational cost that scales proportionally with the size of the data (Gurney, 2018). They are however typically black box models (McGovern et al., 2019), which has prompted active research into explainable artificial intelligence (Samek et al., 2019) and ML (e.g., Labe and Barnes, 2021; Wang et al., 2021). Neural networks have been readily applied in atmospheric science, including helping quantify causes of discrepancies in hydroxyl radical predictions in CCMs (Nicely et al., 2020), weather forecasting (Fente and Singh, 2018) and downscaling precipitation (Groenke et al., 2020).

Hybrid models created by coupling ML with physical models have found application in the bias correction of models (e.g., Dhomse et al., 2021) and the replacement of model components and parameterisations with ML models to refine and speed up simulations (e.g., O'Gorman and Dwyer, 2018; Keller and Evans, 2019). Hybrid approaches also have the potential to create models that rely on numerical physical models for their predictive abilities and knowledge of the physical system, whilst leveraging ML to apply corrections and fill in knowledge gaps (Pathak et al., 2018). Beyond hybrid models, recent research is leading the way towards totally data-driven environmental models, which rather than relying on explicit programming of a system such as in climate models, instead learn and model system behaviour from data (Schneider et al., 2017). Aside from possible benefits of increased model speed, data-driven methods can also learn to simulate not yet understood behaviours in a way that process models cannot. Environmental data-driven models are already being used to model complex and chaotic dynamical oceanic and atmospheric systems with physics informed neural networks (Chattopadhyay et al., 2020; deWolff et al., 2021).

1.3 Model ensembles

One section of atmospheric science where ML and data science approaches are particularly applicable is in the analysis and assimilation of model ensembles, as the data is complex, large, and contains many patterns within highly nonlinear systems. Ensembles of climate and Earth system models are fundamental for international

climate change assessments (Knutti et al., 2010; IPCC, 2013b; Hoegh-Guldberg et al., 2018) and ensembles of CCMs are similarly important in determining the future of atmospheric composition (Dhomse et al., 2018; WMO, 2018). An ensemble is a set of comparable simulations from multiple models (e.g., Taylor et al., 2012; Lamarque et al., 2013a; Eyring et al., 2016b). These can be used to improve the accuracy of projections compared to a single model (Gleckler et al., 2008), and to quantify uncertainty (Reichler and Kim, 2008; Annan and Hargreaves, 2011b), estimated by the variance of the simulations across the ensemble (IPCC, 2010). However, the statistical assumptions made within the construction and analysis of these ensembles can lack statistical rigour, resulting in the ensemble being an *ensemble of opportunity* (Tebaldi and Knutti, 2007), where modelling groups able to take part submit as many simulations as they can, rather than a systematic sample of model uncertainty, where an ensemble would be designed to fully sample the structural, parameter and initial condition uncertainties (IPCC, 2010). How results from model ensembles should be interpreted and to what extent the ensemble prediction is probabilistic is complex and still an area of much discussion (Tebaldi and Knutti, 2007; Knutti, 2010; Curry and Webster, 2011; Knutti et al., 2017; Herger et al., 2018).

In the analysis and assimilation of ensembles there are several confounding factors. Firstly, models are not created equally, some are better at capturing trends or recreating past atmospheric states (Gleckler et al., 2008; Reichler and Kim, 2008; Knutti et al., 2013), and individual models themselves might be more successful in their simulations of certain regions, time periods or particular atmospheric conditions (Eyring et al., 2006; Baker and Taylor, 2016). Secondly, models are not independent, duplicating design ideas, code and even entire model components (Masson and Knutti, 2011; Abramowitz et al., 2019). These features of differing model performance and similarity are shown in Figure 1.3 which shows modelled total ozone column from multiple CCMs from the CCMI (Morgenstern et al., 2017) ensemble alongside the output from four ensemble assimilation methods: a multi-model mean, a multi-model mean within 1 standard deviation, a performance weighted mean, and a weighted mean that accounts for performance and similarity.

The multi-model mean, or 'one model one vote', is a traditional and widely used approach that combines model projections in an ensemble by averaging them (e.g.,

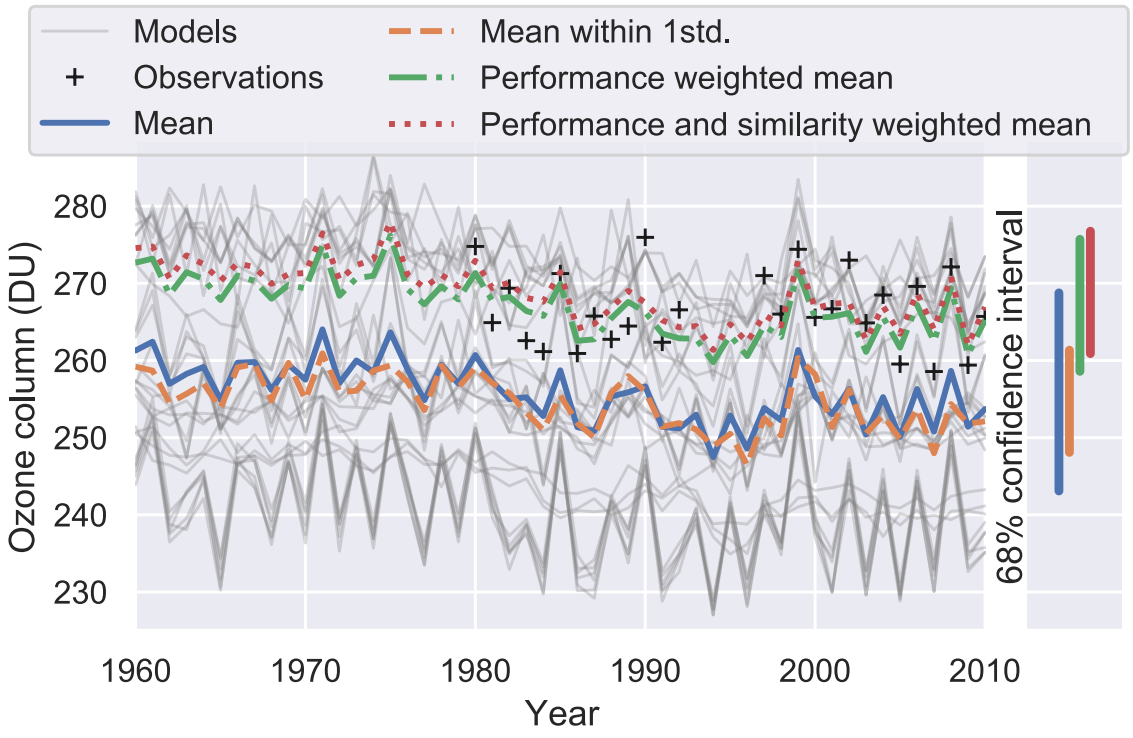

Figure 1.3: Annual total ozone column at 0°latitude and 0°longitude. Historic modelled total ozone column is shown (in light grey) for 14 models, totalling 28 individual simulations, from the Chemistry-Climate Modelling Initiative (Morgenstern et al., 2017). Observations from the Bodeker Scientific total ozone column record are shown as black crosses (Bodeker et al., 2018). Coloured lines show four simple methods for assimilating multiple model outputs into a single ensemble output. The mean (blue line) is a simple arithmetic average across all the model simulations, commonly referred to as the multi-model mean. The orange dashed line shows the same mean but excluding models that are more than 1 standard deviation away from the ensemble mean. A weighted mean, where model weights are calculated as $w_i = 1/(\sum(\text{obs} - \text{model})^2)$ is shown in green. A weighting strategy much like one implemented by Knutti et al. (2017) that accounts for model similarity and model performance is shown in red. The right hand panel shows the average 68 % confidence interval of the assimilated prediction for each of the four methods.

IPCC, 2013b; Hoegh-Guldberg et al., 2018). However, as described above, this method produces a biased projection if the ensemble contains unrealistic or badly performing models and relies on model simulations being symmetrically distributed around the truth, otherwise the mean ensemble output will be biased (Annan and Hargreaves, 2011a). It is further skewed by the inclusion of multiple simulations from a single model, giving that model unequal representation.

A more constrained multi-model-mean only considers simulations that fall within 1 standard deviation of the ensemble mean (e.g., Dhomse et al., 2018), thus removing any extreme outliers. This approach however, similarly fails to account for variable model similarity and performance in models that are not outliers but will still produce biased and potentially unrealistic projections (Knutti, 2010). Other approaches weight models dependent on their success at replicating the past, assuming that model performance is transitive from past to future, and account for model similarity (e.g., Waugh and Eyring, 2008; Knutti et al., 2017; Brunner et al., 2019, see also Chapter 2). Models that successfully recreate observations get increased weighting, whilst models that are highly similar to others, e.g., they have submitted multiple simulations from the same model or two models have the same underlying general circulation model, are down-weighted. The figure shows two weighting methods, one where models are weighted only on their ability to reproduce observations and one where they are also weighted on how similar they are. Although these approaches account for variable model performance within the ensemble, they fail to account for spatially and temporally varying performance for individual models, thus missing out on an opportunity to greatly improve the accuracy of the ensemble projection.

In Figure 1.3 observed ozone values are at the top of the range of the model simulations, indicating individual model biases of up to 30 DU and highlighting the problem of variable model performance. The issue of model similarity can also be seen in the bottom most model which has 5 ensemble members which are near identical, therefore unfairly biasing the ensembled output towards this model. It is clear that analysis and assimilation of this ensemble will be confounded by unequal model performance and varying levels of model dependence.

Which method is chosen to analyse model ensembles has a large impact on the ensemble output. This is seen in Figure 1.3 where the standard mean methods project

an ozone column approximately 10 DU less compared to the weighted means. Different methods also affect the uncertainty of the projection, calculated by a weighted or unweighted measure of model spread, shown by the average confidence interval of the projection on the right hand side of the figure. These discrepancies motivate a need for rigorous ensembling[1] methods that account for model performance and similarity and to develop new ones that increase the utility of under-analysed model ensembles. A central theme of this thesis is to develop new ensembling methods drawing on advances from ML to aid analysis, boost predictive capabilities and better estimate uncertainty.

1.4 Thesis contributions

This thesis seeks to address methodological shortcomings in the combination of CCM model ensembles with observations. Current methods rely on inappropriate, limited or overly simple methods, such as multi-model means and linear regression, to produce important projections and predictions of stratospheric ozone. By merging existing chemistry-climate modelling and observations with advances in ML and data science, we can improve upon the current standard to produce more accurate, robust and uncertainty aware projections of ozone recovery and predictions of historic ozone. Contributions detailed within this thesis are both methodological and application focussed, and are motivated by both a broader desire to better use pre-existing model ensembles across environmental science disciplines and to improve our understanding of modelled and observed stratospheric ozone.

Chapters 2 and 3 have appeared as peer reviewed and published papers, Chapter 4 is unpublished work expanding on the publication in Chapter 3 and Chapter 5 is a paper which is in the submission process. As a result, each chapter contains detailed introduction and motivation. The content and contribution of the chapters are as follows.

Chapter 2 develops a model weighting strategy which accounts for model performance and similarity across a suite of metrics. Using multiple sources of

[1]Throughout this thesis the term *ensembling* will be used as a verb to mean the assembling or combining of multiple model outputs within an ensemble to form a new estimate or prediction. This is a very similar definition to one used in the statistical community to describe the combination of statistical models to generate an improved prediction.

observations and CCM output from CCMI, weighted projections of Antarctic ozone recovery are constructed.

Chapter 3 presents the development of a novel ML Bayesian neural network framework to better combine ensembles of geophysical models by weighting them spatiotemporally. The framework was tested against synthetic data and applied to the problem of infilling sparse historic total ozone column data using the CCMI ensemble, generating a continuous record of total ozone column 1980–2010 with principled uncertainties.

Chapter 4 contains an exploration into the explainability and interpretability of the Bayesian neural network. This chapter discusses and investigates the quality of information that can be elucidated from the Bayesian neural network about the ensembled CCMs' performance and similarity, and the uncertainty of the total ozone column observations.

Chapter 5 presents a continuous vertically resolved ozone dataset, produced using a further developed version of the Bayesian neural network. The dataset is compared with existing infilled datasets to confirm that it captures historic ozone depletion and ozone trends.

Chapter 6 concludes the thesis with a summary of all chapters including discussion on the contributions to ensembles of atmospheric models, the study of stratospheric ozone, and more widely environmental science as a whole. Limitations of this work are also discussed alongside future work.

Chapter 2

Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence

Matt Amos[1], Paul J. Young[1,2], J. Scott Hosking [3], Jean-François Lamarque[4], N. Luke Abraham[5,6], Hideharu Akiyoshi[7], Alexander T. Archibald[5,6], Slimane Bekki[8], Makoto Deushi[9], Patrick Jöckel[10], Douglas Kinnison[4], Ole Kirner[11], Markus Kunze[12], Marion Marchand[8], David A. Plummer[13], David Saint-Martin[14], Kengo Sudo[15,16], Simone Tilmes[4] and Yousuke Yamashita[7].

[1]Lancaster Environment Centre, Lancaster University, Lancaster, UK
[2]Centre of Excellence in Environmental Data Science, Lancaster University, Lancaster, UK
[3]British Antarctic Survey, Cambridge, UK
[4]National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA
[5]Department of Chemistry, University of Cambridge, Cambridge, UK
[6]National Centre for Atmospheric Science (NCAS), Leeds, LS2 9PH, UK
[7]National Institute for Environmental Studies (NIES), Tsukuba, Japan
[8]LATMOS, Institut Pierre Simon Laplace (IPSL), Paris, France

[9]Meteorological Research Institute (MRI), Tsukuba, Japan
[10]Institut für Physik der Atmosphäre, Deutsches Zentrum für Luft- und Raumfahrt (DLR), Oberpfaffenhofen, Germany
[11]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany
[12]Institut für Meteorologie, Freie Universität Berlin, Berlin, Germany
[13]Environment and Climate Change Canada, Montreal, Canada
[14]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[15]Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan
[16]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

Correspondence:
Matt Amos (matt.r.amos@outlook.com, m.amos1@lancaster.ac.uk)

The following work was published in Atmospheric Chemistry and Physics on 26$^{\text{th}}$ August 2020 (citation: Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence, Atmos. Chem. Phys., 20, 9961-9977, doi:10.5194/acp-20-9961-2020, 2020). The authors contributions are listed below.

Statement of contribution

Matt Amos developed the methods and led the analysis, and conceived the study alongside **J. Scott Hosking** and **Paul J. Young**, who made major contributions as the work progressed. **Matt Amos** drafted the manuscript, with the guidance of **Paul J. Young** and **J. Scott Hosking** and input from **Jean-François Lamarque**. **Jean-François Lamarque** and all the other co-authors provided model simulation data. All the co-authors provided model output and helped with finalising the manuscript.

Abstract

Calculating a multi model mean, a commonly used method for ensemble averaging, assumes model independence and equal model skill. Sharing of model components amongst families of models and research centres, conflated by growing ensemble size, means model independence cannot be assumed and is hard to quantify. We present a methodology to produce a weighted model ensemble projection, accounting for model performance and model independence. Model weights are calculated by comparing model hindcasts to a selection of metrics chosen for their physical relevance to the process or phenomena of interest. This weighting methodology is applied to the Chemistry-Climate Model Initiative (CCMI) ensemble, to investigate Antarctic ozone depletion and subsequent recovery. The weighted mean projects an ozone recovery to 1980 levels, by 2056 with a 95 % confidence interval (2052–2060), 4 years earlier than the most recent study. Perfect model testing and out-of-sample testing validate the results and show a greater projective skill than a standard multi model mean. Interestingly, the construction of a weighted mean also provides insight into model performance and dependence between the models. This weighting methodology is robust to both model and metric choices and therefore has potential applications throughout the climate and chemistry-climate modelling communities.

2.1 Introduction

Global chemistry-climate models (CCMs) are the most comprehensive tools to investigate how the global composition of the atmosphere develops, both naturally and under anthropogenic influence (Flato et al., 2014; Morgenstern et al., 2017; Young et al., 2018). As with projecting climate change, consensus views of the past and potential future evolution of atmospheric composition are obtained from coordinated CCM experiments (Eyring et al., 2008; Lamarque et al., 2013b; Morgenstern et al., 2017) and subsequent analysis of the ensemble of simulations (Iglesias-Suarez et al., 2016; Dhomse et al., 2018). Although not a complete sample of structural and epistemic uncertainty, these ensembles are an important part of exploring and quantifying drivers of past and future change, and evaluating the success of policy interventions, such as stratospheric ozone recovery resulting from the Montreal Protocol and its amendments (Dhomse et al., 2018; WMO, 2018). Typically, analysis of an ensemble investigates the behaviour and characteristics of the multi-model mean and the inter-model variance (Tebaldi and Knutti, 2007; Butchart et al., 2010; IPCC, 2013a), rather than accounting for individual model performance or lack of model independence (Knutti, 2010; Räisänen et al., 2010). Methods to address these shortcomings have been proposed for simulations of the physical climate (e.g., Gillett, 2015; Knutti et al., 2017; Abramowitz et al., 2019), but this topic has received less attention in the atmospheric composition community. Here, we demonstrate a weighting method for the CCM simulation of Antarctic ozone loss and projected recovery, where the weighting accounts for model skill and independence over specified metrics relevant to polar stratospheric ozone. We apply this to the recent Chemistry-Climate model initiative (CCMI) (Morgenstern et al., 2017) ensemble and demonstrate the impact of the weighting on estimated ozone hole recovery dates.

Many years of scientific studies and assessments have tied stratospheric ozone depletion to the anthropogenic emission and subsequent photochemistry of halogen-containing gases, such as chlorofluorcarbons (CFCs), hydrofluorocarbons (HCFCs) and halons (WMO, 2018). This science guided the development of the Montreal Protocol, and its subsequent amendments, to limit and ban the production of these ozone-destroying gases, and stratospheric ozone is now thought to be recovering (Solomon et al., 2016; Chipperfield et al., 2017). Of particular concern is the Antarctic "ozone

hole": a steep decline in high latitude stratospheric ozone during austral spring that can reduce ozone concentrations to near zero at particular altitudes, driven by polar night-time chemistry, cold temperatures and heterogeneous catalysis on polar stratospheric clouds (PSCs) (Solomon, 1999). While the ozone hole continues to appear in each austral spring, it appears to be showing signs of recovery (Langematz et al., 2018). The strong cooling associated with Antarctic ozone depletion (Thompson and Solomon, 2002; Young et al., 2012) has driven circulation changes in the stratosphere and in the troposphere, particularly in austral summer. This has notably included an acceleration and poleward movement of the southern high latitude westerly winds and associated storm tracks (Perlwitz et al., 2008; Son et al., 2008), leading to summertime surface climate changes through many lower latitude regions including the tropics ("Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change" 2011).

The recovery process is slow due to the long atmospheric lifetimes of ozone depleting substances, and could be hampered by releases of ozone depleting substances (ODSs) not controlled by the Montreal Protocol, such as short-lived halogens (Claxton et al., 2019; Hossaini et al., 2019) or nitrous oxide (Portmann et al., 2012; Butler et al., 2016), or instances of non-compliance, such as the recent fugitive emissions of CFC-11 (Montzka et al., 2018; Rigby et al., 2019). Recovery itself is often defined as the date at which the ozone layer returns to its 1980 levels, and this is the benchmark used by the WMO (WMO, 2018) to assess the progress due to the implementation of the Montreal Protocol.

The assessment of when the ozone layer will recover is conducted using an ensemble of chemistry-climate models, forced by past and projected future emissions of ozone depleting substances (ODSs) and climate forcers (Eyring et al., 2010; Dhomse et al., 2018). Such ensembles are used to establish the robustness of the model results for a particular scenario: when several models agree, the prevailing assumption is that we can have greater confidence in the model projections. Yet, there has been much discussion about how true this assumption is (Tebaldi and Knutti, 2007; Sanderson et al., 2015a; Abramowitz et al., 2019). In an ideal scenario, every model within an ensemble would be independent and have some random error. In this case, we would expect that increasing the ensemble size would decrease the ensemble uncertainty and allow us to better constrain the mean value. However, in modern model inter-comparison projects

this is not the case: although often developed independently, models are not truly independent, often sharing components and parametrisations (Knutti et al., 2013); models are not equally good at simulating the atmosphere (Reichler and Kim, 2008; Bellenger et al., 2014); and lastly, models do not have a predictable random error but instead have layers of uncertainty extending from uncertainties in parametrising sub-grid processes (Rybka and Tost, 2014) to structural uncertainties from the design of the model (Tebaldi and Knutti, 2007; Knutti, 2010).

Given these issues, there is currently no consensus on how best to combine model output when analysing an ensemble. Probably the most widely used and simplest is to take a multi model mean where each model contributes equally, and indeed it has also been established that an ensemble mean performs better than any single model (Gleckler et al., 2008; Pincus et al., 2008; Reichler and Kim, 2008; Knutti et al., 2010). A more sophisticated method is to weight individual ensemble members, accounting for model performance as well as the degree of a model's independence. Weighting methods of various forms have been developed and implemented on global physical climate model ensembles (Tebaldi et al., 2005; Räisänen et al., 2010; Haughton et al., 2015; Knutti et al., 2017), but seldom for atmospheric composition. In most cases the weights are calculated from comparison of model hindcasts to observational data, either for a single variable of interest or over a suite of diagnostics. Additionally, reliability ensemble averaging (REA) (Giorgi and Mearns, 2002) is an alternative weighting technique which gives higher weights to those models near the multi model mean. The main motivation for using a weighted mean is to encapsulate model skill and model independence, such that we down-weight models which perform less well and/or are more similar.

Quantifying model skill (or performance) against comparable observations forms an important part of the validation and analysis of multi-model ensembles (Gleckler et al., 2008; Flato et al., 2014; Harrison et al., 2015; Hourdin et al., 2017; Young et al., 2018). Many CCM inter-comparison projects feature validation and assessment through the use of observation-based performance metrics, which may capture model performance for particular atmospheric variables (e.g., temperature, chemical species concentrations, jet position), or be a more derived quantity which gets closer to evaluating the model against the process it is trying to simulate (e.g., ozone trends vs. temperature trends,

chemical species correlations, chemistry-meteorology/transport relationships) (Eyring et al., 2006; Waugh and Eyring, 2008; Christensen et al., 2010; Lee et al., 2015). Performance metrics are chosen based upon expert knowledge of the modelled system to ensure that metrics are highly related to the physical or chemical processes that the models are being evaluated on.

In this study we develop a weighting methodology, originally presented by Sanderson et al. (2017) and Knutti et al. (2017), for CCM ensembles that accounts for model performance and model independence. We apply it to the important issue of estimating Antarctic ozone recovery using several well-established metrics of model performances, where previously only unweighted means have been used. We first describe our weighting framework in section 2, before describing the model and observational data in section 3. Section 4 presents the application of the weighting framework to Antarctic ozone depletion and the corresponding results. Sections 5 and 6 present a summary and our conclusions.

2.2 The model weighting framework

In this study, we develop and exploit a framework to calculate model weights based on recent work in the physical climate science community (Sanderson et al., 2015a,b; Knutti et al., 2017; Sanderson et al., 2017; Lorenz et al., 2018; Brunner et al., 2019). Here, for an ensemble of N models, the weight for model i (w_i) is given by

$$w_i = \exp\left(-\frac{D_i^2}{n_i\sigma_D^2}\right) \Big/ \left(1 + \sum_{j\neq i}^{N} \exp\left(-\frac{S_{ij}^2}{n_i\sigma_S^2}\right)\right). \tag{2.1}$$

The numerator captures the closeness of the model to observations. D_i^2 is the squared difference between a model and the corresponding observation, which is a measure of performance. The denominator captures the closeness of a model to all other models by comparing the squared difference between them (S_{ij}^2). Both σ_D and σ_S are constants which allow tuning of the weighting to preference either independence or performance (see discussion below). Put more simply, a model has a larger weighting if it closely matches observations and is suitably different to the other models in the ensemble. Finally, equation 2.1 differs from similar versions (e.g., Knutti et al., 2017)

through the addition of n_i, which is the size of the data used to create the weighting. This could be the amount of grid points for a spatial field, the number of points in a time series, or just one for a single-valued statistic, and it normalises the data by length allowing for comparison between models and variables with time series of different length and time invariant parameters.

Investigating and evaluating a phenomenon or complex process often relies on identifying multiple metrics since it can only be partially expressed by any single variable. Expert understanding of the physical process is needed to select a set of relevant metrics with which to develop the process-based weighting. Including multiple metrics, provided they are not highly correlated, has the further benefit of giving less weight to models which perform well but do so for the wrong reasons. In this framework, ensuring that these metrics influence the weighting proportionally is done by normalising the model data using a min-max scaling between 0 and 1.

When combining multiple metrics into a weighting, the weight of the i^{th} model can be found from

$$w_i = \left(\sum_{k=1}^{M} \exp\left(-\frac{D_{ik}^2}{n_{ik}\sigma_D^2} \right) \right) / \left(M + \sum_{k=1}^{M} \sum_{j \neq i}^{N} \exp\left(-\frac{S_{ijk}^2}{n_{ik}\sigma_S^2} \right) \right), \tag{2.2}$$

where M is the total number of metrics and k is the index of the metric. Note that the summation is performed separately over the numerator and the denominator. This means that we calculate the performance and independence scores over all the metrics combined before merging the scores to create the final weighting which, as before, is normalised over all the models to sum to 1.

We take the combined weights for each model and apply them to our parameter or process of interest (the evolution of stratospheric ozone here). As with the metrics this parameter needn't be a time series and could be a spatial distribution or a single measure. The weighted projection is therefore $x = \sum_{i=1}^{N} w_i x_i$, where x_i is an individual model projection and w_i is the associated weight.

2.2.1 Choosing sigma values

The two scaling parameters (σ_S, σ_D) represent a length scale over which two models, or a model and observation, are deemed to be in good agreement. For example, a large σ_S would spread weight over a greater number of models as more models would lie within the length scale of σ_S. On the other hand, a small σ_S sets a higher tolerance for measuring similarity. The choice of the sigma values needs to be considered carefully to strike a balance between weighting all models equally, thus returning to a multi model mean, versus weighting just a few selected models. As the same values of sigma apply across all metrics it is necessary for the data to be normalised to the same values, ensuring that metrics impact the weightings equally. Figure 2.1 shows how the weighting function depends on σ_S, σ_D, model performance and model independence.

As noted in Knutti et al. (2017) there is not an objective way of determining optimal sigma values. Our method of selecting appropriate parameter values was to consider a training and a testing set of data, much like a machine learning problem. We determined the values of sigma using the training data, which in this case is the refC1SD simulations, such that the weighted training data gave a good fit to the observations. The testing data (refC2 simulations) allowed us to test the weights and sigma values out of the temporal range of the training data, which avoids performing testing on data that was used to tune the parameters.

2.3 Applying the weighting framework to the Antarctic ozone hole

We demonstrate the applicability of this weighting framework by applying it to the important and well-understood phenomenon of the Antarctic stratospheric 'ozone hole', for which we can use several decades of suitable observations to weight the models. Below, we describe the model and observation data used and the metrics selected, against which we measure model performance and independence.

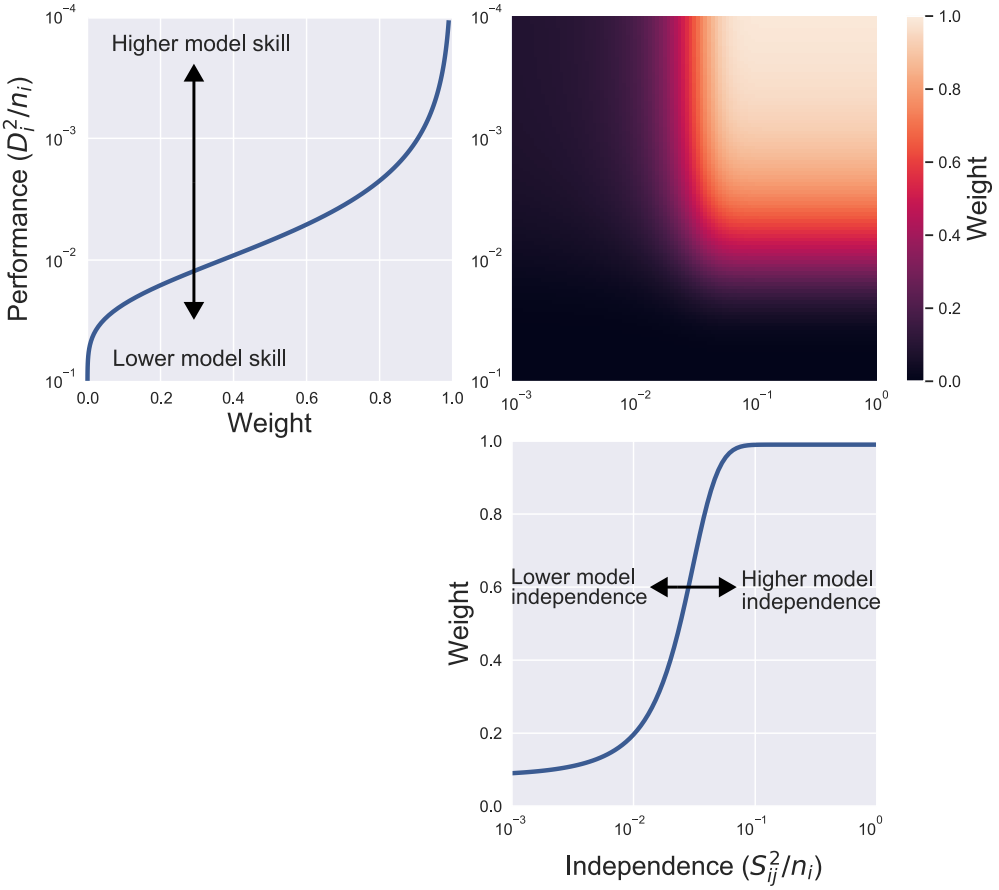

Figure 2.1: Top right shows the overall weighting function w_i (equation 2.1), plotted for 11 models ($N = 11$) with $\sigma_D = 0.1$ and $\sigma_S = 0.1$. Top left shows the contribution to the weighting due to model performance (at $S_{ij}^2/n_i = 1$) and bottom right show the contribution due to model independence (at $D_i^2/n_i = 10^{-4}$). A model which has higher independence and higher skill receives a larger weight. For the weight due to performance (top left) we can see that the weight equals e^{-1} when $D_i^2/n_i = \sigma_D^2$. This shows how σ_D acts as a length scale that determines how close a model has to be to observations to receive weight. σ_S works similarly, setting the length scale that determines similarity.

2.3.1 Model and observation data sources

CCM output was taken from the simulations conducted under Phase 1 of the Chemistry-Climate Model Initiative (CCMI) (Morgenstern et al. (2017) and refs. therein),

which represents an ensemble of 20 state-of-the-art CCMs (where chemistry and atmospheric dynamics are coupled) and chemistry transport models (CTMs, where the dynamics drives the chemistry, but there is no coupling). A detailed description of the participating models is provided by Morgenstern et al. (2017), and here we briefly review their overarching features. Most models feature explicit tropospheric chemistry and have a similar complexity of stratospheric chemistry although there is some variation in the range of halogen source gases modelled. Horizontal resolution of the CCMs ranges from between 1.125°×1.125°to 5.6°×5.6°. Vertically, the atmosphere is simulated from the surface to near the stratopause by all models, and many also resolve higher in the atmosphere. Vertical resolution varies throughout the models, both in the number of levels (34 to 126) and their distribution. All models simulate the stratosphere, although they differ in whether they have been developed with a tropospheric or stratospheric science focus.

We focus on two sets of simulations, called refC1SD and refC2, and for the weighting analysis we only consider models which ran both simulations. Table 2.1 details the exact model simulations used. The refC1SD simulations cover 1980–2010 and represent the specified dynamics hindcast, where the models' meteorological fields are nudged to reanalysis datasets in order that the composition evolves more in line with the observed inter-annual variability of the atmosphere. In addition to being nudged by meteorology the refC1SD runs are forced by realistically varying boundary conditions, including greenhouse gas (GHG) concentrations, ODS emissions, and sea surface temperatures (SSTs) and sea-ice concentrations (SICs). The refC1SD simulations are used to create the model weightings since these are the models' best attempt at replicating the past, giving reasonable confidence that any down-weighting arises due to poorer model performance or strong inter-model similarity. It must be noted that the nudging process is not consistent across the models (Orbe et al., 2018) and we should be mindful that it has the capability to influence the weighting. We discuss the choice to use refC1SD simulations in greater detail in section 2.5.

The refC2 simulations cover 1960–2100 and are used to construct weighted projections of Antarctic ozone recovery, using the weights calculated from refC1SD. The forcing from GHGs and anthropogenic emissions follows the historical scenario conditions prescribed for the fifth coupled model Inter-comparison project (CMIP5)

Table 2.1: The CCMI model simulations used in this analysis and their key references.

Model	refC1SD realisation	refC2 realisation	Reference(s)
CCSRNIES-MIROC3.2	r1i1p1	r1i1p1	Imai et al. (2013) Akiyoshi et al. (2016)
CESM1-CAM4Chem	r1i1p1	r1i1p1	Tilmes et al. (2015)
CESM1-WACCM	r1i1p1	r1i1p1	Marsh et al. (2013) Solomon et al. (2015), Garcia et al. (2017)
CHASER-MIROC-ESM	r1i1p1	r1i1p1	Sudo et al. (2002) (Sudo and Akimoto, 2007) Watanabe et al. (2011) Sekiya and Sudo (2012) Sekiya and Sudo (2014)
CMAM	r1i1p1	r1i1p1	Jonsson et al. (2004) Scinocca et al. (2008)
CNRM-CM5-3	r1i1p2 r2i1p2[a]	r1i1p1	Michou et al. (2011) Voldoire et al. (2013)
EMAC-L47MA	r1i1p1 r1i1p2[a]	r1i1p1	Jöckel et al. (2010) Jöckel et al. (2016)
EMAC-L90MA	r1i1p1 r1i1p2[a]	r1i1p1	
IPSL	r1i1p1	r1i1p1	Marchand et al. (2012) Szopa et al. (2013), Dufresne et al. (2013)
MRI-ESM1r1	r1i1p1	r1i1p1	Deushi and Shibata (2011) Yukimoto (2011), Yukimoto et al. (2012)
UMUKCA-UCAM	r1i1p1	r1i1p1	Morgenstern et al. (2009) Bednarz et al. (2016)

[a] Represents the simulations used in the similarity analysis, but that did not form part of the model weighting.

Table 2.2: The observational products and respective variables used to construct metrics on which to weight the models.

Product	Variable	Metric/s	Citation
MSU	Lower stratosphere temperature (TLS)	TLS TLS gradient Ozone-temperature	Mears and Wentz (2009)
NIWA-BS	Total column ozone V3.4 (TCO)	TCO gradient Ozone-temperature	Bodeker et al. (2018)
GOZCARDS	Hydrogen chloride concentration	Antarctic hydrogen chloride concentration	Froidevaux et al. (2015)
ERA-Interim	Eastward wind speed	Polar vortex breakdown trend	Berrisford et al. (2011)

(Lamarque et al., 2010) up to the year 2000, and subsequently follows representative concentration pathway (RCP) 6.0 for GHGs and tropospheric pollutant emissions (Vuuren et al., 2011); the ODS emissions follow the World Meteorological Organisation (WMO) A1 halogen scenario (WMO, 2011). From CCMI this is the only scenario which estimates the future climate change and developments to stratospheric ozone.

Model performance was evaluated against a series of well-accepted metrics (see below), drawing from widely used observational and reanalysis datasets listed in Table 2.2. Assessing models and ensembles using observational data is a principal way of validating models (Eyring et al., 2006; Waugh and Eyring, 2008; Dhomse et al., 2018) and this is the methodology we follow, with the addition that we create the weights based upon this skill, alongside model independence.

Like many ozone recovery studies, we utilise TSAM (time series additive modelling) (Scinocca et al., 2010) to quantify projection confidence, which produces smooth estimates of the ozone trend whilst extracting information about the inter-annual variability. Here, the TSAM procedure involves finding individual model trends for the refC2 simulations by removing the inter-annual variability using a generalised additive model. Each model trend is then normalised to its own 1980 value. The weighted mean (WM) is created by summing model weights with individual model trends. Two uncertainty intervals are created: a 95 % confidence interval, where there is a 95 % chance that the WM lies within; and a 95 % prediction interval, which captures the

uncertainty of the WM and the inter-annual variability.

2.3.2 Metric choices - How best to capture ozone depletion

The first step in the weighting process is to identify the most relevant processes that affect Antarctic ozone depletion to allow for appropriate metric choice. Suitable metrics require adequate observational coverage and for the models to have outputted the corresponding variables. The metrics we chose are as follows:

Total ozone column gradient. This is the first derivative with respect to time of the total ozone column. Given the discontinuity in the total ozone column record, the years 1992–1996 are excluded. It is a southern polar cap (60°S–90°S) average over austral spring (October and November). September is not included due to discontinuous coverage in the observations.

Lower stratosphere temperature. The lower stratosphere temperature for all of the models are constructed using the MSU TLS-weighting function (Mears and Wentz, 2009). The MSU dataset extends to 82.5°S, and therefore the southern polar cap average ranges from 60°S to 82.5°S and is temporally averaged over austral spring (Sept, Oct, Nov).

Lower stratosphere temperature gradient. This is the first derivative with respect to time of the lower stratospheric temperature found above.

Breakdown of the polar vortex. The vortex breakdown date is calculated as when the zonal mean wind at 60°S and 20 hPa transitions from eastward to westward as per Waugh and Eyring, 2008. We find the trend of the breakdown date between the years 1980–2010 and the gradient of the trend forms the polar vortex breakdown metric.

Ozone-temperature gradient. Both the lower stratosphere temperature and the total ozone column are separately averaged over 60°S to 82.5°S and the October and November mean was taken. We determined a linear relationship between temperature and total ozone column and the gradient of this linear relationship forms the ozone-temperature metric (Young et al., 2013).

Ozone trend-temperature trend gradient. This is similar to the metric above except that we first calculated the time derivative of the total ozone column and temperature polar time series before calculating the linear relationship. The gradient of the linear relationship is the total ozone column trend temperature trend gradient

metric.

Hydrogen chloride. The hydrogen chloride concentration was averaged over the austral spring months and over the Southern Polar cap, for areas which have observational coverage. We consider a pressure range of 316 hPa to 15 hPa to capture the concentration in the lower stratosphere.

These metrics capture two of the main features of ozone depletion, namely: 1) the decrease in temperature over the poles caused by the depletion of ozone, and 2) the breakdown of the vortex which has a major role of isolating the ozone depleted air mass. The chlorine metric encapsulates the anthropogenic release of ODSs and the main chemical driver of ozone depletion. Ozone-temperature metrics allow us to look at model success in reproducing the temperature dependency in ozone reaction rates and stratospheric structure. By looking at the instantaneous rate of change as well as the overall trends, we can gather a picture of both short-term and long-term changes for a range of chemical and dynamical processes.

The metrics are not highly correlated, except for the total ozone column gradient and the lower stratosphere temperature gradient, which are correlated because of the strong coupling of ozone and temperature in the stratosphere (e.g., Thompson and Solomon, 2008). Although this could be cause to discard one of the metrics, to avoid potential double counting, we retain and use both to weight because the models may not necessarily demonstrate this coupling that we see in observations. By considering this variety of metrics, the approach aims to demonstrate that models do not just get the 'right' output, but that they do so for the right reasons.

2.3.3 Evaluating the weighting framework

Two types of testing were used to investigate the usefulness of the weighted prediction and to validate metric choices. Firstly, we performed a simple out-of-sample test on the weighted prediction against the total ozone column observations from NIWA-BS. Although the weights are generated from comparison between the specified dynamics runs (refC1SD) and observations, it does not necessarily follow that the weighted projection created using the free running (refC2) runs will be a good fit for the observations. To test this, we compared the refC2 multi model mean and weighted projection to the observations. Due to the large inter-annual variability in the total

column ozone (TCO) observations, we do not expect the weighted average to be a perfect match; after all, free running models are not designed to replicate the past. However, we need to test the level of agreement between the weighted mean and the observations for an out-of-sample period (2010–2016). This serves a secondary purpose of determining transitivity between the two model scenarios used: i.e., that the weightings found from refC1SD apply to refC2.

Secondly, we used a perfect model test (also known as model-as-truth or a pseudo model test) to determine whether our weighting methodology is producing valid and robust projections. In turn, each model is taken as the pseudo truth and weightings are found in the same way as described in section 2.2 except the pseudo truth is used in place of observations. From these weightings we can examine the skill with which the weighted mean compares to the pseudo truth. We are normally limited to a single suite of observations, but a perfect model test allows us to test our methodology numerous times using different pseudo truths, demonstrating robustness.

Perfect model testing also allows us to test transitivity between scenarios since, unlike with the obvious temporal limit on observations, the pseudo truth exists in both the hindcast and forecast. If a weighting strategy produces weighted means which are closer to the pseudo truth than a multi model mean, then we can have some confidence that we can apply a weighting across model scenarios. Herger et al. (2019) compare the perfect model test to the cross validation employed in statistics, but note that although necessary, perfect model tests are not sufficient to fully show out-of-sample skill which in this case is scenario transitivity. It should be backed up by out-of-sample testing as described above.

2.4 Applying the weighting framework to Antarctic ozone simulations

2.4.1 Antarctic ozone and recovery dates

Figure 2.2 shows the October weighted mean (WM) total column ozone (TCO) trend from the refC2 simulations for the Antarctic (60–90°S). The weights are calculated using equation 2.2, and are based on both model performance and independence. All

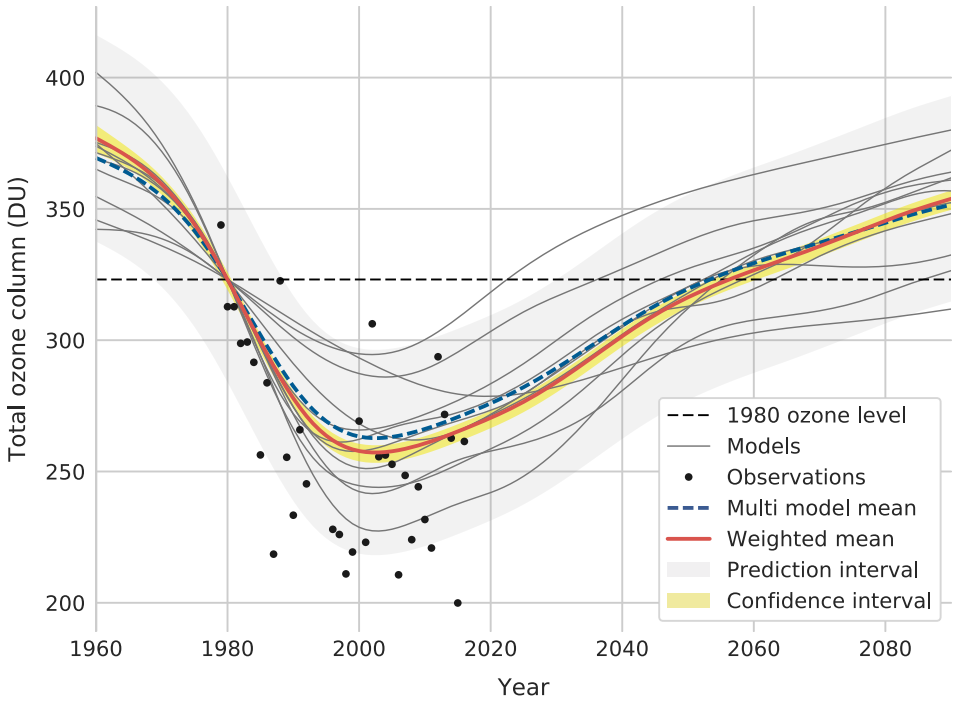

Figure 2.2: Antarctic (60–90°S) October TCO. The weighted mean (refC2 simulations weighted upon refC1SD performance and independence) is shown in red, the multi model mean (refC2 simulations) is shown in blue, and individual refC2 model trends are shown in grey. The NIWA-BS observations are shown in black. All model projections and ensemble projections are normalised to the observational 1979–1981 mean shown as the black dashed line. 95 % confidence and prediction intervals for the weighted mean are also shown with shading.

models simulate ozone depletion and subsequent recovery but with large discrepancies in the absolute TCO values and the expected recovery to 1980 levels (see Dhomse et al., 2018), from here on referred to as D18. The WM and multi model mean (MMM) are similar, given the small number of models considered from the ensemble ($N = 11$). At maximum ozone depletion, around the year 2000, the WM projects a significantly lower ozone concentration (5 DU) than the MMM. This steeper ozone depletion seen in the WM fits the observations better than the MMM, although the modelled inter-annual variability seems to under predict the observations.

The WM predicts a return to 1980 TCO levels by 2056 with a 95 % confidence interval (2052–2060). For comparison the recovery dates presented in D18 were 2062

with a 1σ spread of (2051–2082). Although taken from the same model ensemble (CCMI), the subset of models in this analysis is smaller than that used in D18 meaning that difference in recovery dates between the two works is attributable to both the methodology and the models considered. The smaller number of models used in this study could lead to a narrower confidence interval than the one reported in D18.

The confidence interval for recovery dates is formed from the predictive uncertainty in the WM from the TSAM (for which the 95 % confidence interval is 2054–2059) and the uncertainty associated with the weighting process. Choices made about which models and metrics to include influence the return dates and therefore introduce uncertainty. This is similar to the concept of an "ensemble of opportunity", which is that only modelling centres with the time, resources or interest take part in certain model ensembles. To quantify this uncertainty, we performed a dropout test where a model and a metric were systematically left out of the recovery date calculation. This was done for all combinations of models ($N = 11$) and metrics ($M = 7$), providing a range of 77 different recovery dates between 2052 and 2058. Combining the TSAM and dropout uncertainties produces a 95 % confidence interval of 2052–2060. We additionally tested dropping out up to three metrics at a time and observed that the confidence interval did not notably increase in size.

Figure 2.3 shows the model weights for individual metrics and in total as found using equations 2.1 and 2.2. Good agreement is shown between the models for the metrics of lower stratospheric temperature, the temperature gradient, and the TCO gradient. There is one exception of UMUKCA-UCAM which exhibits a colder pole compared to the ensemble and observations. Resultantly, UMUKCA-UCAM is down-weighted for its lower performance at replicating the historic lower stratospheric temperature. Dissimilarity to the rest of the ensemble will contrastingly increase the weighting but to a lesser effect than the down weighting for performance, due in part to the values of the sigma parameters. In spite of a bias in absolute lower stratospheric temperature, UMUKCA-UCAM does reproduce the trend in the lower stratospheric temperature with similar skill to the other models.

Due to the nudging of temperature that takes place in most of the specified dynamics simulations, we would expect stratospheric temperatures to be reasonably well simulated. However, variation exists in nudging methods in addition to inter

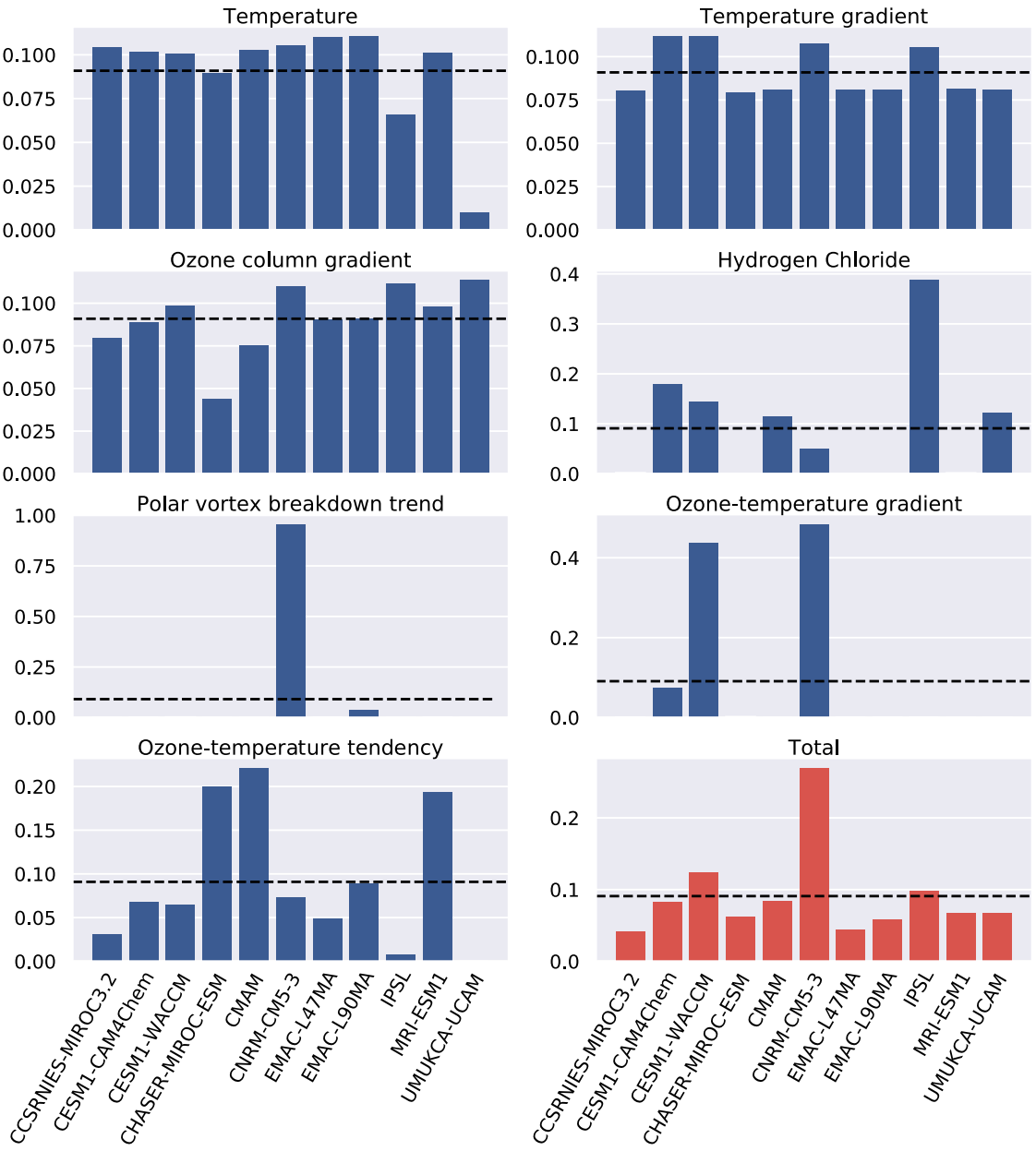

Figure 2.3: Model weights for each of the seven metrics are all shown in blue. The weights account for both performance and independence and are found using equation 2.1. The total weights, as found from equation 2.2, are shown in red and were the weights used to construct the weighted mean shown in Figure 2.2. The black dashed line indicates a uniform weighting as prescribed by a multi model mean.

model differences and this leads to part of the variability in weights (Orbe et al., 2018; Chrysanthou et al., 2019). For the ozone-temperature metrics, which although formed from variables linked to nudged fields are more complex in their construction, we see a much less uniform spread of weights. Furthermore, for processes not directly linked to nudged variables (hydrogen chloride, ozone, and the polar vortex breakdown trend) there is much less agreement between models. This is captured in the weights of these metrics which show just a few models possessing large weights.

The total weighting, formed from the mean of individual metric weights per model, is largely influenced by CNRM-CM5-3, which has a weight of 0.27 (297 % the value of a uniform weighting). The CNRM-CM5-3 simulations are more successful at simulating metrics whilst being reasonably independent from other models, leading to a weight with greater prominence than the other models. This does not mean that CNRM-CM5-3 is the most skilful model. For example, if two nearly identical models had the highest performance, their final weights would be much lower as they would be down-weighted for their similarity. All models are contributing towards the weighted ensemble mean providing confidence that our weighting methodology is not over-tuned and returning model weights of zero. The lowest total model weight is 45 % the value of a uniform weighting.

2.4.2 Testing the methodology

We performed a perfect model test (section 2.3.3) to assess the skill of the weighted mean projection, the results of which are shown in Figure 2.4. The perfect model test shows that, on average, using this weighting methodology produces a WM which is closer to the 'truth' than the MMM by 1 DU. In addition to improvements in projections, the pseudo recovery dates are better predicted on average, with a maximal improvement of 6 yr.

Three models, when treated as the pseudo truth, do not show an improvement of the WM with respect to the MMM. Note that this is not poor performance of the model in question, rather that the weighting methodology does not do an adequate job of creating a weighted projection for that model as the pseudo truth. Using CHASER-MIROC-ESM as the pseudo truth gives a worse WM projection than if we used the MMM. However, the average correlation between the CHASER-MIROC-

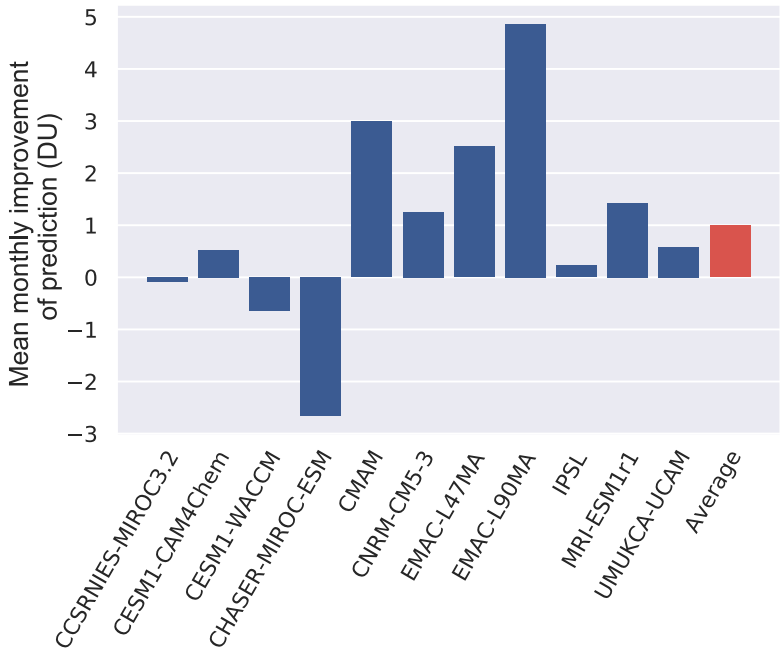

Figure 2.4: Results of the perfect model test. The mean monthly improvement in the Antarctic October TCO projection (1960–2095) of the WM compared to the MMM for each model taken as the pseudo truth. The average shown in red is the improvement across all the perfect model tests. No conclusions about overall model skill should be drawn from this plot.

ESM-simulated TCO and other models in the ensemble is the lowest at 0.65, compared to the average ensemble cross correlation score of 0.81. Since a weighted mean is a linear combination of models in the ensemble, it is understandable that models with low correlation to CHASER-MIROC-ESM will be less skilful at replicating its TCO time series. This is why an improvement is not seen for CHASER-MIROC-ESM as the pseudo truth in the perfect model testing.

We also performed out-of-sample testing on the WM projection for the years 2011–2016 inclusive by comparing it to the TCO observational time series which was smoothed as described in section 2.3.1 to remove inter annual variability. The mean squared error (MSE) was used as the metric for goodness of fit. This range of years is chosen as it is the overlap between the TCO observations, and the years not used in the creation of the weighting. The MSE of the WM is on average 202 DU^2 less than

the MMM per year and the RMSE values were 1510 DU^2 and 2720 DU^2 for the WM and MMM respectively for the out-of-sample period.

2.4.3 Model independence

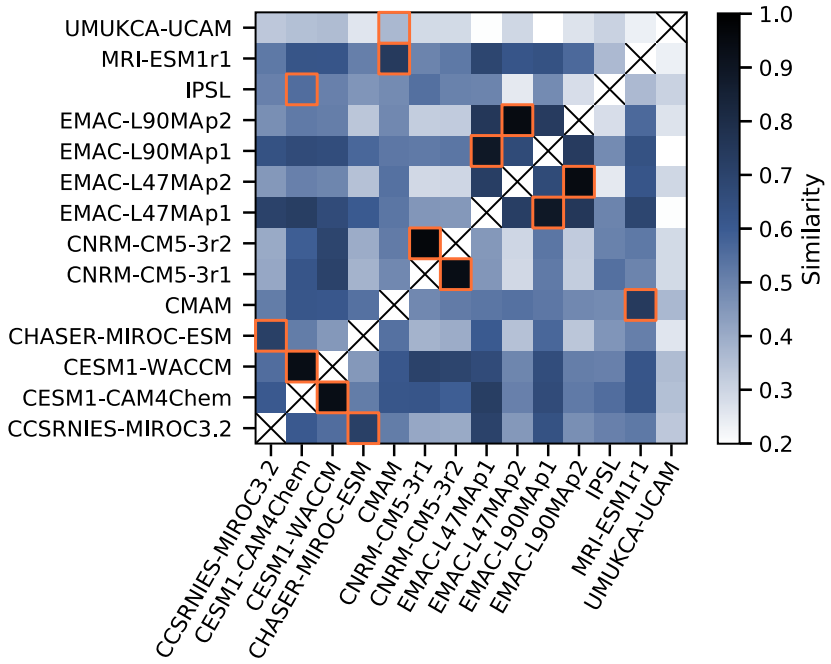

Figure 2.5: Inter-model similarity across all refC1SD models as calculated by equation 2.3. A similarity of 1 denotes models which are identical for all the metrics, whereas a lower similarity shows a greater independence. The orange boxes highlight the model most similar to the model on the y-axis.

The current design of model inter-comparison projects does not account for structural similarities in models, ranging from sharing transport schemes to entire model components. Therefore, a key part of generating an informed weighting is considering how alike any two models are. The weighting scheme presented here accounts for model independence through the denominator in equation 2.1.

The refC1SD scenario from CCMI consists of 14 different simulations, some of which are with different models, whereas others are just different realisations of the same models. Note that there are more models used here than in the creation of the Antarctic ozone projection. This is because for the weighted projection we require both

a refC1SD and a refC2 simulation for each model, but for similarity analysis we can use all the refC1SD simulations. For these model runs we calculated a similarity index s_{ij} (shown in equation 2.3) which is the similarity between models i and j averaged across all the performance metrics, where n_k is the size of the data for metric k.

$$s_{ij} = \frac{1}{M}\sum_{k=1}^{M} \exp\left(\frac{-S_{ijk}^2}{n_k \sigma_S^2}\right) \tag{2.3}$$

Similarities between all refC1SD models are shown in Figure 2.5. We also found the maximum value of s_{ij} for each model, indicating the model which model i is most similar to. The most alike models are the two realisations of CNRM-CM5-3, which are the same models running with slightly different initial conditions. We also see high similarity between the two variations of the CESM model, CESM-WACCM and CESM-CAM4Chem. CESM1-CAM4Chem is the low-top version of CESM1-WACCM, meaning that up to the stratosphere the two models should be much alike (Morgenstern et al., 2017). Analysing the EMAC models this way presents an interesting observation: changing the nudging method, has a greater impact on model similarity than changing the number of vertical levels (the difference between EMAC-L47MAr1i1p1 and EMAC-L47MAr1i1p2, and likewise the 90 level model variant, is that the p1 variant additionally nudges to the global mean temperature (Jöckel et al., 2016)). CHASER-MIROC-ESM and CCSRNIES-MIROC3.2 are two other models which are identified as similar albeit at a lower value. Considering that these two models are built upon the same MIROC general circulation model it is not a surprise that we see a similarity. That the weighting framework can identify all of the models with known similarities (same institution, or realisations) confirms confidence in the methodology and means that we are down-weighting similar models.

2.5 Discussion

The projection of the ozone hole recovery date presented here makes use of an ensemble of the latest generation of CCMs and a weighting methodology that accounts for complexities within model ensembles. While the ozone recovery date found in this work (2056) is different to that found by Dhomse et al. (2018) (2062), these two

dates are not easily comparable as they are created from different subsets of the same ensemble. For our subset of models, the MMM recovery date was 3 years earlier (2053) than the WM. Although the return dates are not significantly different, for the period of peak ozone depletion (especially between 1990 and 2030) the MMM projection is significantly different to the WM. As the model subsets in this work, for the WM and MMM remain the same, the variation in the projections is entirely due to the construction of the WM.

The CNRM-CM5-3 model received the largest weight of 0.27, giving it three times the influence in the WM than in a MMM. Initially this may seem as if we are placing too much importance on one model, but consider that in a standard MMM, a model which runs three simulations with different combinations of components will have three times the influence of a model with a single simulation. Furthermore, CNRM-CM5-3 is not weighted higher because it ran more simulations, it is weighted higher because it is skilful at simulating hindcasts whilst maintaining a level of independence.

Central to the weighting methodology is the selection of metrics requiring expert knowledge. The set of metrics we chose, were grounded in scientific understanding and produce a good improvement of the weighted projection compared to the MMM. There are numerous other metrics of varying complexity which could be considered, such as the size of the ozone hole or the abundance of polar stratospheric clouds. These extra metrics could improve the model weighting and give a more accurate projection, but testing an exhaustive collection of metrics was not our aim, and there are not always appropriate measurements to validate the metrics with. We have shown a weighting framework which improves upon the current methodology for combining model ensembles, and is also flexible and adaptable to which ever metric choices the user deems reasonable. Furthermore, the low range in return dates produced from the dropout testing shows that the results produced in this weighting framework are robust to metric and model choices. This is a desirable effect of a methodology to provide stable results irrespective of fluctuations in the input.

It is reassuring to know that the methodology is robust to metric choices which are compromises between the availability of observational data and expert domain knowledge. For example, the processes known to be important for ozone hole formation are not all measured. In this work we benefit from the decades of interest in polar

ozone which have led to datasets of a length suitable for constructing model weights. This highlights the importance of continued production of good observational datasets because, although perfect model testing allows us a form of testing which forgoes the need for observations, weighting methodologies must be grounded in some estimate of the truth.

Abramowitz et al. (2019) discuss approaches for assessing model dependence and performance, and mention caveats around the notion of temporal transitivity: is model behaviour comparable between two distinct temporal regions? Here, we rephrase the question to be: are the weights generated from the hindcast scenario relevant and applicable to the forecast scenario? This not only questions temporal transitivity, but also that models may have codified differences between scenarios in addition to differences in physical and chemical regimes. In this study, scenario transitivity (as we call it) is demonstrated through perfect model testing. On average the WM produced a better (closer to the pseudo truth) projection than if we had considered the MMM. This shows that weights calculated from the refC1SD hindcasts produce better projections from the refC2 forecasts and are therefore transitive between the two scenarios.

We generated weights from the refC1SD simulations which means that some metrics we chose are based on nudged variables, such as the lower stratospheric temperature gradient. As a result, one might expect that the model skill for these metrics should be equal, although given Figure 2.3 this is not true. One may then expect that the weighting is not capturing model skill, but instead the skill of the models' nudging mechanisms; the models are nudged on different timescales ranging from 0.5 h to 50 h and from varying reanalysis products (Orbe et al., 2020). We use the perfect model test to show that the utility of the weighting methodology is not compromised by using models with such a variety in nudging time-scales and methods.

As the perfect model test produces better projections, for models which are nudged in a variety of ways, we can conclude that the weighting is not dominated by nudging. Take for example UMUKCA-UCAM which is nudged quite differently compared to the ensemble, as evidenced by a southern pole significantly colder than the ensemble. When we take UMUKCA-UCAM as the pseudo truth (temporarily assuming the UMUKCA-UCAM output is the observational truth) we generate weights based upon the refC1SD

simulations and test them on the refC2 simulations. The weights generated are based on the dynamical system simulated in refC1SD which includes any model nudging. We can test how well these weights apply to a different dynamical system without nudging (refC2). As we see an improvement in the WM compared to the MMM we can conclude that the weights generated from the refC1SD dynamical system can be applied to the refC2 dynamical system. If there hadn't been an improvement, then the dynamical systems described by refC1SD and refC2 may be too dissimilar for this weighting methodology and the weights may instead have been dominated by how well models are nudged. Nudging may be influencing the weights, but not to a degree that the accuracy of the projection suffers. Orbe et al. (2020) highlight the need for care when using the nudged simulations and we would like any future work on model weighting to quantify the impact of nudging upon model weights to reflect this.

We justified using the nudged refC1SD simulations, despite these considerations, for two reasons. Firstly, these nudged simulations give the models the best chance at matching the observational record, by providing relatively consistent meteorology across the models. The free running CCMI hindcast simulations (refC1) have a large ensemble variance and, despite producing potentially realistic atmospheric states, are not directly comparable to observational records. Secondly, the perfect model testing discussed above, demonstrates that the nudging doesn't have a detrimental effect on the model weighting.

Although we were not seeking to grade the CCMs as per Waugh and Eyring (2008), the construction of a weighted mean provides insight into model performance which would not be considered in an MMM. This is of some relevance as the CCMI ensemble has not undergone the same validation as its predecessors, such as CCMVal (Eyring et al., 2008). Additionally, we gain insight into model dependence shown in section 2.4.2. Whilst this approach may not be as illuminating as Knutti et al. (2013), where they explored the genealogy of CMIP5 models through statistical methods, or Boé (2018), who analysed similarity through model components and version numbers, it successfully identified the known inter-model similarities. More complex methods are desirable, especially those that consider the history of the models' developments. Nevertheless, the simplicity of quantifying inter-model distances as a measure of dependence lends itself well to model weighting.

2.6 Conclusions

We have presented a model weighting methodology, which considers model dependence and model skill. We applied this over a suite of metrics grounded in scientific understanding to Antarctic ozone depletion and subsequent recovery. In particular we have shown that the weighted projection of the total ozone column trend, with inter-annual variability removed, predicts recovery by 2056 with a 95 % confidence interval of 2052–2060. Through perfect model testing we demonstrated that on average a weighted mean performs better than the current community standard of calculating a multi model mean. Additionally, the perfect model test, a necessary step in validating the methodology, showed a level of transitivity between the free running and the specified dynamics simulations.

This methodology addresses the known shortcomings of an ensemble multi model mean which include, the problem of ensembles including many similar models, and the inability to factor in model performance. It does this by quantifying skill and independence for all models in the ensemble over a selection of metrics which are chosen for their physical relevance to the phenomena of interest. This weighting methodology is still subject to some of the same limitations of taking an ensemble mean: i.e., we are still limited by what the models simulate. For example, in the case of ozone depletion, a weighted mean is no more likely to capture the ozone changes due to the recent fugitive CFC-11 release (Rigby et al., 2019). Instead, it allows us to maximise the utility of the ensemble and, provided we are cautious of over-fitting, it allows us to make better projections.

Addressing the shortcomings and presenting possible improvements of methods for averaging model ensembles is timely given the current running of CMIP6 simulations (Eyring et al., 2016a). That ensemble could arguably be the largest climate model ensemble created to date, in terms of the breadth of models considered. Therefore, the need for tools to analyse vast swathes of data efficiently for multiple interests is still growing. The models within CMIP6 are likely not all independent, which could affect the robustness of results from the ensemble by biasing the output towards groups of similar models. The similarity analysis within this work would allow users of the ensemble data to understand if ensemble biases are emerging from similar models and acknowledge how this may impact their results.

In summary, we have presented a flexible and useful methodology, which has applications throughout the environmental sciences. It is not a silver bullet for creating the perfect projection for all circumstances; however, it can be used to construct a phenomenon-specific analysis process that can account for model skill and model independence, both of which can improve ensemble projections compared to a multi-model mean.

Code and data availability

The jupyter notebook used to run the analysis, along with a collection of functions to produce weightings from ensembles, can be found at `https://doi.org/10.5281/zenodo.3624522`. The CCMI model output was retrieved from the Centre for Environmental Data Analysis (CEDA), the Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation (`http://data.ceda.ac.uk/badc/wcrp-ccmi/data/CCMI-1/output`), and from NCAR's Climate Data Gateway (`http://www.earthsystemgrid.org`).

Acknowledgements

This work was supported by the Natural Environment Research Council [NERC grant reference number NE/L002604/1], with Matt Amos's studentship through the ENVISION Doctoral Training Partnership. Paul J Young is partially supported by the Data Science of the Natural Environment (DSNE) project, funded by the UK Engineering and Physical Sciences Research Council (EPSRC; grant number EP/R01860X/1) as well as the Impact of Short-Lived Halocarbons on Ozone and Climate (ISHOC) project, funded by the UK Natural Environment Research Council (NERC; grant number: NE/R004927/1). We would like to thank Bodeker Scientific, funded by the New Zealand Deep South National Science Challenge, for providing the combined NIWA-BS total column ozone database. We also acknowledge the GOZCARDs team for the production of the HCl record and Remote Sensing Systems for the MSU TLS record. We acknowledge the modelling groups for making their simulations available for this analysis, the joint WCRP SPARC/IGAC Chemistry-Climate Model Initiative (CCMI) for organising and coordinating the model data analysis activity, and the British Atmospheric Data Centre (BADC) for collecting and archiving the CCMI model output. The EMAC simulations have been performed at the German Climate Computing Centre (DKRZ) through support from the Bundesministerium für Bildung und Forschung (BMBF). DKRZ and its scientific steering committee are gratefully acknowledged for providing the HPC and data archiving resources for this consortial project ESCiMo (Earth System Chemistry integrated Modelling). The CCSRNIES-MIROC3.2 model computations were performed on NEC-SX9/A(ECO) and NEC SX-ACE computers at the CGER,

NIES, and supported by the Environment Research and Technology Development Funds of the Ministry of the Environment, Japan (2-1303) and Environment Restoration and Conservation Agency, Japan (2-1709).

Chapter 3

Ensembling geophysical models with Bayesian neural networks

Ushnish Sengupta*[1], **Matt Amos***[2], J. Scott Hosking[3], Carl Edward Rasmussen[1], Matthew P. Juniper[1], Paul J. Young[2].

*The first two authors contributed equally to this work.
[1]University of Cambridge, Cambridge, UK
[2]Lancaster University, Lancaster, UK
[3]British Antarctic Survey, Cambridge, UK

The following work was published in Advances in Neural Information Processing Systems 33 (NeurIPS 2020) on December 2020 (citation: Ensembling geophysical models with Bayesian neural networks, Advances in Neural Information Processing Systems, 33, 2020, `https://proceedings.neurips.cc//paper_files/paper/2020/hash/0d5501edb21a59a43435efa67f200828-Abstract.html`). The published supplementary material can be found in Appendix A, although the plots in section A.4 are expanded upon in Chapter 4 in greater detail accompanied by further analysis. As noted in the manuscript, a small coding oversight led to an increased compute time. For the updated model code see `https://github.com/mattramos/Toy-bayesian-neural-network-ensemble` rather than the links in the manuscript.

The author contributions are listed below.

It should also be noted that the Bayesian neural network described in this chapter is referred to as the BayNNE, though in the thesis introduction and subsequent chapters it is referred to as the BNN. These terms are interchangeable and reflect the evolution of explaining and presenting the BNN to multiple scientific communities.

Statement of contribution

Matt Amos and **Ushnish Sengupta** co-developed the methods (from the initial investigations of **Matt Amos**), performed the analysis and compiled the manuscript. **J. Scott Hosking**, **Matthew P. Juniper**, **Carl E. Rasmussen** and **Paul J. Young** commented on method development, analysis and helped with finalising the manuscript.

Abstract

Ensembles of geophysical models improve prediction accuracy and express uncertainties. We develop a novel data-driven ensembling strategy for combining geophysical models using Bayesian Neural Networks, which infers spatiotemporally varying model weights and bias, while accounting for heteroscedastic uncertainties in the observations. This produces more accurate and uncertainty-aware predictions without sacrificing interpretability. Applied to the prediction of total column ozone from an ensemble of 15 chemistry-climate models, we find that the Bayesian neural network ensemble (BayNNE) outperforms existing methods for ensembling physical models, achieving a 49.4% reduction in RMSE for temporal extrapolation, and a 67.4% reduction in RMSE for polar data voids, compared to a weighted mean. Uncertainty is also well-characterised, with 91.9% of the data points in our extrapolation validation dataset lying within 2 standard deviations and 98.9% within 3 standard deviations.

3.1 Introduction

Climate models are the primary tool for predicting the evolution of Earth's uncertain climate and their output informs international policy (Hoegh-Guldberg et al., 2018). Based on multi-scale physical and chemical processes they allow us to simulate conditions outside of the observational record, both spatially and temporally. Coordinated experiments with an ensemble of multiple models (Taylor et al., 2012; Eyring et al., 2016b) are typically used to increase the accuracy of prediction and to quantify predictive uncertainty, with studies often reporting predictions based on the ensemble average and uncertainty from the ensemble spread.

Such approaches assume that each separate climate model within the ensemble is independent and able to simulate the Earth system with equal skill, neither of which are true (Knutti et al., 2013; Eyring et al., 2019). Climate model skill is established through comparisons against a wide variety of remotely-sensed or in situ observations. Weighted measures of skill from these comparisons provide a more sophisticated method of combining an ensemble compared to simple averaging, and these weighted means are used to constrain ensemble predictions for single or multiple variables of interest (Knutti et al., 2017; Amos et al., 2020). However, there are limitations of these approaches: they assume that historic climate model skills and behaviours can be translated to a future prediction; they rely on observations with the same spatiotemporal coverage as the models and cannot learn model weights for regions where data is sparse; they generally, do not account for the varying quality of observations upon which ensembles are weighted; and they do not account for spatially varying skill in individual models.

Exploiting machine learning methods to analyse and process ensembles of climate models is an emerging area of research. Given the complexity and scale of the data involved, neural networks have obvious benefits. Key examples include emulating climate model ensembles (Knutti et al., 2003), thus saving on high computational costs; identifying regional patterns of climate change (Barnes et al., 2019); and examining and quantifying differences between models' underlying representations of atmospheric physics and chemistry (Nicely et al., 2020). The comprehensive review of Reichstein et al. (2019) provides further examples of how earth system science can – and has – benefited from neural networks and deep learning. Beyond deep learning, causal

inference has provided a new way to analyse the skill of climate model ensembles (Nowack et al., 2020), and newly proposed ensembling methods have improved climate model predictions (Monteleoni et al., 2011; Xu et al., 2019; Ahmed et al., 2020; Merrifield et al., 2020).

In this paper, we address the limitations of current ensembling methods and develop a method which provides more accurate and uncertainty aware predictions. Our approach combines geophysical models within a Bayesian ensembling framework, which assigns spatiotemporally varying weights to models and accounts for heteroscedastic aleatoric uncertainty in observational data, as well as epistemic uncertainty where data is unavailable. By fusing models, which codify our best physical knowledge, with observations, we can better interpolate and extrapolate geophysical data. This provides more accurate future predictions as well as spatiotemporally continuous and observationally constrained historic states. A key strength of our approach is the data model's interpretability, extending its use beyond its predictive capabilities to bring insight and understanding to the climate models.

The code and pretrained models accompanying this paper are hosted in a Github repository `https://github.com/Ushnish-Sengupta/Model-Ensembler`. The dataset of total column ozone observations (Bodeker et al., 2018) and chemistry-climate model predictions from 1980 to 2010 (Morgenstern et al., 2017) are processed, combined and made available as a resource (`https://osf.io/ynax2/download`) for future studies in geophysical model ensembling.

3.2 Methods

3.2.1 Problem formulation

We assume that observations $y(\mathbf{x},t)$ can be modelled as a sum of n physical model predictions $M_i(\mathbf{x},t)$ weighted by their respective weights $\alpha_i(\mathbf{x},t)$, a bias term $\beta(\mathbf{x},t)$ and a heteroscedastic aleatoric noise term $\sigma(\mathbf{x},t)$.

$$y(\mathbf{x},t) = \sum_{i=1}^{n} \alpha_i(\mathbf{x},t) M_i(\mathbf{x},t) + \beta(\mathbf{x},t) + \sigma(\mathbf{x},t) \tag{3.1}$$

Model weights form a partition of unity, i.e., $\alpha_i(\mathbf{x},t) > 0$ and $\sum_{i=1}^{n} \alpha_i(\mathbf{x},t) = 1 \; \forall$ $\mathbf{x},t$. The weights, bias and noise are modelled as probabilistic functions by specifying distributions over the parameters of a neural network. The basic architecture of a neural network embodying these assumptions is shown in Figure 3.1. The physical model weights are represented by the output of a softmax layer which enforces the partition of unity constraint and the weighting of model predictions is performed by a subsequent dot product layer. We choose tanh activations for the hidden layer because its mean output is zero-centered, simplifying our prior design. The extrapolation behaviour of a Bayesian Neural Network with tanh activations outside the training set is also predictably flat, which means the predicted model bias and aleatoric noise does not assume unrealistic values even when we extrapolate.

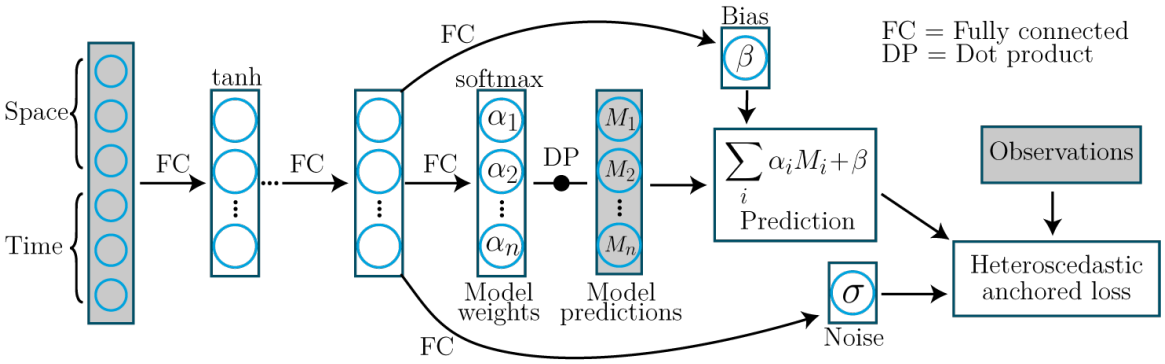

Figure 3.1: Architecture of a neural network where shading depicts external inputs or data. This single neural network is one in a set of identically designed neural networks, that together form the Bayesian ensemble.

3.2.2 Prior design

The computational expense of Bayesian inference is only justified if we are able to encode our domain knowledge into the prior. This can be a challenge since our function space intuitions about the modelled quantity require translation to distributions over parameter values or architecture choices. Input warping is an essential first step if we want to restrict our prior function space to physically realistic functions (Pearce et al., 2019). While only three numbers – latitude, longitude and time – suffice to uniquely identify a datapoint, using these directly as inputs would be problematic

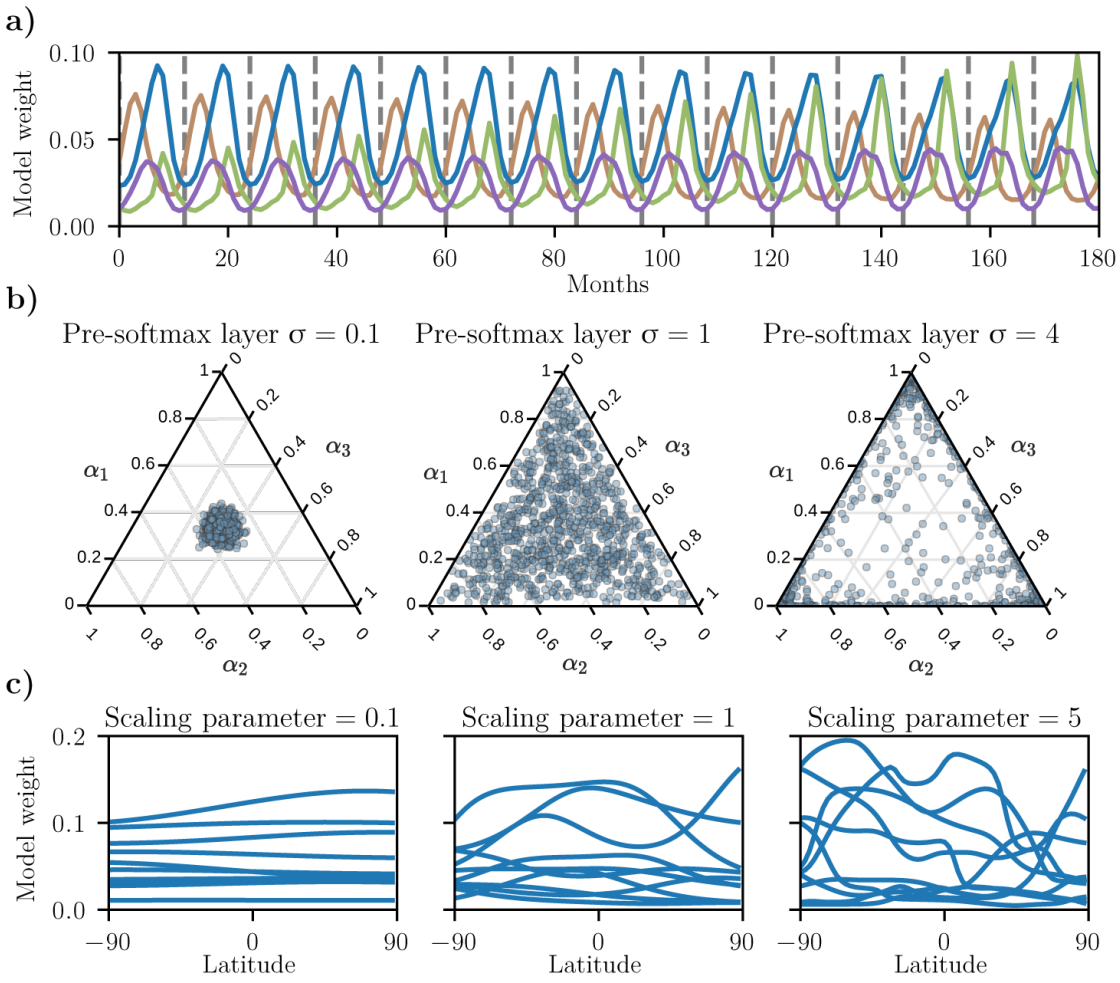

Figure 3.2: Visualizing the prior. a) shows how the warping of the time coordinate enforces desired quasiperiodicity in the physical model weights generated by the prior. Different coloured lines depict the model weights for four samples from the untrained prior and grey dashed lines split the plot into 12 month intervals. b) shows the effect of pre-softmax layer output variance on the prior distribution of physical model weights in the 2-simplex at a particular point in time and space, using 1000 samples from the neural network prior with 3 physical models. c) shows the impact of scaling spatial input on the lengthscales of physical model weights produced by the neural network prior.

because the physical model weights, biases and noise generated by such a prior network would be discontinuous across the 180th meridian and would not respect seasonality. In our case study, locations are represented by their Euclidean coordinates (u, v, w) and the time variable is warped onto the 3D helix $(\cos(2\pi t/T), \sin(2\pi t/T), t)$, where $T = 1\,\text{year}$. This transformation of the time variable makes our prior network generate weights and biases with both a strong annual periodicity and a slow variation over the years (Figure 3.2a), consistent with our expectations. The input variables also need appropriate scaling to ensure that our neural network outputs have the desired characteristic lengthscale. Model skills and biases are likely to vary over the typical lengthscales spanned by climatic or geographic regions that and the scaling of the spatial coordinates should be such that this is reflected in the prior. Scaling is also important for temporal input variables and it can be used to magnify or suppress seasonal or yearly variations. Figure 3.2c demonstrates the effect of three different scaling choices for the w coordinate, $[-0.1, 0.1]$, $[-1, 1]$ and $[-4, 4]$ on randomly drawn samples from the prior: increasing the scale factor favours functions with larger high frequency components.

Another key component of a well-formulated prior is the variance of the pre-softmax layer. Since our physical models are competitive, a priori we should assume that any model combination should be possible at any point in time and space and the α_i-s generated by the prior network should be distributed approximately uniformly in the physical model weight simplex. We should therefore choose a prior variance for the incoming connections to the pre-softmax layer such that the variance of the untrained layer outputs is close to 1.0. Figure 3.2b shows random samples of α_i at an arbitrary point from an untrained neural network with 3 models. It helps us visualize how a small pre-softmax layer standard deviation (0.1) constrains the prior α_i to lie close to the naive multi-model mean, whereas a standard deviation that is too large (4.0) pushes all the prior probability mass towards the corners of the simplex.

Similarly, the prior variances of incoming connections to the model bias term should be scaled to restrict the bias term to zero prior mean and a small variance. While the model bias is necessitated by the fact that certain regions can be modelled poorly by all the physical models, we would prefer to have our combination of physical models do the bulk of the modelling. The prior variances of connections to the heteroscedastic

noise term should likewise be scaled. However, unlike the bias term whose distribution should be zero-centered, the noise should have a positive mean added to it that is informed by our knowledge of the average quality of our observations.

Finally, the number of units in the hidden layer(s) should increase commensurately with the size and resolution of our dataset. Overparameterization leads to desirable Gaussian process-like behaviour (Lee et al., 2017) whereas an underparametrized Bayesian neural network, unable to produce the intricate functions demanded by a large dataset, will have its posterior collapse to a single point and lose all epistemic uncertainty outside of the training dataset. Judicious use of prior predictive checks in conjugation with domain knowledge can thus circumvent the need for expensive hyperparameter tuning or hyperpriors entirely for our simple network and create a well-informed prior.

3.2.3 Approximate inference using randomized MAP sampling

Inference is complicated by the size of a typical geospatial dataset for example, the ozone column dataset used as our case study contains over 2 million datapoints. This rules out more expensive gold-standard inference techniques such as Markov Chain Monte Carlo (Neal, 2012) or Hamiltonian Monte Carlo (Chen et al., 2014). Mean-field variational inference (Blundell et al., 2015) or Monte Carlo dropout (Gal and Ghahramani, 2016), while scalable, are also inappropriate because one of the objectives of ensembling models is to fill-in missing data and it has been demonstrated (Foong et al., 2020) that these techniques are excessively overconfident between well-seperated clusters of training data. Here, we have chosen a state-of-the-art approximate inference technique: approximately Bayesian ensembling using randomized maximum a posteriori (MAP) sampling (Pearce et al., 2018). Ensembling had already been shown (Lakshminarayanan et al., 2017) to be empirically effective at providing calibrated estimates of uncertainty for neural networks and the randomized MAP sampling approach grounds this in Bayesian theory. For the j-th neural network ensemble member, we draw a sample from the prior distribution over parameters (assumed multivariate normal) $\boldsymbol{\theta}_{anc,j} \sim \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})$ and compute the MAP estimate corresponding to a prior re-centered at $\boldsymbol{\theta}_{anc,j}$.

$$\boldsymbol{\theta}_{MAP,j} = \text{argmax}_{\boldsymbol{\theta}_j} \log(P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j)) - \frac{1}{2}\|\boldsymbol{\Sigma}_{prior}^{-1/2}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})\|_2^2 \tag{3.2}$$

If we now consider a dataset of N_D observations $\{y_i, \mathbf{x}_i, t_i\}$ and specify the data likelihood $P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j)$ for our regression task by assuming heteroscedastic Gaussian noise $\sigma(\mathbf{x}, t)$, we may equivalently minimize the following loss function for the j-th neural network

$$\text{Loss}_j = \sum_{i=1}^{N_D} \frac{(y_i - \hat{y}_j(\mathbf{x}_i, t_i))^2}{\sigma_j^2(\mathbf{x}_i, t_i)} + \sum_{i=1}^{N_D} \log(\sigma_j^2(\mathbf{x}_i, t_i)) + \|\boldsymbol{\Sigma}_{prior}^{-1/2}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})\|_2^2 \tag{3.3}$$

The prediction of a trained ensemble with n_e neural networks is therefore a mixture of n_e Gaussians, each centered at $\hat{y}_j(\mathbf{x}_i, t_i)$ with a variance of $\sigma_j^2(\mathbf{x}_i, t_i)$. For computational convenience, we approximate this mixture as a single Gaussian with mean $\frac{1}{n_e}\sum_j \hat{y}_j$ and variance $\frac{1}{n_e}\sum_j \sigma_j^2 + \frac{1}{n_e}\sum_j \hat{y}_j^2 - (\frac{1}{n_e}\sum_j \hat{y}_j)^2$, following similar treatment in (Lakshminarayanan et al., 2017). This also allows us to decompose the total predictive uncertainty into an aleatoric component (first term) and an epistemic component (second and third terms).

To disambiguate any reference to ensembling or ensembles in this paper, we refer to the combination of geophysical models as the "physical model ensemble" and to the set of neural networks used for approximate inference as the "neural network ensemble".

An outline of the algorithm used to train our BayNNE is provided below.

3.3 Experiments

3.3.1 Synthetic data

To validate the Bayesian neural network ensembler (BayNNE), we create a toy problem where the ground truth is known. The "monthly observations" are generated by the function $0.5\left(\frac{\text{lat}}{90}\right)^2 + 0.25\sin\left(2\pi\frac{\text{lon}}{180}\right) - 0.2\cos\left(\pi\frac{\text{mon}}{12}\right)$ (lat, lon and mon are latitude, longitude, and month number respectively) with varying levels of added Gaussian noise in different regions to simulate heteroscedasticity– the noise standard deviation is 0.01 in the northern region (north of 30°N), 0.02 in the tropics (between 30°S and 30°N) and 0.03 in the southern region (south of 30°S). The four "physical models" replicate the observations but only in distinct geographical regions: model 1 is correct in the northern region where it has a bias of +0.03 w.r.t. the observations, models 2

Algorithm 1: Algorithm for initialising and training the BayNNE

Input : Training dataset of N_D observations and physical model predictions corresponding to locations and times $\{\mathbf{x}_i, t_i\}$, physical model predictions for N_T locations and times with missing observations $\{\mathbf{x}_k, t_k\}$

Output : Mean and variance predicted by metamodel for $\{\mathbf{x}_k, t_k\}$

1 Transform latitude, longitude and time of each datapoint to 6-dimensional space-time input $(\cos(lat_i)\sin(lon_i), \cos(lat_i)\cos(lon_i), \sin(lat_i), \cos(2\pi t_i/T), \sin(2\pi t_i/T), t_i)$.

2 Rescale each column of space-time inputs to the range $[-a, a]$. Use larger scales a for input variables on which we expect model weights/ bias to have stronger dependence.

3 Set prior variances of the fully connected layer weights to l_i/n_{input}, where n_{input} is the number of nodes in the previous layer.

4 Tune l_i by performing prior pushforward checks – the output of each fully connected layer should have mean ~ 0 and variance ~ 1.0, except those that feed the bias and noise terms, whose output variance should be small.

5 Initialize n_e neural networks by drawing samples from the prior over parameters.

6 **for** $j \leftarrow 1$ **to** n_e **do**

7 Draw a random sample $\boldsymbol{\theta}_{anc,j}$ from the prior over parameters.

8 Anchor the loss function of j-th neural network to $\boldsymbol{\theta}_{anc,j}$, so that $\text{Loss}_j = \sum_{i=1}^{N_D} \frac{(y_i - \hat{y}_j(\mathbf{x}_i, t_i))^2}{\sigma_j^2(\mathbf{x}_i, t_i)} + \sum_{i=1}^{N_D} \log(\sigma_j^2(\mathbf{x}_i, t_i)) + \|\boldsymbol{\Sigma}_{prior}^{-1/2}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})\|_2^2$.

9 Train with ADAM until convergence.

10 **end**

11 **for** $k \leftarrow 0$ **to** N_T **do**

12 $\mu_{pred,k} = \frac{1}{n_e}\sum_j \hat{y}_j(\mathbf{x}_k, t_k)$

13 $\sigma_{pred,k} = \frac{1}{n_e}\sum_j \sigma_j^2(\mathbf{x}_k, t_k) + \frac{1}{n_e}\sum_j \hat{y}_j^2(\mathbf{x}_k, t_k) - (\frac{1}{n_e}\sum_j \hat{y}_j(\mathbf{x}_k, t_k))^2$

14 **end**

15 Compute negative log-likelihood of predictions on test data. If NLL not converged, return to step 5 and train more neural networks.

16 return μ_{pred}, σ_{pred}

and 3 are correct and unbiased in the equatorial region and model 4 is correct in the south with a bias of -0.03. In regions where the models are not designed to be skilful, they output random noise. Model predictions and observations are shown in Figure 3.3. The synthetic observations span 20 years, and we train the BayNNE on 85 % of the data from the first 10 years. The last 10 years are left for out-of-sample validation.

Results of a BayNNE with 50 neural network ensemble members with 1 hidden layer of 100 nodes trained on this synthetic dataset are shown in Figure 3.3. We observe that it has successfully recovered the expected physical model weights: models only have weights in the regions where they are skilful and where multiple models are equally skilful (models 2 and 3 in the equatorial region), they are assigned equal weights on average. We also find that the magnitudes of the recovered model biases and aleatoric noise match their engineered values. The uncertainty quantification is excellent out of sample with 68.2, 95.4 and 99.7 percent of points lying within 1, 2 and 3 standard deviations respectively. The overall predictive skill is consistent across the training, testing and out of sample datasets, with near-optimal RMSEs of 0.022, close to the average noise in our observations. This test validates the ability of the BayNNE to successfully capture model skill, bias and aleatoric noise, demonstrating competence in accurate ensembling.

3.3.2 Total column ozone dataset

For a more compelling case study with real-world implications, we consider the problem of predicting monthly averaged total column ozone, which is the integrated amount of ozone from the surface to the atmosphere's boundary with space. Total column ozone provides a good estimate of stratospheric ozone and its variability, as approximately 90 % of ozone resides in the stratosphere. Studying and predicting ozone concentrations is an important scientific endeavour, particularly for monitoring the impacts of the Montreal protocol which was designed to protect the ozone layer from anthropogenic emissions (WMO, 2018). This sustained interest has produced a good coverage of observational records and models suited to simulating ozone, both of which we use.

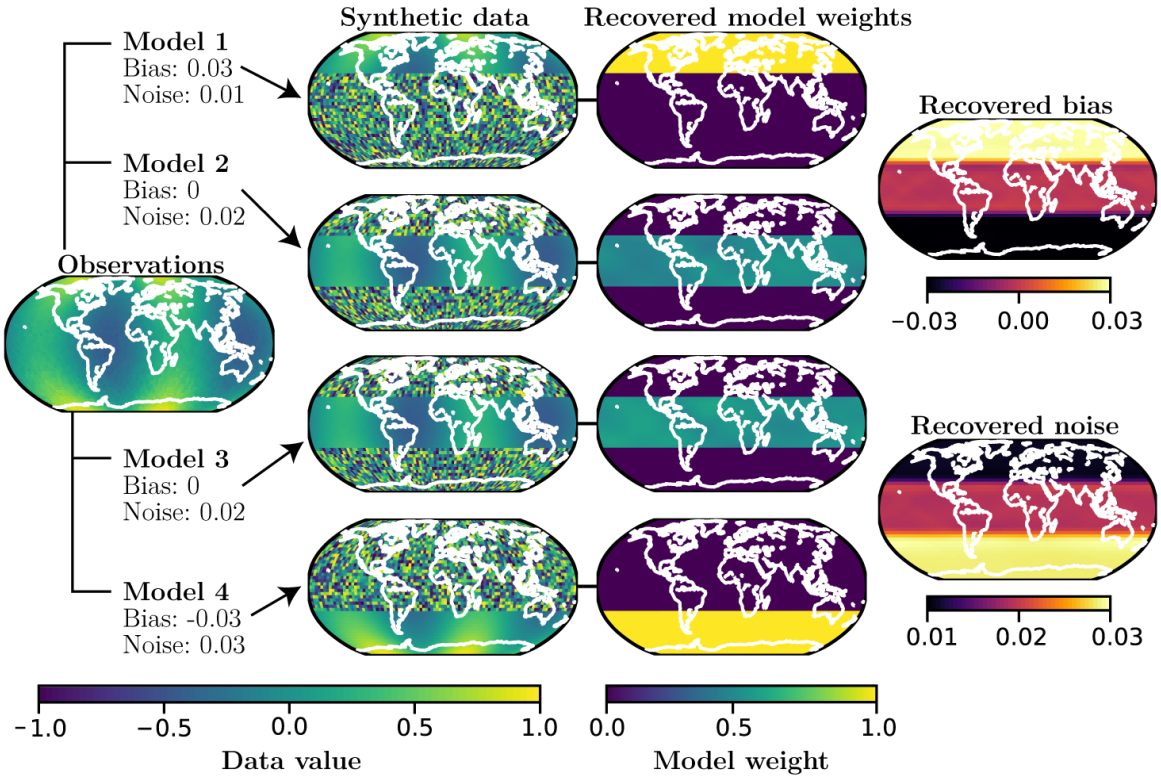

Figure 3.3: Summary of the toy problem results, showing the synthetic observations and model predictions for a single month 5 years out of sample. These are shown alongside the model weights recovered by the BayNNE for each model, the model bias, and aleatoric noise.

3.3.2.1 Description of dataset

We use total column ozone output from 15 chemistry-climate models within the Chemistry-Climate Modelling Initiative (CCMI)(Morgenstern et al., 2017). This ensemble of process models simulates the changing climate, the chemical composition of the atmosphere, and couplings within the chemistry-climate system. We use the hindcast simulation (1980–2010) from CCMI where the models have been nudged (Orbe et al., 2020) so that they replicate the observed meteorology of the past. This simulation represents the models' best attempt at recreating the chemical composition for this 30-year period, making it a suitable comparison to observations.

The observations we use are the NIWA-BS total column ozone record (Bodeker et al., 2018) which is a dataset constructed from satellites and ground-based observations. Coverage is limited by satellite availability and the method of observation, which for some instruments requires daylight. For this reason, a large proportion of missing observations are over polar regions during winter months, an area of interest as this covers the formation of the ozone hole.

3.3.2.2 Constructing the validation set

Significant spatial and temporal correlations present in geospatial data mean that a randomised train-test split will not adequately validate the skill of the BayNNE. Instead, we withold a set of data for validation which mimics the spatial structure and the temporal occurrence of the data missing from the observations. This is a more rigorous test of the intended application of the BayNNE. To test interpolation, we consider 2 forms of data voids in the observations: data missing from tropical regions (often gaps between satellite tracks) and small irregular missing features. Using historic satellite data, we create synthetic patterns resembling those usually associated with missing data due to incomplete satellite coverage, sometimes covering the majority of the tropics. The total data withheld for the purposes of interpolation validation is 24 months of the entire region 30°S to 30°N, 48 months of synthetic data voids due to incomplete satellite coverage and an additional 500 randomly distributed small scale features (up to $15° \times 15°$). We test temporal extrapolation by withholding the last 3 years of data, and spatial extrapolation by withholding data from polar regions. The latter is either an area extending from a latitude of either 60° or 70° to the pole,

which replicates the inability of some instruments to measure ozone at high latitudes during wintertime (Kiesewetter et al., 2010). In total, 24 months of polar cap data for both the north pole and south pole is used for the purpose of validation. Overall, the BayNNE is trained on 77 % (1.8 million datapoints), tested on 4 % (85,000 datapoints) and validated on 19 % (440,000 datapoints) of the available data.

3.3.2.3 Results

The BayNNE used to ensemble the 15 chemistry-climate models for predicting total column ozone comprises of 65 neural network ensemble members, each containing a single hidden layer with 500 nodes. An ensemble of 65 neural networks appropriately sampled the possible combinations of physical models and ensured convergence of the weights, bias, noise and therefore, prediction which can be seen in Figure A.1. The width of the single hidden layer provides adequate overparameterisation as the weight functions the BNN is learning are smooth and simple, and epistemic uncertainty does not collapse out of sample.

Comparisons between BayNNE and commonly used ensembling and interpolation methods are shown in Table 3.1. Interpolation in non-polar regions, including predominantly large gaps in the tropics from incomplete satellite coverage, is compared against bilinear interpolation (Abatzoglou et al., 2018; Meher and Das, 2019), and spatiotemporal kriging (Wardah et al., 2011; Yang and Hu, 2018) using a stochastic variational Gaussian process (Matthews et al., 2017) on 3 year sections of observational data. Spatial and temporal extrapolation skill (root mean squared error) is compared to a uniformly globally weighted multi-model mean (Lamarque et al., 2013a; Dhomse et al., 2018) and 2 weighted means where weights per model are found from the ability of a model to replicate observations in the training set (Knutti et al., 2017; Sanderson et al., 2017). The reader is referred to the supplementary information file for more details on prior design for the BayNNE, training and baseline comparisons.

The BayNNE predictions are significantly better than the baselines in nearly all subsets of the validation dataset (Table 1). Particular improvement over existing methods is seen for ozone predictions over the southern polar cap and for future predictions. Chemistry-climate models are typically less good at simulating ozone over the south pole compared to the north pole due to cold biases (Eyring et al.,

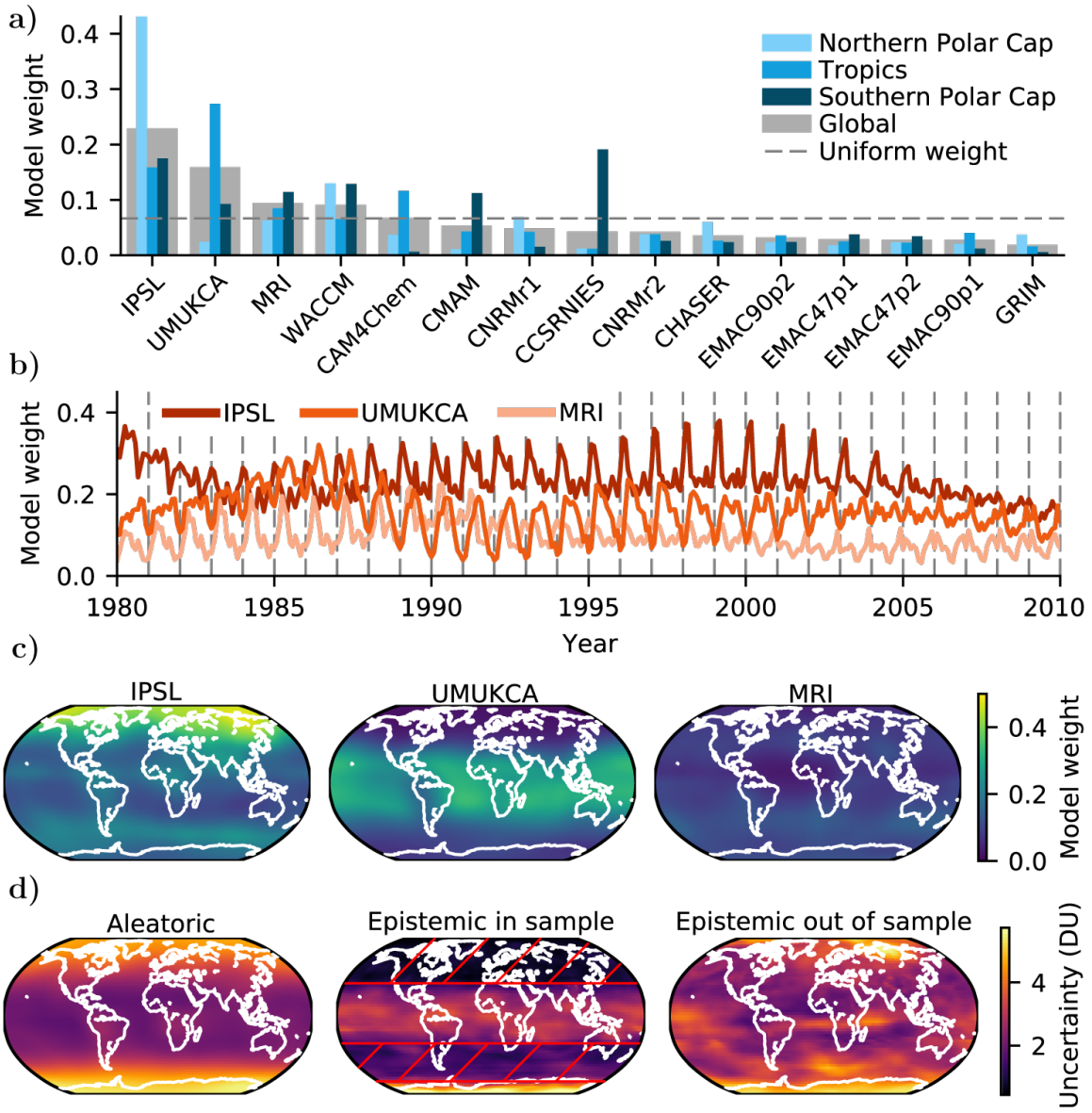

Figure 3.4: Physical model weights and uncertainties recovered by BayNNE when predicting total column ozone. a) shows the average weight per chemistry-climate model globally and for three regions of interest. b) depicts the spatially averaged model weight for the same three models. Vertical dashed lines show the beginning of the year. c) shows the temporally averaged model weight for the three highest weighted models. d) shows the predicted average aleatoric uncertainty and two temporal snapshots of epistemic uncertainty: in sample (with red hatching depicting areas with available training data), and out of sample. DU is a Dobson unit; a measure of the amount of a trace gas in a vertical column.

Table 3.1: Area weighted root mean squared errors of predictions in Dobson units using various methods. NP and SP represent missing north polar and southern polar cap data respectively. For interpolation 'Tropics' covers a block of missing data 30°S to 30°N, 'Satellite voids' (SV) represents the incomplete satellite coverage in the tropics, and small features (SF) are up to $15° \times 15°$.

	Extrapolation			Interpolation		
	Temporal	SP	NP	Tropics	SV	SF
Multi model mean	15.7	30.5	8.8	9.8	9.2	16.4
Weighted mean	8.7	22.1	12.3	8.2	8.5	10.2
Spatially weighted mean	9.8	19.6	6.6	5.5	5.2	10.0
Spatiotemporal kriging*	–	–	–	7.0	2.2	3.4
Bilinear interpolation*	–	–	–	31.2	**1.7**	3.4
BayNNE	**4.4**	**6.6**	**4.7**	**2.7**	2.1	**3.2**

* Used for interpolation only

2006) and discrepancies in simulating the polar vortex (Lin et al., 2017; Gillett et al., 2019). However, by spatiotemporally identifying skilful models, BayNNE is better at predicting southern polar ozone than other ensembling standards in the modelling community. Skill in temporally extrapolating is also much improved, meaning that using BayNNE may provide more accurate future predictions, which is extremely desirable.

Interpretability, a key benefit of using this technique, is highlighted in Figure 3.4, allowing for identification of seasons and regions in which particular models contribute more to the ensemble prediction. For example, we identify the three chemistry-climate models with the largest contributions to the ensemble prediction: IPSL, MRI and UMUKCA. IPSL contributes highly, particularly in far northern regions but also globally, and UMUKCA is the most dominant model in the tropics. Also of note (not shown), is the CCSRNIES model, whose weight for the southern pole in austral winter is approximately half, overwhelmingly making it the dominant model for predicting the build-up period before the springtime Antarctic ozone hole.

Figure 3.4d shows how the BayNNE successfully handles uncertainty. Epistemic uncertainty is increased for regions lacking observations and temporally out of sample. This is highly desirable behaviour as we do not want an ensembling method to be

overly confident beyond data it has seen. To quantify success at handling uncertainty we compare the number of predictions which lie within 1,2 and 3 standard deviations of the truth, which ideally should be 68.3 %, 95.5 % and 99.7 % respectively. These values are: 61 %, 92 % and 99 % for temporal extrapolation; 62 %, 94 % and 99 % for the south pole; and 56 %, 87 % and 97 % for the north pole. The negative log likelihoods for these three regions are -2.56, -2.19 and -2.33.

3.4 Conclusions

We have presented Bayesian neural network ensembling (BayNNE), a principled approach to geophysical model ensembling, which learns spatiotemporally varying model weights and bias based upon the physical models' ability to replicate observations. This uncertainty-aware approach incorporates a heteroscedastic aleatoric uncertainty, which accounts for the varying quality of observational data and other sources of irreducible uncertainties. Additionally, the epistemic uncertainties inherent in a Bayesian framework prevent overconfident extrapolation when BayNNE is not constrained by observations.

We have validated BayNNE on a synthetic dataset where we demonstrated its ability to recover the correct model weights, biases and noise. We then applied it to the more challenging problem of ensembling 15 chemistry-climate models to predict total column ozone. BayNNE predictions were significantly more accurate than current ensembling techniques, both temporally out-of-sample and for infilling historic observations. As a result, we have produced an accurate and complete gridded reconstruction of total column ozone for the period 1980–2010, which offers new insights, particularly for the ozone hole. Interpretability is maintained and model weights/ biases offer an understanding of localised model performance, allowing diagnosis by modellers. Considering that most physical model ensemble weighting techniques do not vary weights in space and time and do not take account of uncertainties in the observations used to create model weights, this ensembling technique represents a significant improvement.

Accurately ensembling geophysical models (e.g. climate models) improves the predictive capability of the ensemble, allows for better investigation of historic

conditions through the imputation of discontinuous observations and in the case of climate models, is vital for investigating the evolution of the climate. Moreover, quantifying the certainty of predictions is fundamental for constraining future change and describing our confidence in the predictions. Future work should not only look at applying this tool to other climate modelling problems but also to problems in other disciplines, such as hydrology, where competing model predictions need to be similarly combined in light of observational evidence. We note that the nudged chemistry-climate models in our case study have their behaviour partially constrained by observed meteorology, whereas free running models predicting the future cannot have this constraint. A proper treatment of the chaos-induced uncertainty in free running models would be worth investigating for use in forecasting using ensembles. Finally, it would also be interesting to consider physical model weights as a function of model variables (e.g. ozone-temperature gradient) that causally impact model skill, instead of proxies like location and time, as this may improve forecast accuracy.

Broader Impact

We created an ensembling technique which takes into account the limitations of observations and models. This method is applicable to many geophysical models (e.g. hydrological, regional climate and chemistry-climate models) although nuances in each field and model ensemble mean the BayNNE should not be blindly used.

Positive impacts include more accurate and better constrained predictions from model ensembles. This could shift the standard of how model ensembling is performed, leading to this method (or derivatives) influencing scientific understanding and downstream policy decisions. The greater understanding offered by combing models and observations in this way, has the potential to open up sparse historic observational records, through fusion with geophysical models. This would, for example, allow for greater understanding of historic climate states.

The response to climate change is influenced by predictions formed from model ensembles, and although accurate and appropriately certain ensembling could result in more definitive and correctly concentrated mitigation efforts, highly certain but wrong predictions could lead to an incorrect pooling of resources and result in negative socio-economic impacts. For these reasons we must be mindful about dangers of extrapolating and unknown errors in observational datasets which incorrectly bias results.

Acknowledgements

This work was supported by the Natural Environment Research Council [NERC grant reference number NE/L002604/1], with Matt Amos's studentship through the ENVISION Doctoral Training Partnership. Ushnish Sengupta is an Early Stage Researcher within the MAGISTER consortium which receives funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766264. The project was also supported with research credits provided by Google Cloud.

Paul J Young is partially supported by the Data Science of the Natural Environment (DSNE) project, funded by the UK Engineering and Physical Sciences Research Council [EPSRC grant number EP/R01860X/1].

We acknowledge the modelling groups for making their simulations available for

this analysis, the joint WCRP SPARC/IGAC Chemistry-Climate Model Initiative (CCMI) for organising and coordinating the model data analysis activity, and the British Atmospheric Data Centre (BADC) for collecting and archiving the CCMI model output. We would like to thank Bodeker Scientific, funded by the New Zealand Deep South National Science Challenge, for providing the combined NIWA-BS total column ozone database.

Chapter 4

Geophysical interpretability of the Bayesian neural network

This unpublished work extends the published technical paper in Chapter 3, providing an atmospheric science-led investigation into the interpretability of the Bayesian neural network for the purpose of ensembling chemistry-climate models. It is entirely my own work.

4.1 Introduction

The previous chapter detailed the development of a Bayesian neural network (BNN) that implemented a weighting strategy for an ensemble of geophysical models to produce uncertainty aware ensemble predictions (Sengupta et al., 2020). We demonstrated the utility of the BNN by producing a continuous historic record of total ozone column by combining sparse observational data (Bodeker et al., 2018) with chemistry-climate models (CCMs) from the Chemistry-Climate Model Initiative (CCMI) (Morgenstern et al., 2017). That chapter was predominantly methods focused and through rigorous statistical testing highlighted the predictive skill of the BNN and its ability to successfully capture uncertainty.

This chapter presents an investigation into the interpretability of the BNN, asking questions about what information can be extracted about observational quality and CCMs from the BNN which produced the infilled historic total ozone column. While machine learning methods can make informed and accurate decisions and predictions, they are often seen as opaque, particularly deep learning, with the reasons leading to a particular decision or prediction obfuscated in largely undecipherable statistical model parameters (McGovern et al., 2019; Samek et al., 2019). With this in mind, we designed the BNN to be readily interpretable to provide useful information about the physical models and observations it assimilates, and to build trust. One of the main strengths of the Bayesian neural network is that it breaks down complex predictions into a more readily understood notion of ensembling: linearly combining weighted models.

As described in Chapter 3, the BNN models observations $y(\mathbf{x},t)$ as a sum of n physical model predictions $M_i(\mathbf{x},t)$ weighted by their respective weights $\alpha_i(\mathbf{x},t)$, a bias term $\beta(\mathbf{x},t)$ and a heteroscedastic aleatoric noise term $\sigma(\mathbf{x},t)$.

$$y(\mathbf{x},t) = \sum_{i=1}^{n} \alpha_i(\mathbf{x},t) M_i(\mathbf{x},t) + \beta(\mathbf{x},t) + \sigma(\mathbf{x},t) \tag{4.1}$$

Here we explore the BNN parameters defined above to better understand the quality of observational data, gain insight into the complex landscape of CCMs and to investigate how interpretable the BNN is. The BNN parameters are discussed in turn in the following sections: the model weights, bias and uncertainties. The infilled dataset we

examine here, and the BNN used to produce it, are the same as described in Chapter 3 for infilling the total ozone column record, and so the reader is referred there for the technical details.

Throughout this chapter the term *model* is used in reference to physical models, often interchangeably with *CCM*. This clarification is important to avoid confusion because the BNN is itself a statistical *model*. The term *weight* also has two meanings in this thesis. It can refer to the numerical weights of neural network layers or the weight of a physical model, calculated within a weighted mean or by the BNN. It is the latter definition we use in this chapter.

4.2 Model weights

Model weights (α) represent how much individual CCMs are contributing to the overall ensemble prediction. In Chapter 2 weights represented a combined measure of model performance and independence, whereas weights learnt by the BNN are not explicitly designed to quantify performance. This is because the BNN learns how to optimally combine models in order to match the observations, rather than comparing each individual model prediction to the observations. An important aspect in the construction of the BNN is its ability to learn weights that vary spatially, temporally and seasonally, unlike weights in Chapter 2 which are time and space invariant.

In this section we dissect the BNN derived model weights as a function of space and time and investigate if the BNN learning is representative of model behaviours. An initial spatiotemporal summary of model weights is shown in Figures 4.1 and 4.2, which is followed by a comparison between weights and model similarity or performance to investigate to what extent model performance and similarity can be inferred from the BNN derived model weights.

Figure 4.1 shows the average weight in space for each of the 15 CCMs. Large variation in the spatial patterns is evident between the models, indicating the strong influence of spatial coordinates on the weighting, particularly the dependence of weights on latitude. Some models exhibit very localised weighting, such as CCSRNIES that only has non-negligible weight over the southern pole and UMUKCA that has the majority of its weight over tropical regions. Nonuniform spatial weighting for different

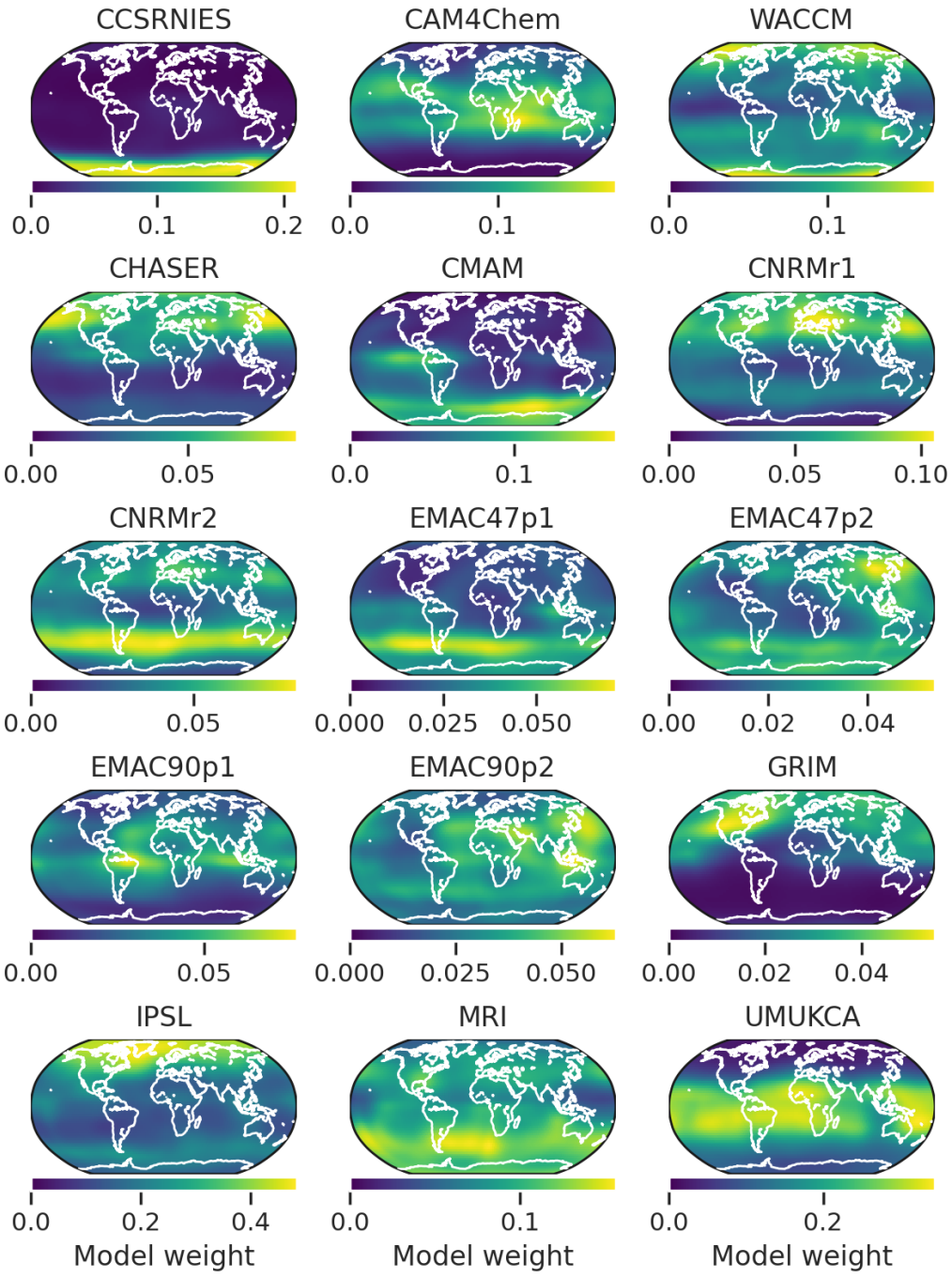

Figure 4.1: Model weights in space (averaged over 1980–2010) from the total ozone column BNN for each of the 15 CCMs. Note the different colour bar scale for each subplot.

models implies that there is a benefit to permitting the spatial variation of weights, which is seen through an improvement in BNN prediction accuracy.

Average weight in time for each of the 15 CCMs is shown in Figure 4.2. All weights display strong oscillatory behaviour with an annual periodicity, which justifies our prior belief used in the design of the BNN that weights are seasonally dependent. CCSRNIES and CHASER, built upon the same general circulation model, exhibit particularly strong seasonality that means their weights are negligible during boreal summer. Some weights display large scale temporal dependence, with some particularly abrupt changes, such as MRI in 1992, suggesting a similarly quick change in the observational quality or model prediction.

Both Figures 4.1 and 4.2 show how varied weights can be across the model ensemble. The maximum weight at a single spatiotemporal point is about 0.6, meaning that for that grid square a single model prediction contributes about 60 % of the entire ensemble prediction. Conversely, it not uncommon for weights to be approximately 0 and therefore the corresponding model has no contribution to the ensemble prediction. That weights are highly spatiotemporally dependent is a result of the fact that the BNN prediction improves significantly when assuming that model contributions should not be constant in time and space. Therefore, more generally, the predictions from model ensembles are improved when we assume that model performance, and subsequently model contribution, is spatiotemporally varying.

4.2.1 Estimating model similarity from weights

In any physical model ensemble, it is unlikely that all models are independent of one another (Abramowitz et al., 2019; Amos et al., 2020). For ensembles, such as those within CCMI or the coupled model intercomparison project (CMIP6) (Eyring et al., 2016b), institutions often submit similar models that may be based on the same underlying general circulation model or share large sections of model code. Similarly, model components such as an ocean module, may be shared between institutions and be part of multiple models. Although attempts have been made to show (Masson and Knutti, 2011) and account for (Knutti et al., 2017; Amos et al., 2020) model similarity, quantifying it is challenging because, for example, it is not known how similar two models that use the same ocean module are, given that these similarities are often

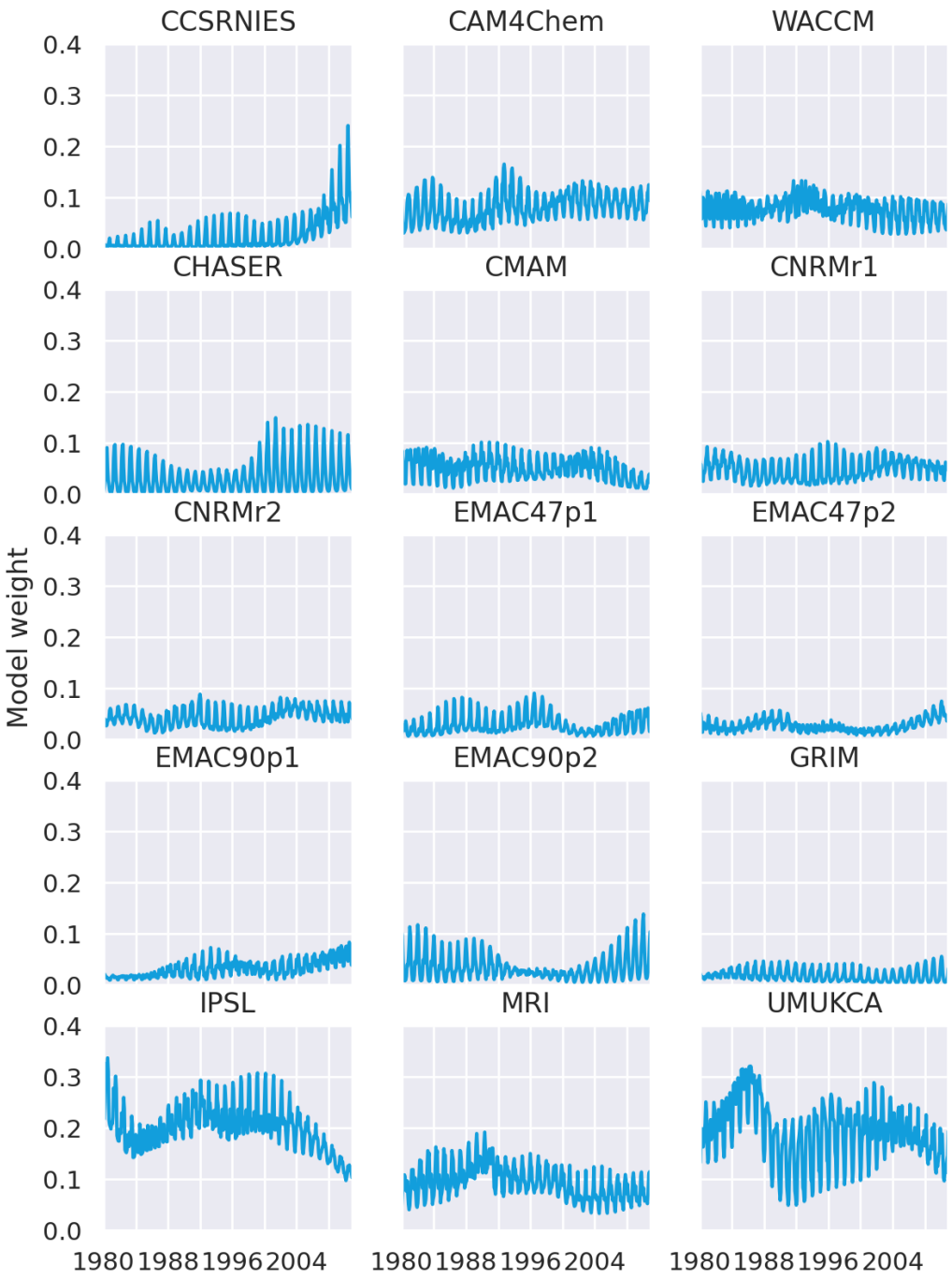

Figure 4.2: Model weights in time (globally averaged) from the total ozone column BNN for each of the 15 CCMs.

hidden in extremely long and complex model code. In the BNN's predictions, we do not need to consider the similarity of models because models are weighted simply by how much their inclusion improves the assimilated ensemble prediction. However, similarity in the spatiotemporal patterns of the model weights from the BNN may be indicative of underlying model similarities. We explore this here and investigate whether the BNN derived weights are consistent with known model similarities and if they match those calculated in Chapter 2 (Amos et al., 2020).

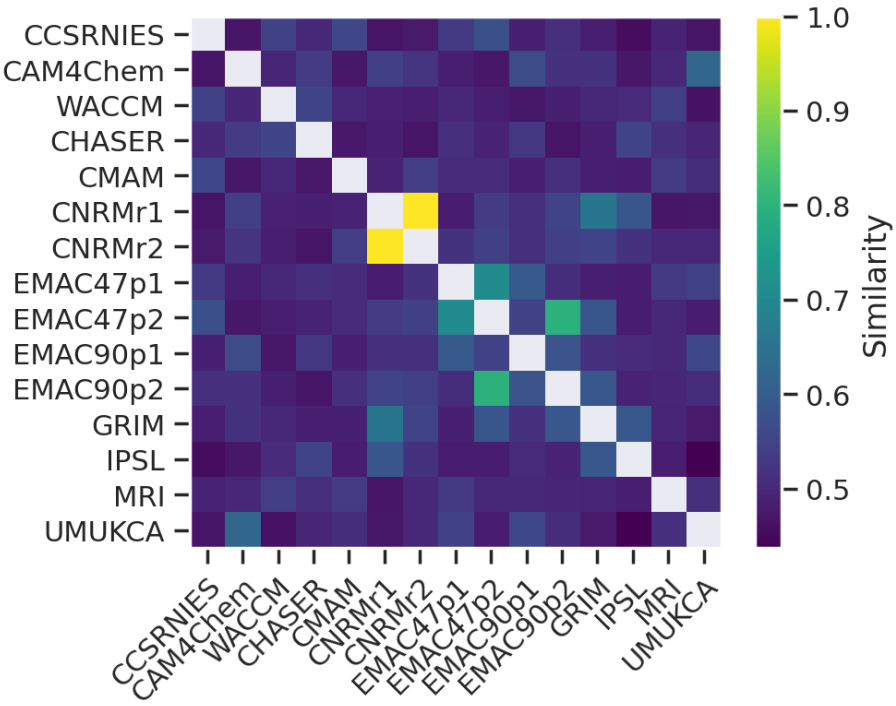

Figure 4.3: Inter-model similarity calculated from the similarity in CCM weights learnt by the BNN. This similarity measure is calculated as the reciprocal of the root mean squared error between normalised CCM pairs. A similarity of 1 indicates these are the most similar models, not that they are identical.

We define a simple similarity metric as the spatiotemporal average of the reciprocal of the root mean squared error (RMSE) between model pairs that have undergone normalisation. This metric increases for more similar model weights and tends to zero for totally dissimilar weights. We normalise the metric by the highest similarity score such that it is in the range [0,1] where a score of 1 represents the most similar model

pair. Figure 4.3 shows the similarity score between model weights for all the CCMI models used in the total ozone column BNN. The most similar models are CNRMr1 and CNRMr2 which are two realisations of the same model. Other high scoring model pairs are EMAC47p1 with EMAC47p2 and EMAC47p2 with EMAC90p2, which are all from the same model but vary in number of atmospheric model levels (47 or 90) and in the nudging scheme (p1 indicates nudging to the global mean temperature whereas p2 does not) (Jöckel et al., 2016). High similarity between these models is expected, as the underlying physical and chemical processes are simulated the same way and so it is reassuring to see that the weights that the BNN prescribes to similar models are likewise similar. Even though the BNN is not tasked with identifying similar models, nor does it directly compare model outputs, model similarity is an emergent feature of the BNN.

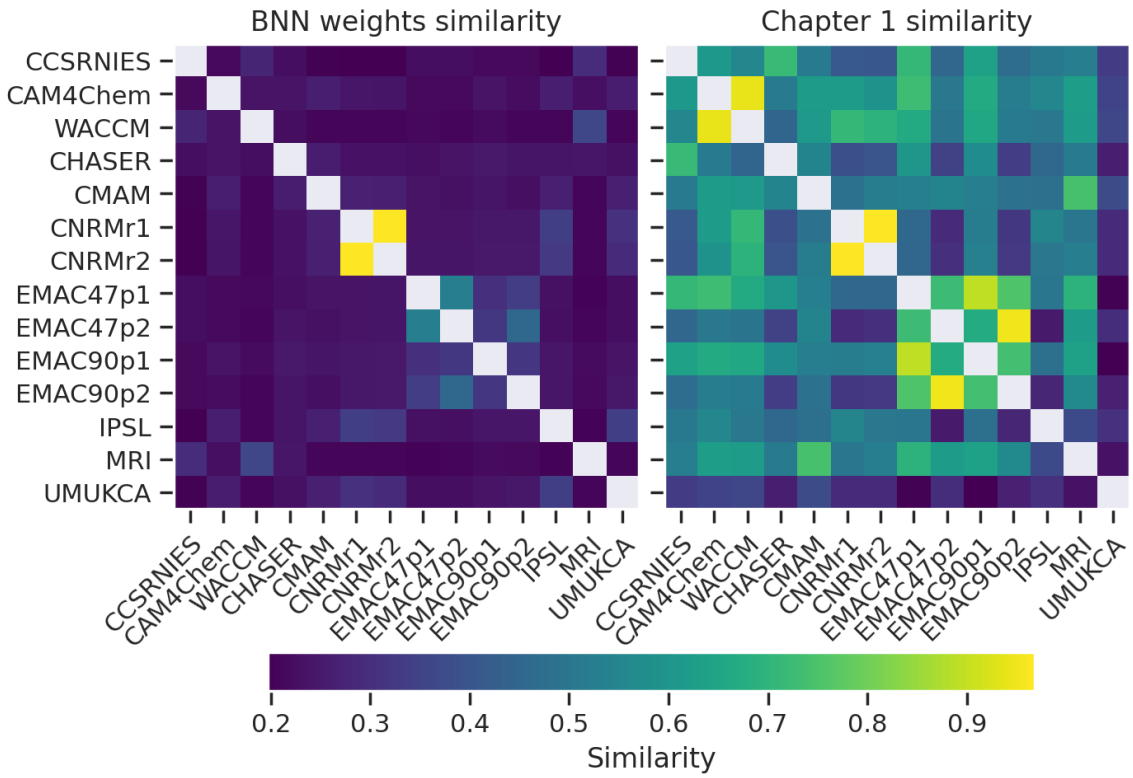

Figure 4.4: Inter-model similarity for the southern polar cap (90°S–60°S) as learnt by the BNN (left) compared to those calculated by the similarity analysis in Chapter 2 (right).

In Chapter 2 we calculated model similarity for CCMI models over the southern polar cap (90°S–60°S) as part of a weighting scheme used to project ozone hole recovery. Figure 4.4 compares the similarity found between BNN model weights with the similarity scores calculated from Chapter 2, with BNN model weights constrained to September, October and November (austral spring) over the southern pole (90°S–60°S) to match that chapter's similarity analysis. Like the global view on similarity in Figure 4.3, known similar models are easily identified as similar (e.g., CNRMr1 and CNRMr2) by both the Chapter 2 analysis and the BNN. However, the similarity from BNN model weights only identifies a few highly similar models whereas the similarity analysis from Chapter 2 highlights many more possible model pairings. One such pairing identified in Chapter 2, but not this analysis, is between CAM4Chem and WACCM which are models built on the same underlying GCM (CESM) (Kay et al., 2015). Though the same circulation model could lead to higher model similarity, in practice however, CAM4Chem is a low-top model meaning that the models are most likely similar only up to the lower stratosphere (Morgenstern et al., 2017).

Discrepancies in identified similarities are likely due to differences in the analysis. For example, the analysis in Chapter 2 only considers the spatially averaged time series of ozone unlike the BNN analysis which is spatially resolved. Additionally, the BNN calculates weights from a non-linear combination of spatiotemporal coordinates whereas the Chapter 2 analysis linearly combines a number of metrics important to stratospheric ozone.

4.2.2 Model weights as a proxy for model performance

Existing model weighting frameworks (e.g., Knutti et al., 2017; Amos et al., 2020) calculate model weights from performance metrics, such as the inverse of the squared difference between model output and observations, that captures an individual model's ability to replicate observations or observed statistics. Instead, the BNN optimises model weights such that it minimises the loss between observations and the ensemble prediction. Therefore, the BNN model weights are not necessarily indicative of performance as the models are not compared directly to observations. The purpose of this section is to determine if some understanding of model performance is retrievable from the BNN model weights.

In Figure 4.5 BNN model weights are compared to a commonly used performance metric, which we call RMSE weight, defined as the reciprocal of the root mean squared error (RMSE) between models and the same total ozone column observations used to train the BNN (Bodeker et al., 2018). There is a weakly significant moderate positive correlation between the BNN derived weights and performance scores (correlation coefficient = 0.40 at 87 % confidence). Although the lowest scoring model (GRIM) also receives the lowest weighting from the BNN, showing some agreement, some high scoring models (WACCM and CCSRNIES) are not similarly weighted highly in the BNN although they still score highly.

The inclusion of multiple highly similar models in the ensemble presents difficulties when analysing model performance as the RMSE weight is calculated independent of other similar model runs, whereas the BNN derived weight is not. This means that if n identical models were included within the BNN we would expect the learnt model weights to be $1/n$ times the value that would be learned if we had included just one of the identical models. This result of identical models being down-weighted is also seen in the synthetic dataset test in subsection 3.3.1. To account for highly similar models within the ensemble, both the family of EMAC models and CNRM models have been averaged to find the average RMSE weights and summed to find the BNN derived weights for EMAC and CNRM. This combination of RMSE and BNN derived weights can be seen in Figure 4.5.

An obvious feature of Figure 3.4 from Chapter 2 is the high weighting the CCSRNIES model of up to 60 % over the southern pole in austral spring, suggestive of the model having a higher predictive skill here. We compare the BNN inferred model weights to the RMSE weight for the average monthly weights for the southern polar cap (90°S–60°S) in Figure 4.6. To avoid an unfair comparison, data were only included in these averages if there were collocated observations, since the BNN derived weights are spatially continuous whereas the model performance metric only includes data where there were observations. In August, September, October and November, CCSRNIES is the most dominant model and also the highest performing model in the ensemble. Outside of these months the contribution to the ensemble from CCSRNIES is much decreased which matches the low values of the performance metric. However, a second model (UMUKCA) with high southern pole weighting, shown in the same figure, does

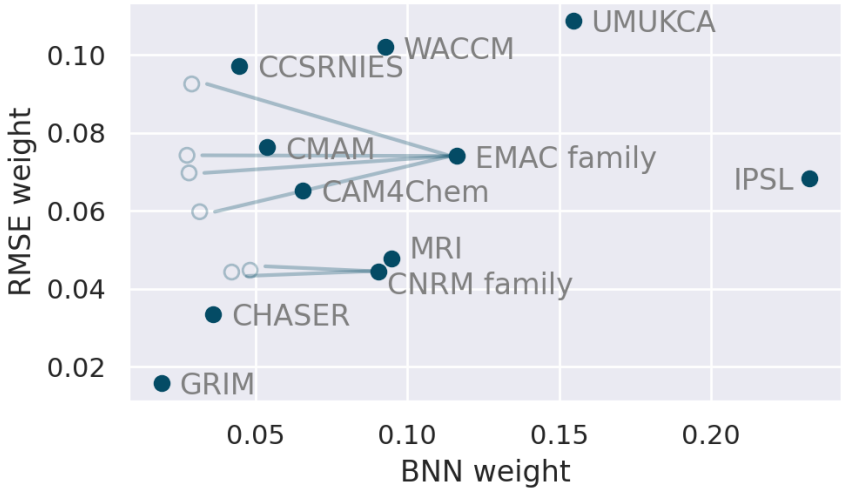

Figure 4.5: Globally averaged model weights for the CCMs learnt by the BNN compared to globally averaged weights derived from a performance metric of the reciprocal of the root mean squared error between the observations and model predictions. Models with highly similar simulations have been grouped to find average weights for each model family (EMAC and CNRM). To find these central point, BNN model weights are summed and RMSE weights are averaged, which are depicted in the plot by the empty circles being combined into a filled circle.

not show a similar relationship between performance and weight, indicating that a rationalisation of model performance given the BNN derived weights is not always possible.

4.2.3 Model weight discussion

Quantifying model performance and similarity is complex. It depends on the quality and availability of the comparative observations, the chosen metric or score, and whether we consider one or multiple model outputs. For these reasons, the above analysis comparing weights to a performance score does not imply that the BNN derived weights are not indicative of model performance, rather that they are not indicative of model performance as measured by the reciprocal of the root mean squared error. Instead, the BNN model weights inform us which models provide the most utility to the weighted ensemble prediction. As the BNN is a complex weighting framework, where the weight of a model is determined by a highly nonlinear combination of

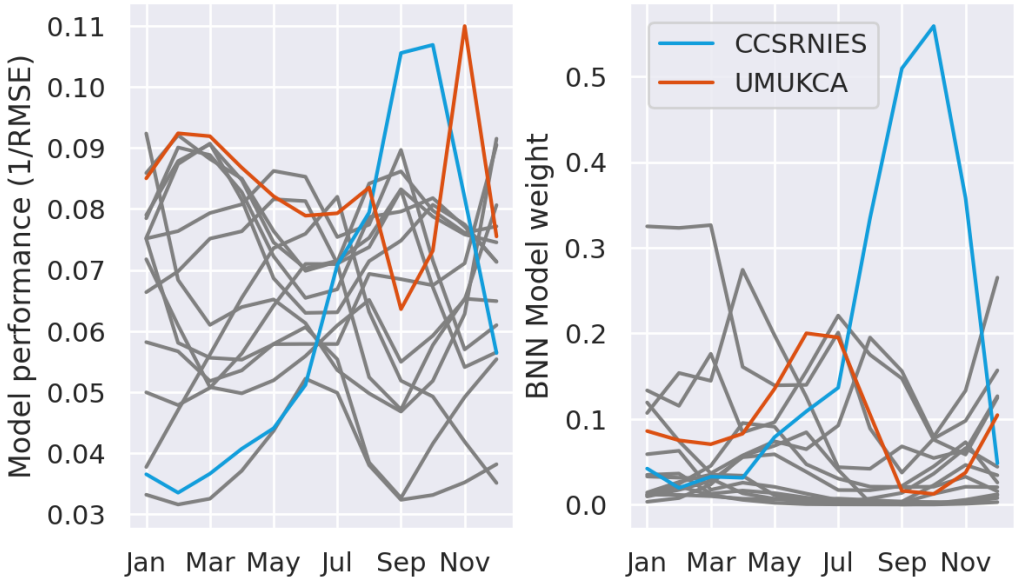

Figure 4.6: Seasonal model performance (left), calculated as 1/RMSE, and seasonal BNN model weight (right) averaged over the southern polar cap (90°S–60°S) for the 15 CCMs. Two models are shown in blue or orange and the other models are shown in grey.

the input coordinates, it is not surprising that there is disagreement between the model weights and a comparatively simple performance measure. The same is true when considering model similarity. Although it is reassuring to recover known model similarities from the BNN derived weights, quantifying true model similarity is highly complex and calculating a simple similarity metric only captures a small part of this. Further confounding the interpretation of BNN model weights is that they are strongly linked to the availability of observational data. The continuous model weights are the BNN's extrapolated and interpolated model weights that, for sparse data, might not be representative of underlying model performance. For these reasons it is suggested that the BNN should not be used as a tool to investigate model performance.

4.3 Model bias

The model bias (β) is the small correction applied to the linear combination of weighted models to rectify any offset between the ensembled model prediction and observations.

Like the model weights, the bias is a function of space and time, encompassing our prior knowledge that model performance varies spatiotemporally. In this section we explore model bias as calculated by the BNN, testing if it corresponds to known model biases in CCMs or observations.

Figure 4.7 shows the BNN bias averaged spatially and temporally. Spatially the bias exhibits very clear boundaries between positive and negative bias, particularly at a constant latitude of about 35°S. Poleward of 35°S and very high northern latitudes have an average negative bias correction. All other regions have positive bias which is especially high in the tropics. This means that the linear weighted combination of models, determined by the BNN, over-predicts in polar regions and under-predicts elsewhere.

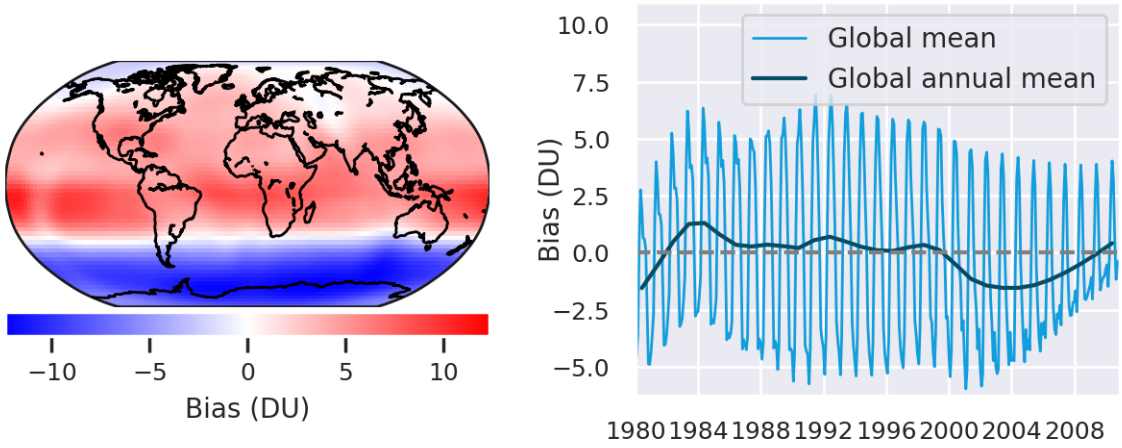

Figure 4.7: BNN bias in Dobson units shown as a temporal average (left) and a spatial average (right). The right panel shows the global average and the global annual average as described in the legend, and a grey dashed line demarcating zero.

Temporally, the bias is very oscillatory with a period of a year showing that the bias has strong seasonal dependence, although the seasonal amplitude decreases after about 2002. When annually averaged the bias is small and approximately zero centred. The seasonality of the BNN bias is shown in Figure 4.8 alongside the seasonal bias from the CCMs, calculated as the observations used to train the BNN (Bodeker et al., 2018) minus model prediction. All CCMs exhibit the same strong seasonal bias which peaks in boreal spring/summer, much like the bias learnt by the BNN. Through the

training process, with the aim of optimally combining models, the BNN has successful learnt a bias which not only improves predictions but also is representative of the biases in the underlying models it is ensembling.

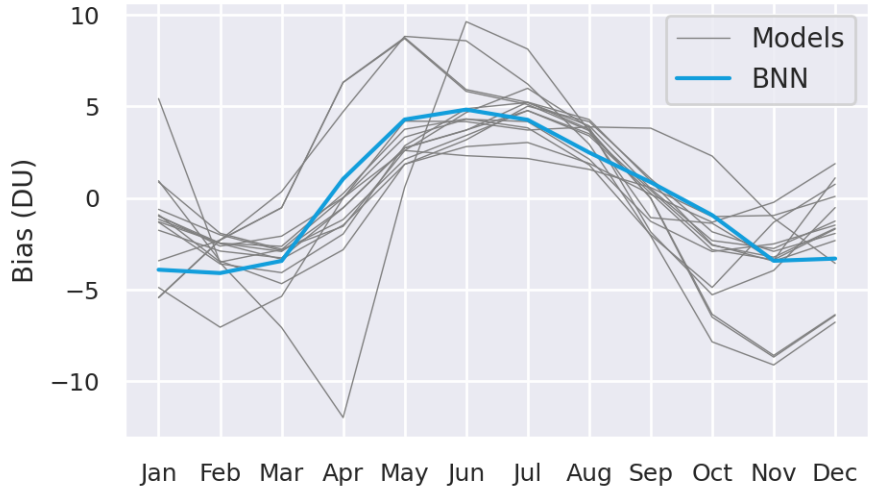

Figure 4.8: Average seasonality of bias, predicted and corrected for by the BNN (blue), and for each CCM (grey), calculated as model output minus observation. To show the biases on a comparable scale each model bias is shown relative to its average bias, hence for this plot the average bias for each model equals zero.

4.4 Model uncertainty

The total uncertainty from the BNN prediction is formed from a combination of two uncertainties: an aleatoric uncertainty (σ), which is an estimate made by the BNN of the uncertainty in the observational data; and an epistemic uncertainty which is caused by a lack of data, such as in the data sparse polar regions. Chapter 3 showed the skill with which the BNN framework accurately determined both uncertainties. As with the model weights and bias, these uncertainty terms are dependent on space and time. This section explores the BNN uncertainty and considers how it provides insight into the variable quality of observations and informs where prediction accuracy is reduced due to data deficiencies.

4.4.1 Aleatoric uncertainty - data uncertainty

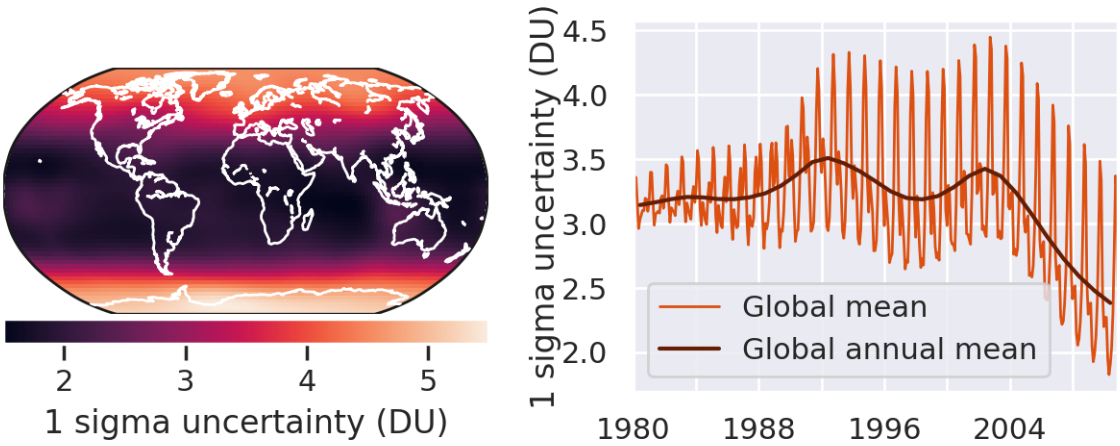

Figure 4.9: Aleatoric uncertainty as predicted by the BNN averaged temporally (left) and spatially (right). The uncertainty is plotted at the 1 sigma level which encompass approximately 68 % of the data. Global mean time series are shown in the right plot for the monthly time series and an annual mean.

The aleatoric uncertainty, or observational noise (σ in equation 4.1) is designed to account for the imprecision of observations and uncertainties that arose in the assimilation of multiple observational sources into the BSTier0 dataset. Figure 4.9 shows how the estimated observational noise varies in space and time. Uncertainty is maximum over polar regions, decreases towards to tropics and has an average value of 3.2 DU, relating to a fractional uncertainty of approximately 1 %. It is very latitudinally dependent but exhibits little longitudinal dependence. Temporally, uncertainty has a strong decrease post-2004 which coincides with the AuraMLS satellite becoming operational (Waters et al., 2006), a satellite product which "performs exceptionally well in most regions" (Tegtmeier et al., 2013). The high uniform coverage of AuraMLS and its improved accuracy (5 % or less throughout the stratosphere (Froidevaux et al., 2008)) has previously been shown to improve the accuracy of other total ozone column datasets (McPeters and Labow, 2012), as it also does here. We further note that these findings are in agreement with analysis summarised in Figure 5.3.

The observational uncertainty has strong seasonal dependence, peaking during both boreal and austral winter/spring periods. As can be seen in Figure 4.10, these peaks

are caused by increased uncertainty over the polar regions during the winter and spring periods, whereas the uncertainty contribution from 60°S–60°N is comparatively small. Winter and spring in polar regions are periods of complex dynamics, including the polar vortex and its subsequent breakdown in spring, which affects ozone distribution and concentrations (Schoeberl and Hartmann, 1991; Butchart, 2014). These dynamics are difficult to simulate accurately, with many models predicting delayed polar vortex breakdowns, which is further confounded by differences in modelling responses and feedbacks between ozone depletion and stratospheric cooling (Lin et al., 2017). Additionally, CCMs have historically suffered from a 'cold pole problem' (Austin et al., 2003), which is particularly prevalent in the southern hemisphere during winter and spring and is still exhibited in UMUKCA (Dennison et al., 2019). The increased aleatoric uncertainty during these periods is most likely a response to the larger differences between models and observations driven by the complex dynamics and coupling between the chemistry-climate system, which increases the uncertainty estimated by the BNN.

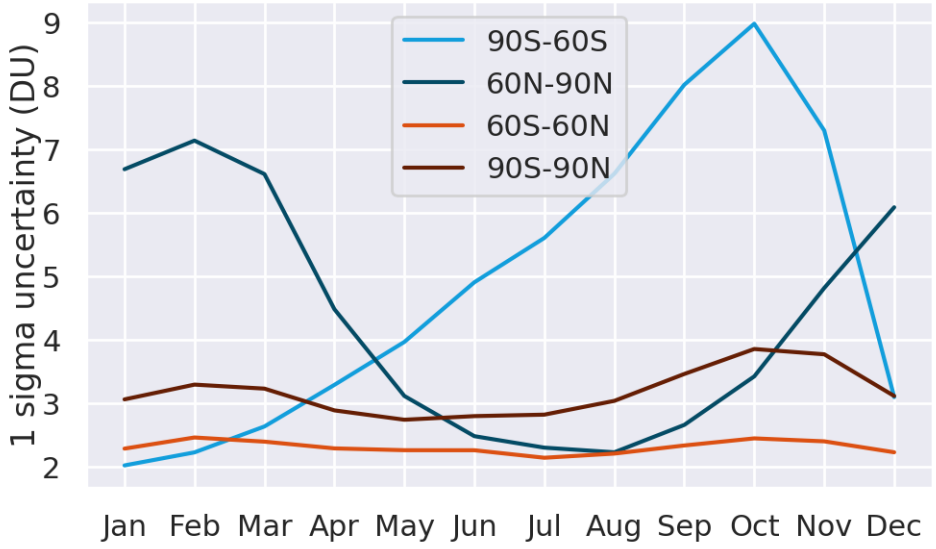

Figure 4.10: Mean aleatoric uncertainty estimated by the BNN shown at a 1 sigma level (approximately a 68 % confidence interval) for both polar regions (90°S–60°S and 60°N–90°N), near-global (60°S–60°N) and global (90°S–90°N).

4.4.2 Epistemic uncertainty - statistical model uncertainty

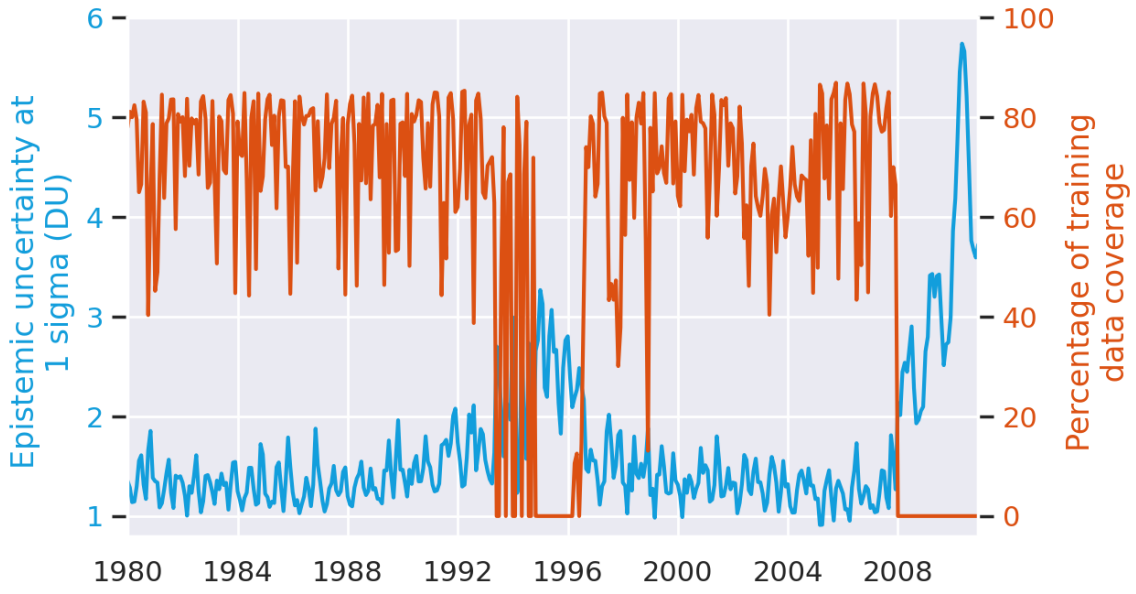

Figure 4.11: Spatially averaged epistemic uncertainty (blue) compared to the percentage of data coverage within the training dataset (orange). There is zero data coverage for the years 2008, 2009 and 2010 because these years were removed from the training dataset for testing purposes.

4.4.3 Epistemic uncertainty - statistical model uncertainty

Epistemic uncertainty is a measure of the reduction in predictive ability driven by a lack of data availability or information. In a BNN epistemic uncertainty arises because each of the BNN's individual neural networks produce different predictions, which vary more from each other in regions of sparse data where the predictions are less strongly constrained. The inverse relationship between data availability and epistemic uncertainty can be clearly seen in Figure 4.11, which shows time series of globally averaged epistemic uncertainty and data coverage. In temporal periods of zero data coverage the epistemic uncertainty quickly increases: 1994–1996, where no observations were available, and 2008 onward which was removed for testing. This is extremely desirable behaviour as we should not expect the BNN to know what the model weights

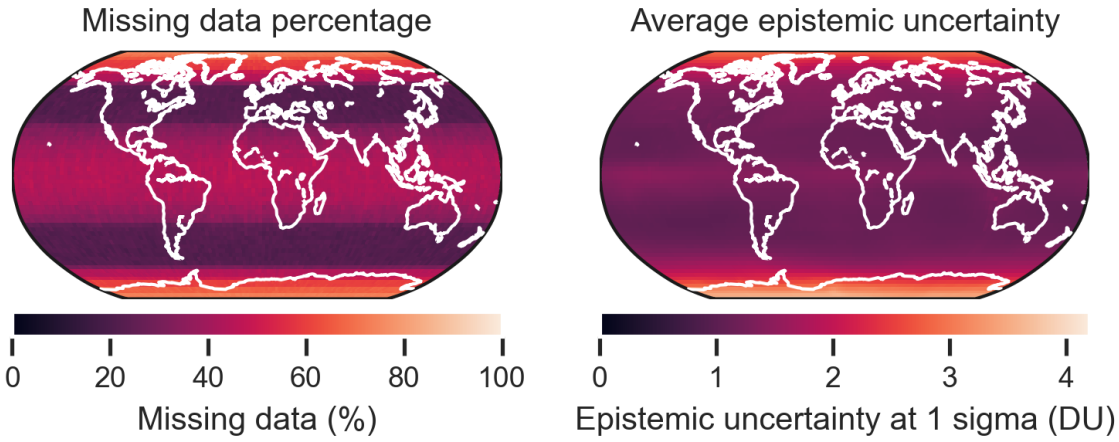

Figure 4.12: Temporally averaged epistemic uncertainty (right) compared to the percentage of data coverage within the training dataset (left).

and bias, and therefore the overall ozone prediction, should be in regions which we have little to no observations. This behaviour is also exhibited spatially in 4.12, resulting in higher epistemic uncertainty over the polar regions, due to their increased sparsity from the absence of observations in polar night, and the tropics.

4.5 Discussion

One of the main strengths of the BNN is that it breaks down predictions of a variable or state into a more regularly understood notion of ensembling: a linear combination of weighted models. The BNN models for every point in time and space, providing a useful resource into investigating and interpreting model contribution. However, model performance and similarity inferred by the BNN are only qualitatively comparable with other performance and similarity metrics. This is in part due to the difficulty in, and lack of agreed approach for, assessing model performance (Katzav et al., 2012) and similarity (Masson and Knutti, 2011). Just because a model output matches the observations does not mean that it is simulating the entire system well, or that it will do in the future.

Despite a lack of rigorous quantification of model performance and similarity, using simple metrics does reveal some ability of the BNN to recover known model

similarities. Although the BNN weights do not match directly represent measures of model performance, the BNN still accounts for variable model performance through the spatiotemporal variability of weights, which greatly improve the accuracy of predictions. In addition to improving upon traditional weighting methods (e.g., Räisänen et al., 2010; Knutti et al., 2017; Brunner et al., 2019), the interpretability of the BNN is an important but not exclusive feature. Interpretability is incorporated into the framework by modelling observations as a linear combination of physical models. Although other ML methods such as Gaussian processes or vanilla neural networks could be applied to this ensembling challenge, none scale efficiently to large datasets whilst also quantifying uncertainty appropriately.

Alongside insight from model weights, the bias and both uncertainty terms provide useful insight into observational quality and the predictive accuracy of the BNN. With this information we could select the geographical placement of new stations or instrumentation that best improves the predictions. The weights, bias and uncertainty, not only help dissect the BNN prediction, but also build trust in this deep learning ensembling method, hopefully encouraging wide community use.

The analysis in this chapter and the previous one emphasise the utility of the BNN as a tool to ensemble models by leveraging the spatiotemporal variability of model performance and the BNN's capabilities in estimating uncertainties. The BNN produces accurate predictions that consider observational quality and availability, through a framework in which model contributions and uncertainties can be analysed. Although this chapter has shown links between BNN derived model behaviours and those derived from traditional analysis, such as performance and similarity, the BNN is designed and excels as a model ensembling tool rather than one designed for ensemble analysis.

Chapter 5

A continuous vertically resolved ozone dataset from the fusion of chemistry climate models with observations using a Bayesian neural network

Matt Amos[1], Ushnish Sengupta[2], Paul J. Young[1], J. Scott Hosking[3].

[1]Lancaster University, Lancaster, UK
[2]University of Cambridge, Cambridge, UK
[3]British Antarctic Survey, Cambridge, UK

The following work will soon be submitted to Earth System Science Data. Full code and output can be found at `https://github.com/mattramos/VertOzone-BNN`. The author contributions are listed below.

Statement of contribution
Matt Amos developed the methodology with intellectual contributions from **Ushnish Sengupta**. **Matt Amos** performed all the data processing, method development

and analyses, and also and wrote the manuscript. All the coauthors commented on analyses and helped with finalising the manuscript.

Abstract

Continuous historic datasets of vertically resolved stratospheric ozone, support the case for ozone recovery, are necessary for the running of offline models and increase understanding of the impacts of ozone on the wider atmospheric system. Vertically resolved ozone datasets are typically constructed from multiple satellite, sonde and ground-based measurements that do not provide continuous coverage. As a result, several methods have been used to infill these gaps, most commonly relying on regression against observed time series. However, these existing methods either provide low accuracy infilling especially over polar regions, unphysical extrapolation, or an incomplete estimation of uncertainty. To address these methodological shortcomings we used and further developed an infilling framework that fuses observations with output from an ensemble of chemistry-climate models within a Bayesian neural network. We used this deep learning framework to produce a continuous record of vertically resolved ozone with uncertainty estimates. Under rigorous testing the infilling framework extrapolated and interpolated skillfully and maintained realistic interannual variability due to the inclusion of physically and chemically realistic models. This framework and the ozone dataset it produced, enables a more thorough investigation of vertically resolved trends throughout the atmosphere.

5.1 Introduction

Ozone in the upper troposphere and stratosphere is monitored predominantly by ozonesondes (e.g., Witte et al., 2017), satellites (e.g., Tegtmeier et al., 2013) and ground based instruments (e.g., Fioletov et al., 2008). Vertically resolved ozone datasets combining measurements from multiple instruments (e.g., Miller et al., 2002; Ball et al., 2017) are regularly used to investigate ozone trends throughout the atmosphere, including the monitoring of ozone recovery in accordance with the Montreal protocol (Chipperfield et al., 2017; Ball et al., 2018; WMO, 2018), and processes such as the impacts of changing ozone concentrations on the atmospheric system, the Brewer Dobson circulation (Polvani et al., 2018) and surface climate (Ivy et al., 2017). In determining ozone recovery it is especially useful to use vertically resolved datasets rather than the total ozone column, as the latter can show ozone recovery when in fact depletion in the lower stratosphere might be disguised by increasing tropospheric ozone (Ball et al., 2018; Gaudel et al., 2018).

Gaps are present in these datasets, originating from gaps in the records of individual instruments, due to the finite lifetimes of instruments, difficulty of measuring during polar night (Bowman and Krueger, 1985; Randel and Wu, 1999a) as well as funding concerns, leading to an incomplete picture of atmospheric ozone that is particularly prevalent over polar regions (Bodeker et al., 2021). Several statistical infilling methods have been developed to address these observational gaps creating spatially and temporally complete ozone datasets (e.g., Bodeker et al., 2013; Davis et al., 2016) for use as continuous climatologies and for the running of climate models which do not simulate chemistry. These data-driven statistical infilling approaches are comparatively simple and are not well suited to infilling regions of sparse observations because they lack the physical and chemical understanding of the system in the way that chemistry-climate models do not. Here, we describe and demonstrate a new methodology to produce a spatially and temporally continuous vertically resolved ozone dataset. We use a Bayesian neural network as described by Sengupta et al. (2020) to fuse together our best physical and chemical understanding of the chemistry-climate system, represented by chemistry-climate models, with discontinuous observations to construct an assimilated product of vertically resolved ozone.

Continuous ozone datasets, such as those of Randel and Wu (2007), Bodeker

Scientific (Bodeker et al., 2013) and SWOOSH (Davis et al., 2016) are generated from statistical models fitted to observations. These observational datasets themselves are constructed from multiple satellite observations, accounting for satellite drift and applying bias corrections, often further supplemented by the addition of ozonesondes or ground based measurements. Satellite measurements of ozone are typically from instruments such as the total ozone mapping spectrometer (TOMS) which measures total column ozone from backscattered UV radiation (Heath et al., 1975), or limb sounders that measure across an atmospheric section providing vertically resolved ozone concentrations (Froidevaux et al., 2008). These satellite measurements are often limited in their coverage, both spatially, particularly over the poles as solar occultation limb sounders require sunlight absent in polar winters, and temporally, by the finite lifetimes of the satellites and their orbits.

Understanding and attributing ozone change and variability is aided by a continuous ozone record that is resolved in time, latitude and height. In addition to the benefits of historic ozone records in monitoring ozone depletion and recovery, ozone records and climatologies are used as offline fields in a variety of climate models that do not compute their own interactive chemistry (Cionni et al., 2011), or for chemical transport models. To this end, there have been several datasets that have produced continuous vertically resolved ozone records by leveraging a variety of infilling techniques. In the following paragraphs we describe infilling techniques used in several of these ozone products, that takes place after the pre-processing, gridding and merging of individual satellite outputs into a single product. These infilling methods represent standard data approaches, which use observational data and supplementary datasets to infer, with some interpolation or regression tool, what the missing values should be.

Randel and Wu (2007) and Bodeker Scientific (Bodeker et al., 2013) (BSTier1.4) created infilled ozone records by using multi-linear regression to model ozone as a sum of global time series such as the quasi-biennial oscillation and the solar cycle expanded into harmonic components (WMO, 2018). Bodeker et al. (2013) additionally expanded the coefficients of the regression model as Legendre polynomials which implemented latitudinal structure to aid spatial infilling. The advantages of infilling based on regression against observed time series is that the infilling is partially grounded in our observations and knowledge of the physical system. The limitations of such an

approach are caused by the linearity of the regression model smoothing the ozone field resulting in an underestimate of interannual variability. Additionally, as with existing filled vertically resolved ozone datasets, sources of uncertainty from both data and infilling method are not fully considered.

In the SWOOSH dataset (Davis et al., 2016), two infilling methods are used. Firstly, gaps near the poles (on regular latitude) are filled by taking data from ozone gridded on an equivalent latitude grid and imputing this data. This method likely underestimates ozone for regions within the polar vortex as any given equivalent latitude will be less than the corresponding geographic latitude. Secondly, data is interpolated for each vertical level using a radial basis function with an inverse multi-quadric function that imputes the mean of the surrounding points adjusted to preference closer points. Where the data is unbounded, such as the poles, interpolation is performed between existing data and the climatological average. This method of interpolation will perform less well over large regions of sparse data and as such the authors comment that the pre-1990 section of the filled dataset should be used with caution. Additionally, bounding the interpolation with the climatology risks not capturing the true trend of the data and as a result the authors recommend not using the filled version for trend analysis.

Another common method used for infilling and data assimilation in atmospheric and climate science is 4-dimensional variational assimilation (4D-Var) (Courtier et al., 1994). This is used in numerical weather prediction to produce historical reanalyses (e.g., Rienecker et al., 2011; Hersbach et al., 2020) where observations are assimilated within the physical constraints of a numerical model. 4D-Var reanalyses are commonly used across environmental disciplines (e.g., Viste et al., 2013; Blunden and Arndt, 2016; Amos et al., 2020) but the method does have limitations. Firstly, it is extremely computationally expensive as the assimilation involves running a numerical model and optimising the model against millions of observations. Secondly, whereas climate projections are typically computed from an ensemble of different models (allowing for variable model performance within the ensemble), in 4D-var the observations are assimilated to a single model which may have inherent biases.

Hybrid modelling approaches, statistical or machine learning models with physical models, are beginning to find application in data infilling for atmospheric composition.

One such approach is bias correcting a physical model with a machine learning algorithm. Dhomse et al. (2021) learn and correct for the bias in a chemical transport model (TOMCAT) when compared to observations. Similarly to 4D-Var, this method is reliant on only a single model, which might introduce biases which would be accounted for by considering an ensemble of models. Furthermore, this approach does not attempt to quantify the uncertainty of estimates, either from the observations or the model which learn the model bias, in this case a random forest.

The described infilling methods have their strengths and weaknesses. BSTier1.4 and 4D-Var predictions benefit from an infilling approach grounded in the physical and chemical understanding of the system. However, for BSTier1.4 this is through a simple linear regression, which cannot capture full variability, and for 4D-Var only a single model with high computational expense is used. The infilling from BSTier1.4 and SWOOSH are suitably quick to run but their simplicity results in less accurate infilling, particularly when extrapolating and infilling large regions. Additionally, not all methods consider the complete uncertainty in the infilled predictions.

In this paper we describe a fundamentally different approach to infill historic ozone to produce a continuous vertically resolved record by using archived chemistry-climate models (CCMs), which grounds the infilling method to our understanding of the physical system. These CCMs are weighted, combined and bias corrected within a Bayesian neural network which accurately infills missing data and alongside providing a principled quantification of uncertainty. This approach is also computationally cheap compared to 4D-Var and, despite being a deep learning algorithm, maintains interpretability such that the predictions can be dissected back to individual model contributions. Section 5.2 introduces the CCMs, ozone observations and the necessary pre-processing steps to prepare the data for the Bayesian neural network, which is described in section 5.3, alongside the process of testing and validation. In section 5.4 we explore the trends and coverage of the infilled ozone dataset and compare it to existing infilled products. Finally, we make concluding remarks in section 5.5.

5.2 Ozone and model data

Our approach to infilling vertically resolved ozone combines observations from the Bodeker Scientific observational dataset (Bodeker et al., 2013) with an ensemble of chemistry climate models from the Chemistry-Climate Model Initiative (CCMI) (Morgenstern et al., 2017). In this section we summarise the characteristics of these data and detail how we processed them in order to build a machine learning-style dataset to train the Bayesian neural network.

5.2.1 Bodeker Scientific vertically resolved ozone

We used the tier 0 data of vertically resolved ozone from Bodeker scientific (Bodeker et al., 2013) (hereafter referred to as BSTier0) as observations. This dataset combines merged measurements from ozonesondes and satellites onto a coarse resolution grid, following quality checking and screening to remove anomalous values as is described by Bodeker et al. (2013). From BSTier0 we extracted a vertical pressure range of {500–0.3 hPa} and screened values that were less than 1 ppb as the original dataset had negative concentrations. This cut-off concentration was chosen because, for the upper troposphere and stratosphere region of interest, concentrations have not been measured to fall below the order of magnitude of 1 ppb even in the ozone deficient tropical tropopause (Newton et al., 2018).

5.2.2 CCMI model output

The CCMI data (Morgenstern et al., 2017) consists of an ensemble of chemistry-climate models (CCMs) that include a detailed description of atmospheric chemistry to better understand the behaviour of compounds whose abundances depend on chemical processes. Similarly to its predecessor initiatives (Eyring et al., 2008; Lamarque et al., 2013b), a focus of CCMI is exploring and understanding stratospheric ozone distributions (Morgenstern et al., 2018) and projections (Amos et al., 2020), making this ensemble highly appropriate for our uses.

We used output from the so-called refC1SD scenario (Morgenstern et al., 2018), which are specified dynamics simulations where CCMs are nudged to historic (1980–2010) meteorological conditions (Orbe et al., 2020). Although the nudging was

performed differently across the ensemble (Orbe et al., 2020), the simulations represent the models' best efforts at replicating past meteorology and past atmospheric composition. This allows for a straightforward direct comparison between models and observations unlike if the models were free running, in which case we would have to compare trends and climate statistics. From the refC1SD ensemble, we selected 13 simulations of monthly average vertically resolved ozone concentrations from the 9 models that cover the vertical pressure range {500–0.3 hPa}. The output was latitudinally regridded to match the resolution of the observations (5°) and were longitudinally averaged, producing a zonal mean.

5.2.3 Making the data machine learning ready

We performed coordinate mappings and data scaling to both the BSTier0 observations and model output. These steps, common in machine learning (ML) applications, ensure that the data is descriptive of the spatiotemporal domain and allow us to encode our prior understanding of the system into machine learning (ML) models.

After pre-processing, the ozone data both modelled and observed is described by time, latitude and a vertical coordinate. The vertical coordinate air pressure p was mapped using a natural logarithm to $\ln(p)$ to provide a consistently decreasing vertical coordinate which is proportional to height. Latitude (θ), the second spatial coordinate, is mapped to $\sin(\theta)$. We mapped time t (in this case the number of months since the start of the observations) onto seasonal harmonic terms ($\cos(2\pi t/12)$ and $\sin(2\pi t/12)$) as well as keeping t as a continuous time coordinate. The seasonal mapping requires two elements to ensure continuity over the first and last months so that the data reflects that January and December are adjacent months.

Subsequent to the mapping, the coordinates were min-max scaled to between -1 and 1, which ensures that all coordinates are treated as equally important. Additionally, the model predictions and observations of ozone concentrations were scaled with a natural logarithm, before being min-max scaled such that the observations are between -1 and 1. The motivation for the log scaling is twofold. Firstly, the log transformation means that ML tools can learn evenly across a domain where ozone concentrations span multiple orders of magnitude. Without performing this log scaling, the learning efforts would be focussed in the regions of larger concentrations (~ 1 ppm) because

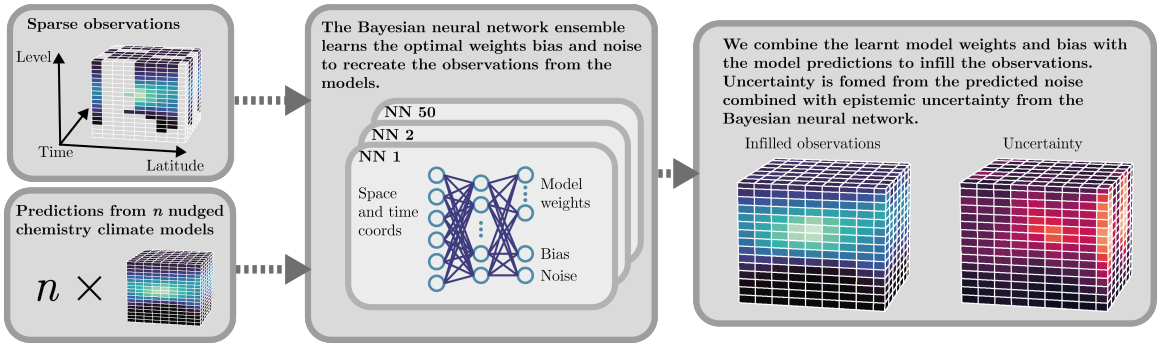

Figure 5.1: **Depiction of the Bayesian neural network.** It shows the overall framework of combining gridded observations with CCMs using the BNN to produce infilled predictions alongside uncertainties. The middle box also contains a simplified diagram of the neural network architecture.

these would be the regions in which an ML model could more easily minimise the absolute error (the difference between the observations and the prediction). Secondly, as negative concentrations are unrealistic, the log scaling constrains the prediction to be strictly positive.

To test the predictive capability of the Bayesian neural network, particularly its ability to recover realistic uncertainties, we split the data into a training (365,000 datapoints) and three testing datasets. The testing datasets are designed to test interpolation over small and large scales, and extrapolation. They are the following: 1) interpolation-testing covering all data over a 1.5 year period between July 1997 and December 1998 inclusive (25,000 datapoints); 2) extrapolation-testing covering all data in the last 1.5 years of the total dataset covering July 2009 to December 2010 (30,000 datapoints); and 3) random-testing which is a randomised sample of 5 % of the data not included in the other testing datasets (20,000k datapoints). The Bayesian neural network is trained only on the training dataset and the three testing sets are held out for validating predictive ability and uncertainty quantification.

5.3 Methods

5.3.1 Bayesian neural networks for fusing CCMs and observations

The Bayesian neural network we used to fuse CCMs and observations together is an extension of our earlier work (Sengupta et al., 2020), and is summarised in Figure 5.1. We took a similar approach to the task of infilling the sparsely sampled ozone observations (*obs*) by assuming that they can be modelled as a linear combination of n nudged CCM outputs (M_i), adjusted by weights (α_i), a bias (β) and heteroscedastic noise (σ):

$$obs(\mathbf{x},t) = \sum_{i=1}^{n} (\alpha_i(\mathbf{x},t) M_i(\mathbf{x},t)) + \beta(\mathbf{x},t) + \sigma(\mathbf{x},t). \tag{5.1}$$

These weights, bias and noise are modelled as spatially ($\mathbf{x}$) and temporally (t) varying, allowing for both the fact that CCMs may better simulate certain spatial regions or certain seasons and that observational noise is not constant. We used the Bayesian neural network (BNN) to learn the optimal weights, bias and noise.

In general, BNNs are neural networks (NNs) over whose parameters a prior is specified and whose posterior distribution can be recovered by a variety of inference techniques (Neal, 2012). In this paper, we used an anchored ensemble of individual NNs to approximate Bayesian inference (Pearce et al., 2020). The BNN learns weights, bias and noise from the spatiotemporal inputs such that, when combined with model predictions, it can successfully replicate the observations over a training period within a specified level of uncertainty. Model weighting is a regression task for which there are numerous capable machine learning models. We justify the choice of a NN approach for reasons of scalability, since gridded model data, even at coarse resolutions, quickly becomes large. A detailed description of the statistical underpinnings of the BNN is presented by Sengupta et al. (2020), including the relevant derivations and a complete discussion of the BNN design. Here we summarise the main concepts of the BNN for combining observations and geophysical models.

The probabilistic nature of the BNN allows us to encode our prior knowledge within it and quantify uncertainty. Our prior for the model weights (α) is that for an

individual NN, at any point in time and space, any combination of models should be equally likely. Averaged over all NNs, this prior belief means the untrained prediction becomes the multi model mean, which is an appropriate starting point for model ensembles (e.g., Reichler and Kim, 2008; Knutti et al., 2010). Other priors are that the bias (β) should be small and zero-centred, because the models should predominantly contribute to the prediction, and that the noise (σ) should similarly be small and positive as the noise should not be so large that it inhibits finding optimal weights and bias. Overall, these priors mean that the output from an untrained BNN will be the multi model mean with a small bias, small noise and a large epistemic uncertainty which spans the range of predictions made by the different CCMs in the ensemble.

To encode our prior knowledge within a NN we have to determine what values the parameters of the NN can take in order to produce our prior distribution. NNs comprise of layers of neurons which have numerical weights and biases and it is these we encode the prior within. There are many combinations of values that the NN layer weights and biases can take to achieve this, meaning that our prior is actually a distribution over neural network parameters.

We encode the prior into the BNN by anchoring the trainable parameters (the NN weights and biases) of each NN ensemble member to a random draw from this prior distribution. This anchoring is a form of regularisation that ensures diversity in the NN predictions in accordance with our prior, from which we can quantify an uncertainty in the learnt weights and bias. Increased disagreement across the BNN in regions with no observations, due to the anchoring, means that we are suitably uncertain about the values of the weight of a CCM, the bias or the noise during extrapolation. This of course is a highly desirable characteristic. After training, individual NNs are sampled from an approximate posterior distribution and therefore, given an adequate number of anchored NNs, we can reconstruct an estimate of the posterior distribution. The BNN prediction is computed by taking the mean of the individual NN predictions and the epistemic uncertainty of the prediction is calculated as the standard deviation across the individual NN predictions. The epistemic uncertainty represents the uncertainty of the BNN. The total predictive uncertainty σ_{TOT}, combining noise predictions with epistemic uncertainty, is calculated as follows:

$$\sigma_{\mathrm{TOT}} = \sqrt{\frac{1}{n_e}\sum_{j=1}^{n_e}\sigma_j^2 + \frac{1}{n_e}\sum_{j=1}^{n_e}y_j^2 - \left(\frac{1}{n_e}\sum_{j=1}^{n_e}y_j\right)^2}, \tag{5.2}$$

where n_e is the number of NN ensemble members, σ_j is the predicted noise for each NN ensemble member, and y_j is the prediction for each NN ensemble member.

5.3.2 BNN training

The BNN trained for the purpose of infilling the vertically resolved ozone consists of an ensemble of 48 NNs, an ensemble large enough to ensure the posterior distribution was suitably resolved. The 48 NNs were independently initialised and anchored according to our priors before conducting prior predictive checks to ensure the untrained BNN output reflected our priors (the prediction of the multi model mean with small noise and small bias). The BNN was trained on the training dataset for 100,000 epochs on a cluster of 4 T4 GPUs, taking 10 hours to complete and costing about £20 in cloud computing credits (as of April 2021). The full training details are provided in the Supplementary Materials (Appendix B).

5.4 The BNN ozone dataset (BNNOz)

The trained BNN optimally fuses the 13 CCM predictions with the sparse BSTier0 observations with full consideration of uncertainty, producing a continuous vertically resolved ozone and uncertainty prediction, spanning {500–0.3 hPa} for the years 1980–2010. A subset of these infilled results is shown in Figure 5.2, highlighting the smoothness of the prediction and the variable nature of the predictive uncertainty.

5.4.1 Testing and validation

We assess the performance of the BNN ozone prediction (hereafter BNNOz) using the root mean squared error (RMSE) between BNNOz and the BSTier0 observations that were not part of the training dataset, as well as the fractional error, which calculates the absolute error relative to the true observational value. Fractional error is used in addition to the RMSE because the latter is dominated by regions of high ozone

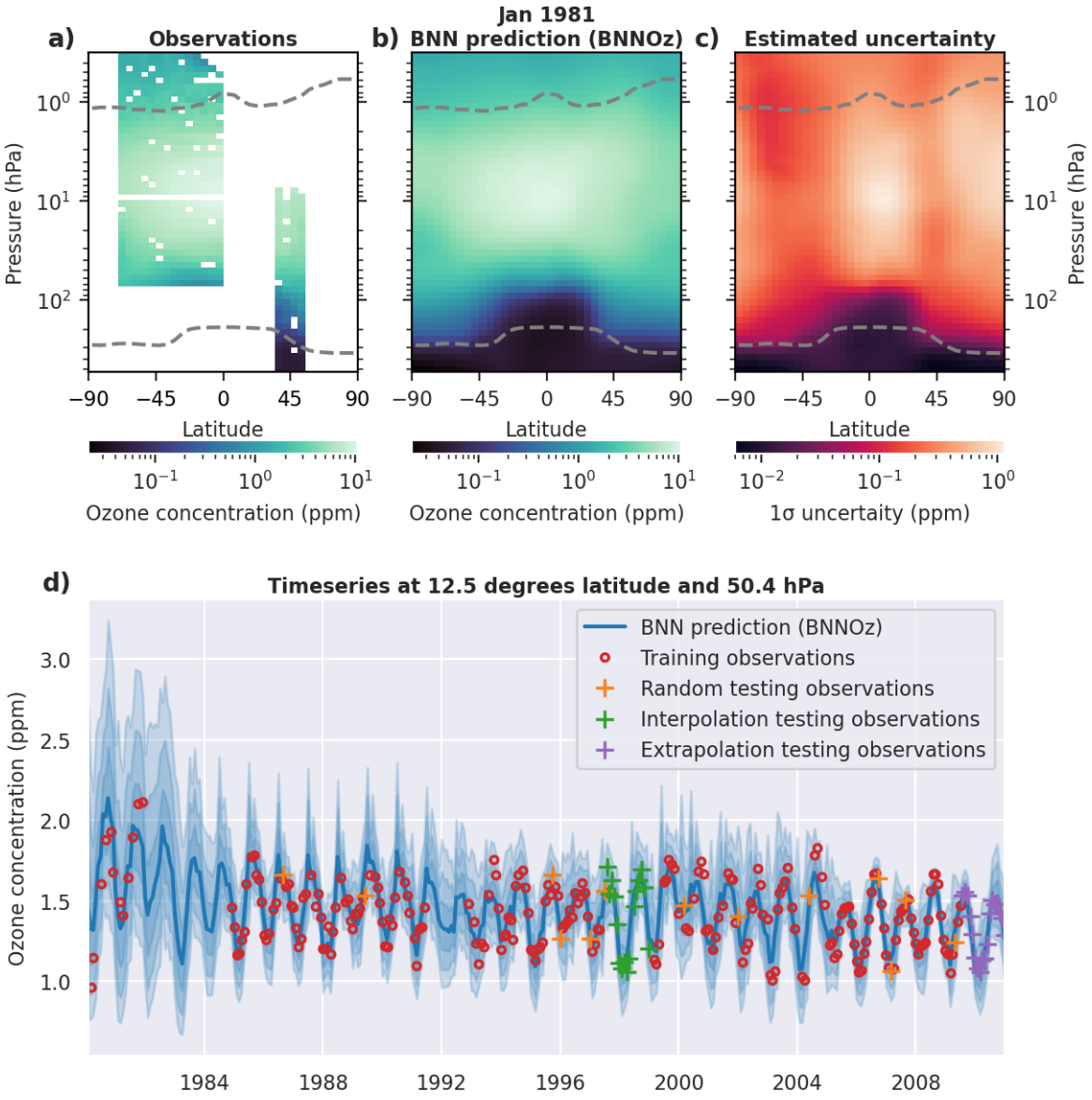

Figure 5.2: **A snapshot of the ozone concentration predictions from the Bayesian neural network (BNN)**. The top row shows zonal mean plots for January 1981 of **(a)** the BSTier0 ozone observations used to train the BNN (Bodeker et al., 2013), **(b)** the infilled ozone prediction from the BNN and **(c)** the estimated uncertainty in that prediction. Dashed grey lines in **(a)** to **(c)** show the approximate positions of the tropopause (lower) and stratopause (upper), defined using NCEP reanalysis (Kalnay et al., 1996) and CCMI model averages respectively. Panel **(d)** shows an example predicted monthly mean ozone timeseries from the BNN at 12.5°N and 50.4 hPa. The markers show the separate sets of observations used for training and testing, as specified in the legend (see section 5.2.3 for details), and the shading indicates the 68%, 95% and 99.7% confidence intervals (moving darker blue to light blue) for the BNN prediction.

levels and is therefore not an informative measure of performance for regions of low levels. Alongside RMSE and fractional error, we measure how well the BNN quantifies uncertainty by calculating what portion of the BNNOz data falls within 1, 2 and 3 standard deviations of the prediction, where the standard deviation is the uncertainty prediction from the BNN. Under a Gaussian assumption, we would expect these values to be 68.3 %, 95.5 % and 99.7 % respectively.

Table 5.1: Analyses of the performance of the BNN predictions and uncertainty quantification for the training and testing data splits. Root mean squared error and fractional error are used to assess predictive skill, and the skill of the uncertainty prediction is calculated by the portion of data that fall within the uncertainty prediction. Under a Gaussian assumption these values would be 68.3 %, 95.5 % and 99.7 % respectively.

Data split	Root mean squared error (ppm)	Mean fractional error (%)	Percentage of points within 1,2,3 standard deviations (%)
Training	0.14	1.2	80.0, 97.8, 99.6
Random testing	0.15	1.5	79.1, 97.4, 99.5
Interpolation testing	0.17	1.7	70.1, 94.7, 99.0
Extrapolation testing	0.11	1.5	77.6, 97.4, 99.5

The BNNOz prediction has been tested and validated against the three testing datasets to ensure that the BNN method can interpolate and extrapolate over large spatial regions and time periods of missing data, similar to those present in the original BSTier0 data. The results of these performance and uncertainty tests are shown in Table 5.1. The mean fractional error is between 1 % and 2 % for the different testing splits, showing that the infilling performance of the BNN is good and consistent across both interpolation and extrapolation tasks. Low fractional error and low RMSE scores across the training and testing datasets demonstrate that the BNN is capable of reconstructing the observations accurately from the ensemble of CCMs. Furthermore, this predictive skill suggests that the BNN can successfully learn meaningful model weights, bias and noise.

The uncertainty quantification scores, displayed in the final column of Table 5.1, show that the BNNOz uncertainty estimates are good and are even slightly under-confident for all the testing datasets. Uncertainty estimates of the BNNOz are

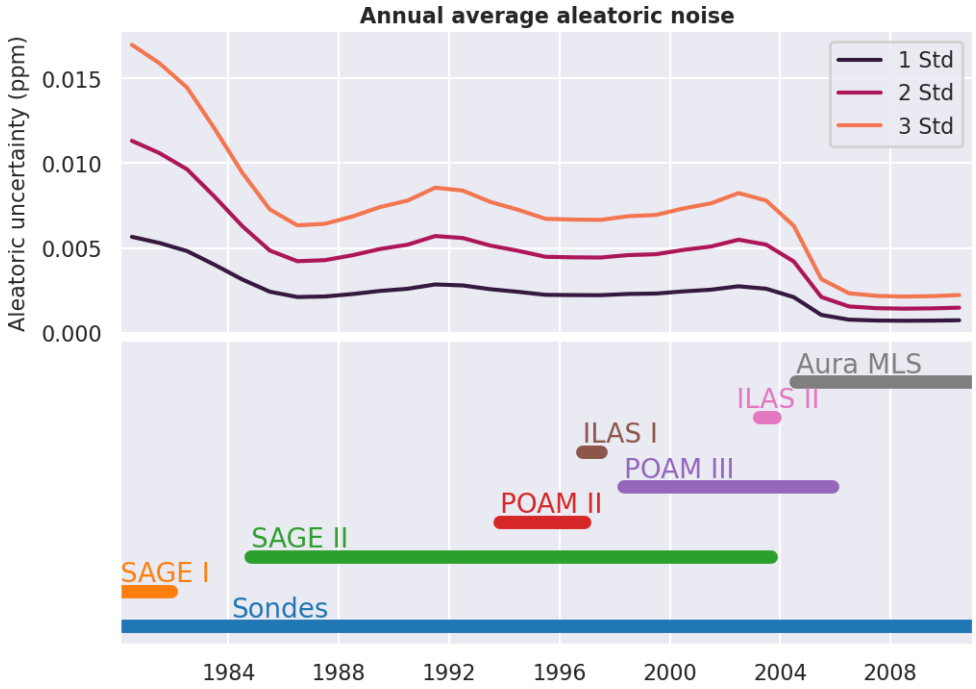

Figure 5.3: The average annual aleatoric uncertainty (standard deviation) calculated by the BNN (top panel), shown relative to the temporal coverage of satellites used in the BSTier0 dataset (bottom panel). The aleatoric noise is a measure of the estimated uncertainty in the observations.

comprised of both aleatoric (observational) and epistemic (knowledge) uncertainties, and vary in time and space to capture that different instruments with different levels of precision are used in the BSTier0 observational dataset. Figure 5.3 shows how the estimated annual aleatoric uncertainty is dependent on which satellite products are used. Uncertainty is high early in the record and decreases to a minimum after 2004 when the high quality AuraMLS (Waters et al., 2006) satellite data becomes the main satellite product used. Although the BNN is not given explicit information about the construction of the original dataset, by allowing BNN parameters to vary temporally and spatially the BNN has captured and accounted for complex features of the underlying data.

We would typically expect the performance of the BNN to be lower for extrapolation than interpolation, as extrapolation tasks are constrained by less surrounding data than interpolation tasks. However, for this set of models and observations the BNN

extrapolates more skillfully. The reason for this can be traced back to the construction of the testing datasets and the original BSTier0 dataset we are infilling, which is a conglomerate of measurements from multiple instruments. Firstly, the observational noise decreases in time, as shown in Figure 5.3, meaning that the latter years covering the extrapolation test (Jul2009–Dec2010) are likely of higher quality than the data used in the interpolation test (Jul1997–Dec1998). Secondly, the observational data coverage increases throughout the record meaning the extrapolation testing period is better constrained. These performance results do not mean that the BNN is better at extrapolating than interpolating; the opposite is true in general and it is due to varying quality and availability of ozone observations.

5.4.2 Comparison to existing vertically resolved ozone datasets

We now compare BNNOz to four existing datasets of vertically resolved ozone datasets, to highlight some shared similarities and differences, and to demonstrate how the extrapolation capabilities of the BNN allow new exploration of under observed regions. Table 5.2 summarises the details of the comparator datasets, and Figure 5.4 indicates their vertical coverage. Their latitudinal coverage varies from global (90°S–90°N) to near global (60°S–60°N), depending on the observations and infilling methods used. The BASIC and SBUV cohesive datasets (acronyms are introduced in Table 5.2) do not extend down into the troposphere whereas SWOOSH and BSTier1.4 do. None of these datasets represent independent comparisons to the BNN prediction as they are all constructed, at least partially, from the same satellite and observational products present in the BSTier0 data used to train the BNN.

5.4.2.1 Ozone anomalies

To compare ozone datasets, demonstrate the improved coverage of BNNOz and highlight uncertainty, Figures 5.5 and 5.6 show ozone anomalies for set latitudinal regions and pressure levels, based on the analyses presented in the Global Ozone chapter (Braesicke et al., 2018) of the World Meteorological Organisation (WMO) Scientific Assessment of Ozone Depletion report (WMO, 2018). Anomalies are calculated as the annual

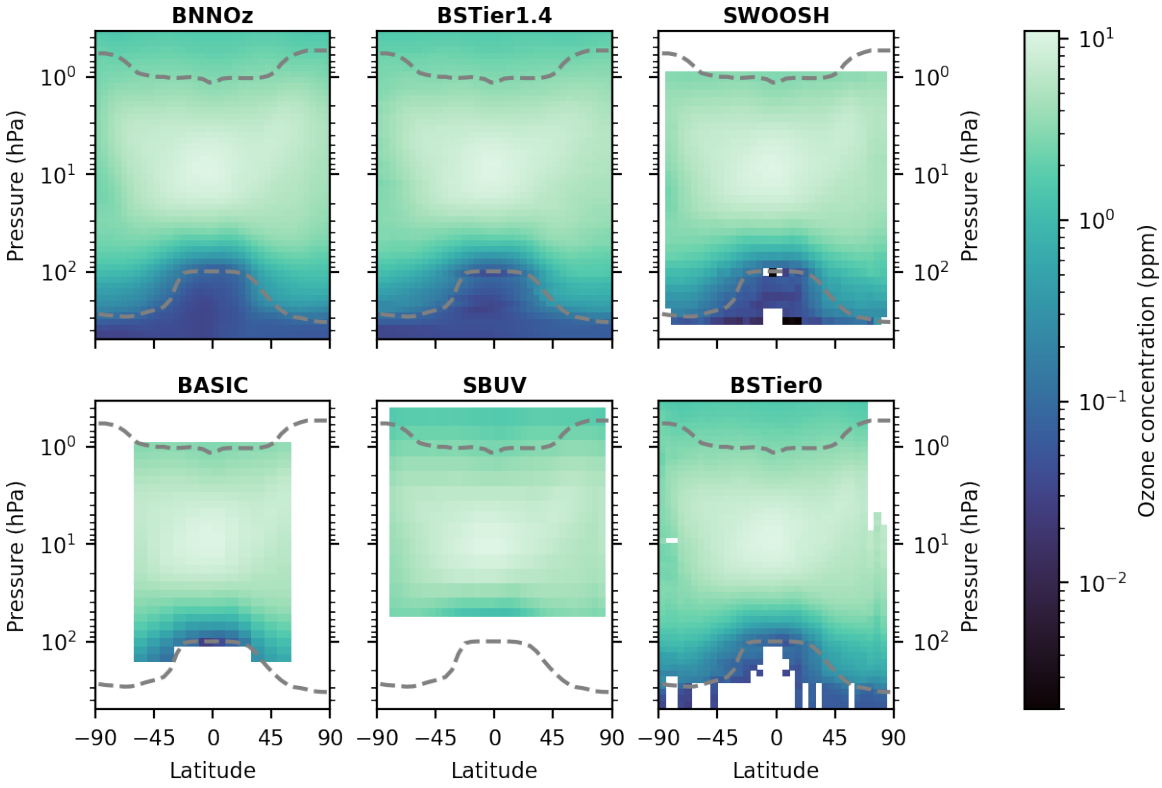

Figure 5.4: **Monthly averaged ozone concentration for March 1995 for 6 datasets showing typical latitudinal and vertical coverage.** Dashed grey lines show the approximate positions of the tropopause and stratopause. Apparent missing data in the tropical tropopause for SWOOSH are caused by the incompatibility of negative ozone concentrations with log scaling.

Table 5.2: Descriptions of the comparative datasets used. BNNOz, BSTier1.4, and SWOOSH are all infilled datasets, whereas BASIC and SBUV are unfilled. We used the version of SWOOSH that was filled by a combination of replacing latitudinal gaps with their corresponding equivalent latitude measurement and radial basis function interpolation to fill smaller gaps, as this is the most complete SWOOSH ozone dataset. BSTier1.4 is filled using a multi-linear regression of global time series expanded into harmonic components. The BASIC (BAyeSian Integrated and Consolidated composite ozone time-series dataset) version we used for comparison did not use SBUV(/2) observations.

Dataset	Acronym	Temporal coverage	Vertical coverage	Reference
BNN prediction	BNNOz	1980–2010	500–0.3 hPa	
Bodeker Scientific Tier1.4	BSTier1.4	1979–2016	878.4–0.046 hPa	Bodeker et al. (2013)
SWOOSH (eq lat and anomaly filled)	SWOOSH	1984–	316.2–1 hPa	Davis et al. (2016)
BASIC (without SBUV obs)	BASIC	1985–2018	146.8–1.2 hPa	Alsing and Ball (2017) Ball et al. (2017)
SBUV cohesive	SBUV	1980–	50–0.5 hPa	Miller et al. (2002)

mean ozone anomaly from deseasonalised time series, relative to the base period of 1998–2008, at pressure levels of 2, 10, 20 and 70 hPa.

Figure 5.5 shows ozone anomalies at pressure levels of 2, 10, 20, 70 hPa for the near global average (60°S–60°N) and averaged over the southern mid-latitudes (60°S–35°S), tropics (20°S–20°N) and northern mid-latitudes (35°N–60°N), which are regions covered by all datasets. Figure 5.6 shows anomalies at the same pressure levels for the southern polar cap (90°S–60°S), northern polar cap (90°S–90°N) and the entire globe (90°S–90°N), which are regions that require some form of infilling and as a result are only shown for BNNOz, BSTier1.4 and SWOOSH. For both figures, the annual mean ozone anomalies are calculated from a deseasonalised (relative to 1998–2008) ozone time series and are relative to a 1998–2008 base period.

Figure 5.5 shows that BNNOz is in good agreement with other vertically resolved ozone products for the four regions within 60°S–60°N. The similarity between the BNNOz and other products is greater at higher pressures (lower altitudes) where the BNNOz uncertainty largely encompasses other datasets, although it decreases with altitude. In time periods of higher data coverage the datasets are more likely to be

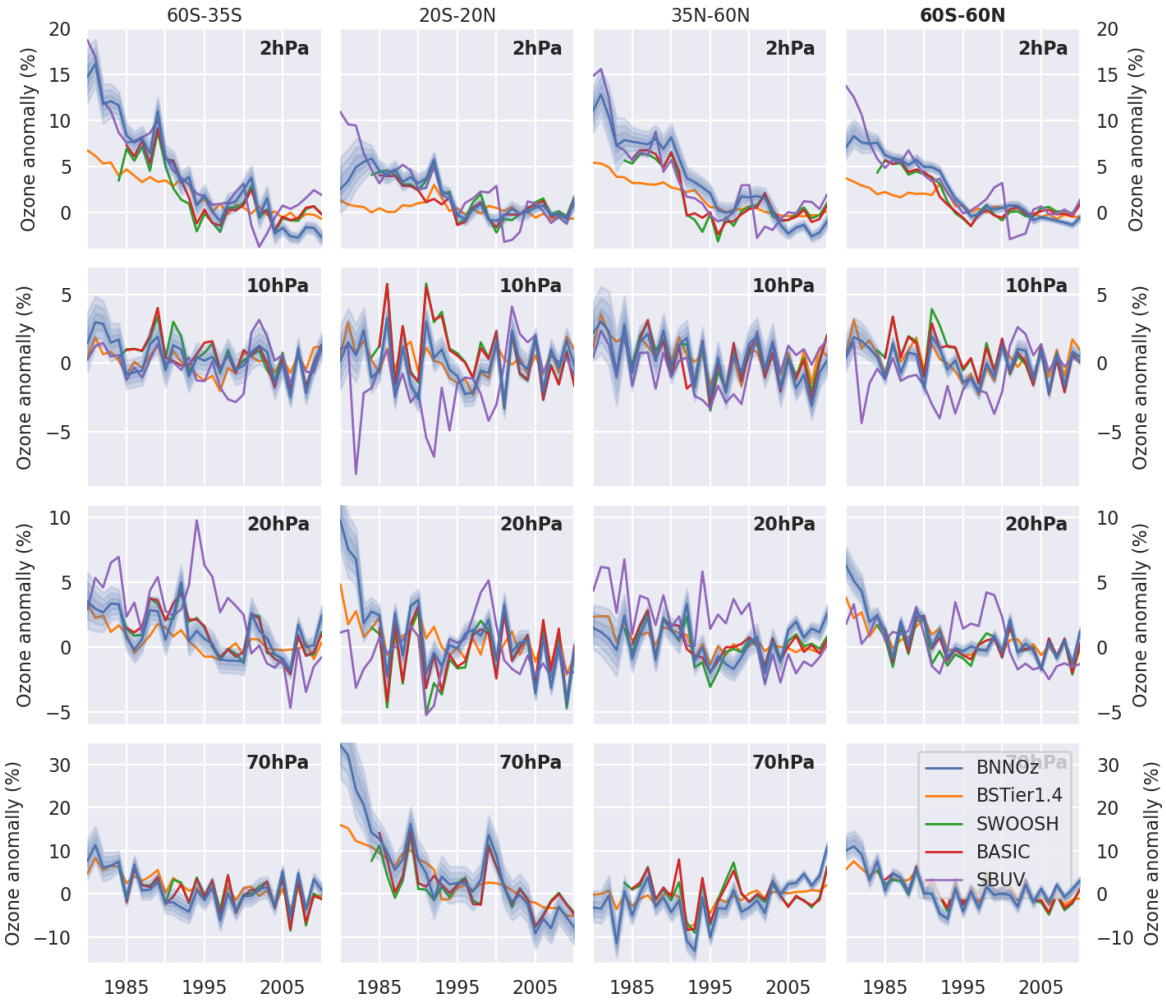

Figure 5.5: **Comparison of annual mean deseasonalised ozone anomalies from a range of ozone datasets: BNNOz, BSTier1.4, SWOOSH, BASIC and SBUV (cohesive).** Anomalies are shown for pressure levels 2, 10, 20 and 70 hPa, averaged over 60°S–35°S (first column), 20S–20N (second column), 35°N–60°N (third column) and 60°S–60°N (last column). The anomalies are calculated relative to the base period 1998–2008 and are area weighted. The shading around the BNNOz shows uncertainty at 68%, 95% and 99.7% confidence intervals. This figure is based on figures 3-15 and 3-16 from the 2018 WMO Ozone Assessment Report (WMO, 2018) itself adapted from the SPARC/IO3C/GAW LOTUS (Long-term Ozone Trends and Uncertainties in the Stratosphere) report (SPARC/IO3C/GAW, 2018).

Figure 5.6: **Comparison of annual mean deseasonalised ozone anomalies from infilled ozone datasets: BNNOz, BSTier1.4 and SWOOSH.** As per Figure 5.4, but with anomalies averaged over 90°S–60°S (left column), 60°N–90°N (middle column), 90°S–90°N (right column). The anomalies are calculated relative to the base period 1998–2008 and are area weighted. The shading around the BNNOz shows uncertainty at 68%, 95% and 99.7% confidence intervals. The BASIC and SBUV datasets are not shown as they have incomplete coverage beyond the range 60°S to 60°N. This figure is based on figures 3-15 and 3-16 from the 2018 WMO Ozone Assessment Report (WMO, 2018) itself adapted from the SPARC/IO3C/GAW LOTUS (Long-term Ozone Trends and Uncertainties in the Stratosphere) report (SPARC/IO3C/GAW, 2018).

in agreement, compared to early in the record where data availability is lower. This drop in data availability in the early part of the datasets is reflected in the increasing uncertainty of BNNOz, particularly between 1980–1985. Overall, for near-global, mid-latitude and tropical regions, the BNNOz successfully captures the long-term trends and smaller fluctuations, caused by forcing such as the solar cycle and the El Niño Southern Oscillation (ENSO) (WMO, 2018), that are seen across the datasets.

Figure 5.6 explores the predictions over polar regions and the global average, where the differences between ozone products should be more apparent due to the need for infilling in these regions. The effect of greater data sparsity over these regions, compared to the more data rich near global predictions in Figure 5.5, is seen in the increased BNNOz uncertainty. For the polar cap predictions (90°S–60°S and 60°N–90°N) the SWOOSH infilled predictions are typically in agreement with BNNOz within the uncertainty. The infilled BSTier1.4 prediction similarly falls mostly with the BNNOz uncertainty but displays much less interannual variability than BNNOz and SWOOSH, particularly at higher altitudes. As the underlying observations are then same between BNNOz and BSTier1.4, this difference is therefore a result of the BSTier1.4 infilling method which has led to a smoothing of the ozone anomaly time series. In addition to maintaining variability, BNNOz has the benefit of extending the other global infilled product SWOOSH to five years earlier.

5.4.2.2 Ozone trends

Ozone trends are typically calculated using a regression model which isolates the trend from background meteorological influences and known drivers of stratospheric ozone (SPARC/IO3C/GAW, 2018). Here, we calculate ozone trends using dynamical linear modelling (DLM) which performs regression against the same variables as the more commonly used multiple linear regression (MLR), but allows for a smoothly varying nonlinear ozone trend, unlike MLR. Using code from Alsing (2019), our implementation of DLM follows Ball et al. (2018) using the same regression time series: a latitudinally resolved stratospheric aerosol optical depth (Thomason et al., 2018), 30 cm radio flux (for solar variability) (Wit et al., 2014), an ENSO index (NCAR, 2019), and an index of the quasi-biennial oscillation (QBO) using zonal winds at both 30 and 50 hPa (Berlin, n.d.). Vertically resolved ozone trends for each dataset are calculated

1985–1997 (pre-1997) and 1998–2010 (post-1997) as 1997 was the approximate peak of halogen-containing ozone-depleting substances (Newman et al., 2007). The year 1997 also bisects the trend, enabling comparison with existing analysis performed with piecewise MLR (SPARC/IO3C/GAW, 2018). We then calculate percentage ozone change relative to the trend value at 1997.

Vertically resolved ozone trends for BNNOz and comparative datasets are shown in Figures 5.7 and 5.8 for the same latitudinal bands used in Figures 5.5 and 5.6. For the pre-1997 trends in the regions between 60°S–60°N (Figures 5.7 a, b, c and d), strong agreement is seen between BNNOz and the other datasets, particularly in the upper stratosphere where all datasets show significant (at 3-sigma) ozone depletion. For post-1997 trends (Figures 5.7 e, f, g and h), BNNOz similarly falls mainly within the dataset spread and shows neither significant widespread depletion or recovery.

For the pre-1997 trends over, and including, the polar regions (Figures 5.8 a, b and c), SWOOSH and BNNOz are in reasonable agreement with one another especially for globally averaged ozone (Figure 5.8 c). Similarly to near global regions, the BSTier1.4 displays a much weaker but more confident trend. BNNOz estimates a peak ozone reduction of 10 % at 1 hPa over the southern polar cap for the period 1985–1997, equating to a loss of $0.26\,\mathrm{ppm\,dec^{-1}}$. At this same location SWOOSH and BSTier1.4 estimate significantly lower losses of $0.094\,\mathrm{ppm\,dec^{-1}}$ and $0.052\,\mathrm{ppm\,dec^{-1}}$ respectively. Disparity in the trend estimates is largely caused by two factors. Firstly, datasets consist of various observational sources compiled and assimilated differently which results in significantly different datasets and therefore trends. Secondly the infilling methods will have an impact on the ozone datasets, which will be particularly exacerbated over polar regions. For example, BSTier1.4 extrapolates using an expanded form of MLR and presents different ozone anomalies and trends compared to BNNOz, itself based on the same observational data (BSTier0). This difference between BSTier1.4 and BNNOz is even more apparent in Figures 5.8 d, e and f, that show post-1997 trends, where the three continuous datasets (BNNOz, BSTier1.4 and SWOOSH) are in disagreement about recovery over the southern pole and globally.

That the calculated trends disagree, whilst having a high confidence of the disagreement, is strongly indicative that the differences are caused by the underlying observations. For that reason, it is difficult to determine whether datasets differ due

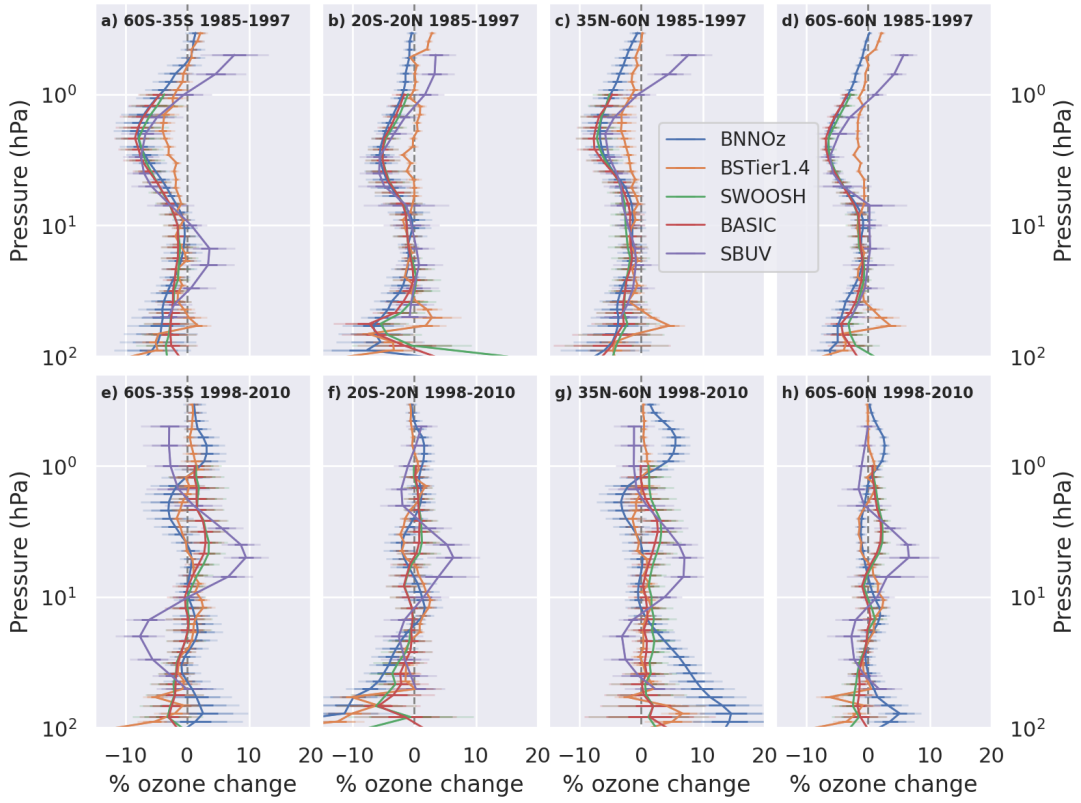

Figure 5.7: **Ozone trends calculated using dynamical linear modelling (DLM) as a percentage ozone change relative to 1997, for 5 ozone datasets exclusive of polar regions.** Trends are shown for pre-1997 (top row) over latitude bands 60°S–35S, 20°S–20°N, 35°N–60°N and 60°S–60°N (a, b, c and d respectively), and for post-1997 (bottom panel) for the same latitude bands (e, f, g and h). The coloured lines represent each dataset and are shown with shaded uncertainty bars (1, 2 and 3 standard deviations) calculated from the DLM trend estimation.

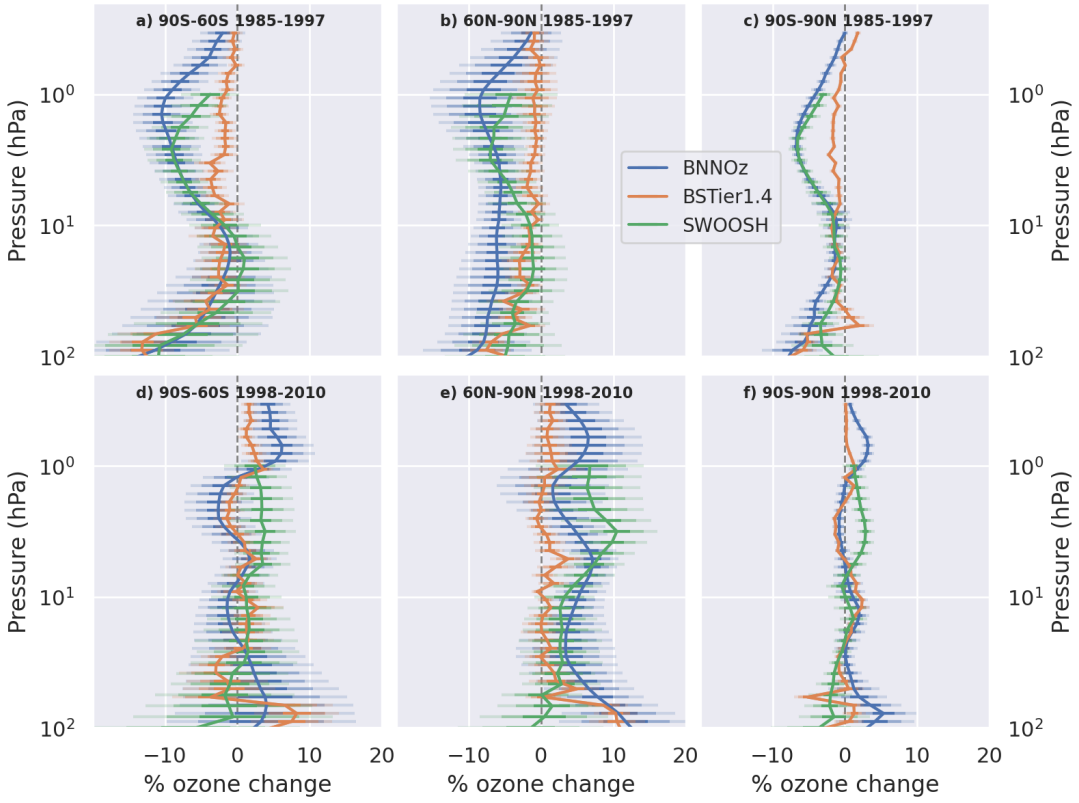

Figure 5.8: **Ozone trends calculated using dynamical linear modelling (DLM) as a percentage ozone change relative to 1997, for 5 ozone datasets inclusive of polar regions**. Trends are shown for pre-1997 (top row) over latitude bands 90°S–60°S, 60°N–90°N, and 90°S–90°N (a, b and c respectively), and for post-1997 (bottom panel) for the same latitude bands (d, e, and f). The coloured lines represent datasets that have consistent near global coverage and are shown with shaded uncertainty bars (1, 2 and 3 standard deviations) calculated from the DLM trend estimation.

to their infilling method or their observational foundation.

5.5 Conclusion

We have described the BNN methodology of model-observation fusion and demonstrate an application of producing a globally continuous dataset of vertically resolved ozone (BNNOz), extending our original application of the BNN in producing a total ozone column dataset (Sengupta et al., 2020). Using the BNN, we can combine physically and chemically realistic chemistry-climate model output with observations to generate predictions of ozone that are accurate and have an associated uncertainty. This Bayesian deep learning method successfully extrapolates over systematically under-observed regions and periods, as it relies on chemistry-climate models rather than a regression model-based method for the extrapolation. We have demonstrated the extrapolation and interpolation ability of the BNN and compared the BNNOz to widely used datasets. We anticipate a more exhaustive comparison and validation of BNNOz against other existing datasets when output from the next phase of CCMI comes available, extending the simulations.

This BNN approach to infilling observations has a number of advantages over other commonly used methods. It is a Bayesian approach which, as highlighted by Ball et al. (2017), is robust to unknown observational uncertainty and quantifies both statistical and observational uncertainty. Given the construction of the prediction, from weighted models and bias, it is interpretable as model contributions can be scrutinised spatially and temporally, to encourage trust beyond other opaque ML methods. This interpretability also furthers our understanding of the underlying observations, accounting for the varied quality of different observational sources and assimilation methods. As the BNN predictions are based on an ensemble of physically realistic models, we maintain accurate interannual variability, unlike linear regression methods that can smooth infilled output (e.g., Bodeker et al., 2013), and can recover trends, unlike infilling techniques that interpolate to climatological averages (e.g., Davis et al., 2016). As such, the BNNOz dataset is less limited by the assumptions made by these other infilling approaches.

The infilled prediction is strongly influenced by the choice of the infilling method.

This is especially seen by comparing BNNOz and BSTier1.4 that use different infilling methods but the same original dataset. For BSTier1.4 the ozone anomalies tend to be more smoothly varying than BNNOz, indicating a lower ability to capture interannual variability. The differences between the approaches are more pronounced for ozone trend analyses, where BSTier1.4 exhibits a smaller trend than BNNOz, particularly in regions such as the poles that require more infilling. Differences in ozone trends and anomalies originate in the infilling methods, demonstrating the importance of investigating and testing the various methods, and providing principled uncertainty quantification.

Separate to differences caused by infilling methods, comparison between datasets in regions with good observational coverage show us that differences in ozone trends and anomalies are also due to the underlying observations. This may originate from the observational products used, retrieval methods or bias correction and as a result, infilled predictions will be highly dependent on these choices (WMO, 2018). Even with a principled infilling technique such as the BNN, the infilled product can only be as good as the ground-truth observational data it learns from. Future efforts therefore, should concentrate on improvements in observational data assimilation and pre-processing, or end-to-end approaches from satellite retrieval to infilling which would allow the propagation of errors throughout all the processing.

Chapter 6

Conclusion and general remarks

This thesis created, demonstrated and explored new sophisticated tools that improve upon current methods of assimilating model ensembles and observations, improving the accuracy and uncertainty quantification of predictions and projections of stratospheric ozone. By drawing on ideas from data science and machine learning (ML), whilst balancing the interests of modellers and atmospheric scientists, these new methods enabled broad investigation of ozone in the upper atmosphere, generating modern ozone hole recovery estimates, creating a continuous vertically resolved ozone dataset crucial in trend analyses, and producing ozone predictions for traditionally under-observed regions or periods. These tools are flexible and can have wide reaching application in many environmental modelling scenarios where ensembles of geophysical models are relied upon.

Chapter 1 introduced the key themes of the thesis: stratospheric ozone – its importance to life, which processes determine its abundance, particularly anthropogenically driven depletion, and how it is modelled and observed; model ensembles – how assimilation from multiple models can provide more accurate predictions, the current methods used to perform assimilation (or ensembling), and the caveats and limitations in these methods; and, ML and data science – the types of problems ML is skilled at solving and recent applications of ML within atmospheric science and environmental modelling. This chapter motivated the need for a better understanding of both historic and modelled stratospheric ozone and suggested that ML approaches used in conjunction with model ensembles could achieve this.

Chapter 2 further developed a model weighting framework used within the climate modelling community that accounted for variable model performance and similarity (Knutti et al., 2017) and detailed its application to projecting Antarctic ozone depletion. Rather than weighting upon a single variable as is commonly done in the climate community, the CCMs were weighted across a suite of metrics relevant to ozone depletion over the southern polar cap, requiring the models to simulate multiple parts of the stratospheric ozone system accurately. This framework was used with CCMs from the Chemistry-Climate Modelling Initiative (CCMI) (Morgenstern et al., 2017) and various reanalysis and observational products, to produce a weighted projection of Antarctic ozone hole recovery. The weighted projection showed ozone hole recovery (to 1980 levels) by 2056 with a 95 % confidence interval of 2052–2060, earlier than the most recent study (Dhomse et al., 2018) that used the same chemistry-climate models (CCMs) but non-weighted multi-model means. Perfect model testing and metric sensitivity testing were conducted which showed this weighting framework's ability to produce more accurate projections than the widely used multi-model mean. This work will feature in the polar ozone chapter of the WMO 2022 Ozone Assessment Report.

Chapter 3 presented a novel framework for fusing geophysical model ensembles with observations within a Bayesian neural network (BNN). This framework is capable of using discontinuous observations to infer spatiotemporally varying model weights and bias, whilst estimating the spatiotemporally varying observational uncertainty. Compared to existing model averaging and weighting methodologies, such as those presented in Chapter 2, the BNN represents a large improvement in model ensemble and observational assimilation through improved accuracy, representation of total uncertainty and a more complete consideration of the ensemble design, or lack thereof. The BNN was rigorously tested using synthetic and real-world data, showing increased predictive ability and uncertainty quantification over existing model ensembling and data imputation methods. An application of this framework for infilling historic records was demonstrated through the infilling of total ozone column using CCMI models and the NIWA-BS record. By learning optimal weights, bias and noise of the nudged CCMI model output, the BNN produced temporally and spatially continuous gridded total ozone column from 1980–2010 complete with uncertainty predictions.

Alongside atmospheric science contributions, this chapter, published at an established ML conference, presented the first use of an anchored BNN which accounts for variable data uncertainty (heteroscedasticity).

Chapter 4 extended the ML focussed Chapter 3 with a discussion about how the results from the BNN can inform our understanding about the CCMI models and observations. This interpretability investigation compared BNN inferred model weights, bias and uncertainties with more traditional measures of model performance, bias and similarity. Strong relationships were shown between uncertainty and data availability, as expected, and between uncertainty and latitude. Polar regions exhibited much higher uncertainty than tropical regions, showing that the BNN appropriately represented uncertainties contained in the original ozone dataset used to train the BNN. The BNN inferred bias matched the seasonality of CCMs' bias well. There were some indications to show that BNN inferred CCM weights were correlated with CCM performance, but this was not consistent across all temporal and spatial scales. However, similarity between BNN weights for known similar models, showed that the BNN was capable of identifying similar models. As well as opening the ML *black box*, this chapter further highlighted the complexity and difficulty in appropriately quantifying model similarity and performance, as was discussed in Chapter 1. Although the BNN displayed some understanding of model performance and similarity, its main utility, and where it excels, is as an ensembling tool, not an ensemble analysis tool.

Chapter 5 presented a specific application of the Bayesian neural network framework, extending it to infilling historic vertically resolved ozone, essential for determining the depletion and recovery of ozone throughout the atmosphere. The gridded infilled dataset that spans 1980–2010 and 500–0.3 hPa addressed flaws in previous infilled datasets that suffered from underestimated interannual variability and trends in polar regions, caused by the difficulty in extrapolating over polar regions and other large areas of missing data. The chapter presented comprehensive trend analysis, using dynamical linear modelling, of all infilled vertically resolved ozone datasets, demonstrating the importance of using principled infilling techniques.

As a whole, this thesis presented how our use of CCMs can be improved using ML and data science methods, which account for the complexity and features of ensembles. The developed tools improved our capabilities to forecast and hindcast stratospheric

ozone leading to datasets that are more accurate and uncertainty-aware than their predecessors.

The ozone datasets and projections produced in this thesis will enable further and more detailed investigation of stratospheric ozone. Continued creation of BNN fused datasets, as and when new CCM ensembles and observations become available, will aid not only stratospheric ozone monitoring for the determination of ozone recovery and adherence to the Montreal Protocol, but also the important impacts on ozone depletion and variability on climate, health and the chemical composition of the atmosphere. The second phase of CCMI simulations are to be released in 2022, which simulate 1960–2016 offering an extension of the BNN into time before ozone monitoring satellites. It will be interesting to see how the BNN handles larger uncertainties 20 years out-of-sample and if the datasets it produces can offer useful and meaningful constraints on ozone abundances pre-1980. Similarly, the increasing use of CCMs to investigate tropospheric composition (e.g., Collins et al., 2017; Young et al., 2018) provides new possibilities for using and testing BNNs and process-based weighted means, applications that will require using point-based observations rather than previously used gridded observations and other methodological innovation to account for the high temporal variability.

Analyses in this thesis have heavily relied upon observational datasets, particularly of ozone. As the use of ML and data-driven approaches grows, the need for large and accurate datasets will similarly grow, requiring continued addition to and maintenance of observational records. With particular reference to ozone, it is particularly important for differences between observational stratospheric ozone products (seen in Chapter 5) to be reconciled, including the removal or correction of erroneous data that currently exist in multiple datasets. Otherwise, ML approaches will learn to fit inaccurate data and in turn produce inaccurate predictions. This of course is not a critique of just ozone datasets, any data-driven method is liable to learn badly from erroneous data, so an ongoing challenge in environmental data science is the curation of accurate datasets for use with data-driven approaches.

Ensembles of environmental models and their outputs are and continue to be a fundamental part of environmental science (IPCC, 2013b; WMO, 2018). Given the continual increase in model participation, model complexity and reliance on analyses of ensembles, the tools used to assimilate model output, analyse petabytes of data and

validate model output need to continue improving. There is some community uptake in more sophisticated ensemble methods. The most recent IPCC report notably considers weighted model means of global mean temperatures, using methods similar to those in Chapter 2. However, methods that account for variable model performance and similarity are not broadly used throughout the report as communities lack a universal and robust framework for weighting model projections.

The tools presented in this thesis, process based weighted means and the BNN for ensembling models, are part of a wider landscape of environmental data science methods and algorithms, often relying on ML, which are being developed to address these challenges and bring new insight to environmental science. ML is not yet set to replace large scale environmental models, although this research is underway (Schneider et al., 2017), but it can quickly provide useful and impactful post-hoc analyses of large observational and model datasets in a way that traditional analyses cannot (e.g., Ryan et al., 2018; Nicely et al., 2020). Although the field of environmental data science is growing, further work is required to establish its tools amongst those commonly used by environmental scientists, including establishing best practices for handling environmental data in ML applications and building reproducible and easy to use community specific tools. This thesis suggested several best practices including how to map geophysical coordinates onto coordinates that are more meaningful to ML tools, how to rigorously test ML algorithms against irregular environmental data and how to encode knowledge of physical systems into ML models.

These contributions, the thesis as a whole, and more broadly environmental data science, are possible through collaborations across environmental science, atmospheric modelling, ML and statistics. Collaborations provide a cross-disciplinary understanding of ML tools and environmental data and problems and underpin many of the advances in environmental data science. Careful balance is required in these collaborations, as improving on the current environmental science standards is unlikely to require state-of-the-art ML or novel statistical innovation. However, the collaboration that led to the BNN in this thesis advanced both environmental science capability and state-of-the-art ML. It is promising to see that so many institutions (industry and academic), funding bodies and publishers are understanding of the need for collaborations and are providing the necessary support.

The process-based weighting methodology and the BNNs relied upon nudged CCM runs where the model is forced to observed meteorology (Orbe et al., 2020). These simulations are models' best efforts at recreating the past and are therefore extremely useful as they allow for direct comparison between observations and model output. However, these types of simulations are not widely used by other environmental modellers, such as climate modellers or oceanographers. To broaden the applicability and flexibility of the BNN, future work is required to incorporate free running models and by doing so, produce probabilistic projections. However, here the benefit of using a Bayesian approach becomes an issue, as far out of sample projections would generate large and uninformative uncertainties. Why would we expect the BNN to predict what the weight of a model should be in 100 years time? To address this issue and enable the BNN to make projections, I propose a reframing of the problem which requires less extrapolation. Rather than asking how a model's weight is expected to vary in time, we can ask how we expect the model weight to vary dependent on physical parameters. For example, in the case of stratospheric ozone, we could infer model weights dependent on a combination of stratospheric temperatures, the amount of equivalent effective stratospheric chlorine and other process-based measures. Weighting on physical and modelled variables would require less, or in some cases no extrapolation, and give a physiochemical sense to the weighting beyond the current spatiotemporal coordinates. This approach blends the process-based weighting from Chapter 2 with the BNN and is an area of active research currently.

ML is not a silver bullet that removes the need for process models or their ensembles. Their grounding in physical and chemical processes and equations is still invaluable for the exploration of the Earth system and its components. ML and data science however, do provide us with better ways to use these model ensembles, improving the robustness, accuracy and utility of their projections and predictions.

Appendix A

Supplementary material for chapter 3: Ensembling geophysical models with Bayesian Neural Networks

A.1 Construction of ozone baselines

We remind the reader that all of these baselines use the same data for training, testing and validation as the Bayesian neural network ensemble. This validation tests the ability of the ensembling methods to interpolate and extrapolate, particularly over regions of interest and sparse data.

A.1.1 Multi-model mean

This is the uniform weighting of all the 15 chemistry-climate models. The prediction is therefore,

$$y_{\mathrm{MMM}}(\mathbf{x}, t) = \frac{1}{15} \sum_{i=1}^{15} M_i(\mathbf{x}, t) \tag{A.1}$$

where $y_{\mathrm{MMM}}(\mathbf{x}, t)$ is the multi-model mean prediction and $M_i(\mathbf{x}, t)$ is the i-th individual model prediction.

A.1.2 Weighted mean

Here the ensemble mean is constructed from model projections weighted by their skill (in replicating observations) and their independence. This is based on work from Knutti et al. (2017) and Sanderson et al. (2017). For an ensemble of N models, the weight for model i (w_i) is given by

$$w_i = \exp\left(-\frac{D_i^2}{n_i\sigma_D^2}\right) / \left(1 + \sum_{j\neq i}^{N} \exp\left(-\frac{S_{ij}^2}{n_i\sigma_S^2}\right)\right), \tag{A.2}$$

where D_i^2 is the squared difference between a model and observation averaged over all space and time, S_{ij}^2 is the squared difference between models averaged over all space and time, n_i is the size of data used in constructing the weight, and σ_D and σ_S are constants which allow tuning of the weighting. The weights w_i are normalised to sum to 1. The weighted prediction is therefore

$$y_{\mathrm{WM}}(\mathbf{x}, t) = \sum_{i=1}^{N} w_i M_i(\mathbf{x}, t). \tag{A.3}$$

Values of σ_D and σ_S were found by minimising the difference between the weighted prediction and the observations over the training data.

A.1.3 Spatially weighted mean

The ensemble is constructed much the same as the weighted mean presented above, except that model weights vary in space. The weights are calculated

$$w_i(\mathbf{x}, t) = \exp\left(-\frac{D_i(\mathbf{x}, t)^2}{n_i(\mathbf{x}, t)\sigma_D^2}\right) / \left(1 + \sum_{j\neq i}^{N} \exp\left(-\frac{S_{ij}(\mathbf{x}, t)^2}{n_i(\mathbf{x}, t)\sigma_S^2}\right)\right), \tag{A.4}$$

and are used to generate the prediction,

$$y_{\mathrm{SWM}}(\mathbf{x}, t) = \sum_{i=1}^{N} w_i(\mathbf{x}, t) M_i(\mathbf{x}, t). \tag{A.5}$$

A.1.4 Spatiotemporal kriging

We performed spatiotemporal kriging using an implementation of a stochastic variational Gaussian process (SVGP) from GPFlow (Matthews et al., 2017). Due to the size of the training data (1.8 million data points) we used a SVGP on 3 year sections of observational data with 5000 inducing points per section. The kernel we used was the product of a Matern3/2 kernel acting over latitude and time, and periodic Matern3/2 kernel acting over longitude. We used an Adam optimiser implemented in tensorflow to train the SVGP.

A.1.5 Bilinear interpolation

Bilinear interpolation over the training data was performed using the SciPy function griddata: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html`.

A.2 Hyperparameter details

The pretrained model weights and the code to run the BayNNE for both the synthetic and ozone experiments can be found here: `https://anonymous.4open.science/r/6bf08e5a-c909-45a3-be63-aa0f5ba187df/`. Table A.1 shows the hyperparameters chosen in the running of the BayNNE for both experiments.

The heteroscedastic loss function is prone to episodes of catastrophic forgetting. To avoid these, we use large batch sizes, small learning rates and a large number of epochs so that each neural network ensemble member may be stably trained till convergence.

The neural network ensemble for the 2 million datapoint ozone dataset were trained across a cluster of 6 P100 GPUs. Each neural network needed 20 hours to be trained till convergence and the entire ensemble needed 8 days of wall clock time (NOTE that with a small tweak in the code post submission, the training time was reduced by an order of magnitude).

Table A.1: Hyperparameter values and priors for BayNNE.

	Synthetic experiment	Ozone experiment
Spatial coord scaling	2	2
Temporal coord scaling (month of year)	1	2
Temporal coord scaling (total months)	1	1
Number of physical models	4	15
Number of neural network ensemble members	50	65
Bias mean. prior	0	0
Bias std. prior	0.01	0.03
Noise mean prior	0.02	0.015
Noise std. prior	0.004	0.003
Number of hidden layers	1	1
Number of hidden nodes	100	500
Optimiser	Adam	Adam
Batch Size	2000	25000
Learning rate	5×10^{-5}	3×10^{-5}
Number of epochs	6000	125000

A.2.1 Neural network ensemble convergence

An important hyperparameter is the size of the neural network ensemble. We used 65 neural network ensemble members for the total ozone predictions. Figure A.1 demonstrates the convergence of the ensemble as early as an ensemble size of 30. However, we ran a larger ensemble to ensure convergence.

A.3 Derivation of loss function

In the following, we derive the anchored ensembling loss function for the heteroscedastic case. For the j-th neural network ensemble member in randomized MAP sampling, we compute the MAP estimate corresponding to a recentered prior over parameters $\boldsymbol{\theta}_{anc,j}$, $P(\boldsymbol{\theta}_j) = \mathcal{N}(\boldsymbol{\theta}_{anc,j}, \boldsymbol{\Sigma}_{prior})$. Here $\boldsymbol{\theta}_{anc,j}$ is a sample from the original multivariate normal prior over parameters, i.e. $\boldsymbol{\theta}_{anc,j} \sim \mathcal{N}(\boldsymbol{\mu}_{prior}, \boldsymbol{\Sigma}_{prior})$.

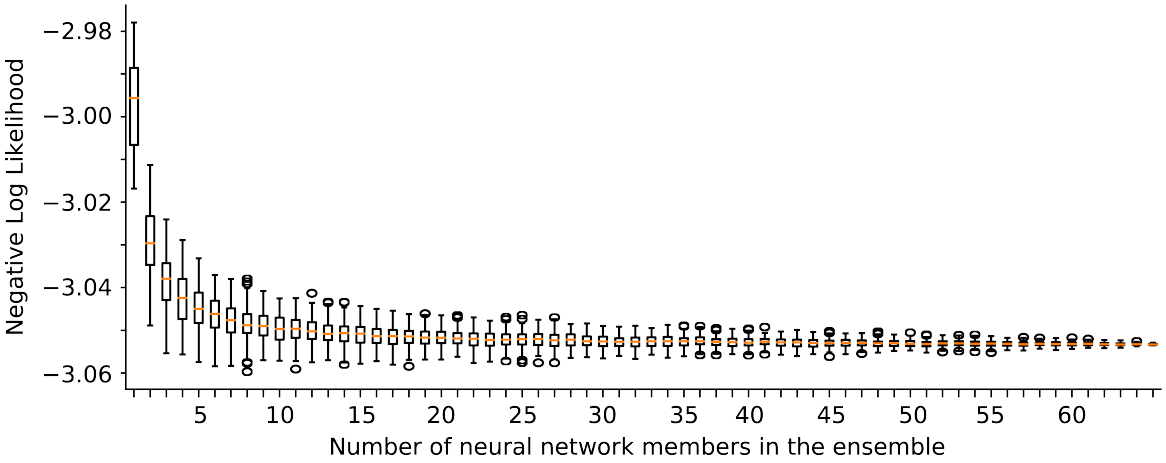

Figure A.1: The distribution of negative log likelihood dependent on the (neural network) ensemble size.

$$
\begin{aligned}
\boldsymbol{\theta}_{MAP,j} &= \text{argmax}_{\boldsymbol{\theta}_j} P(\boldsymbol{\theta}_j|\mathcal{D}) \\
&= \text{argmax}_{\boldsymbol{\theta}_j} P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j) P(\boldsymbol{\theta}_j) \quad \text{(Bayes' Theorem)} \\
&= \text{argmax}_{\boldsymbol{\theta}_j} \log(P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j)) + \log(P(\boldsymbol{\theta}_j)) \quad \text{(log strictly monotonic increasing)} \\
&= \text{argmax}_{\boldsymbol{\theta}_j} \log(P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j)) - \frac{1}{2}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})^T \boldsymbol{\Sigma}_{prior}^{-1} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j}) + \text{const.} \\
&= \text{argmax}_{\boldsymbol{\theta}_j} \log(P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j)) - \frac{1}{2} \|\boldsymbol{\Sigma}_{prior}^{-1/2} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{anc,j})\|_2^2 \quad \text{(diagonal prior cov.)}
\end{aligned}
$$

If we specify the data likelihood for our regression task assuming i.i.d. observations and additive heteroscedastic Gaussian noise i.e., $P_{\mathcal{D}}(\mathcal{D}|\boldsymbol{\theta}_j) = \prod_{i=1}^{N_D} \mathcal{N}(\hat{y}_j(\mathbf{x}_i, t_i), \sigma_j^2(\mathbf{x}_i, t_i))$, we obtain

$$\begin{aligned}\boldsymbol{\theta}_{MAP,j} = \text{argmax}_{\boldsymbol{\theta}_j} & -\frac{1}{2}\sum_{i=1}^{N_D}\frac{(y_i-\hat{y}_j(\mathbf{x}_i,t_i))^2}{\sigma_j^2(\mathbf{x}_i,t_i)} - \sum_{i=1}^{N_D}\log(\sigma_j(\mathbf{x}_i,t_i)) + \text{const.} \\ & -\frac{1}{2}\|\boldsymbol{\Sigma}_{prior}^{-1/2}(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{anc,j})\|_2^2 \\ = \text{argmin}_{\boldsymbol{\theta}_j} & \sum_{i=1}^{N_D}\frac{(y_i-\hat{y}_j(\mathbf{x}_i,t_i))^2}{\sigma_j^2(\mathbf{x}_i,t_i)} + \sum_{i=1}^{N_D}\log(\sigma_j^2(\mathbf{x}_i,t_i)) + \|\boldsymbol{\Sigma}_{prior}^{-1/2}(\boldsymbol{\theta}_j-\boldsymbol{\theta}_{anc,j})\|_2^2 \\ & (\times \text{ -2 throughout})\end{aligned}$$

A.4 Extra ozone experiment plots

In the main text we highlighted the models with the most interesting features and highest weights. For completeness here, we include a wider range of plots looking at model weights and modelled bias and uncertainties, for the ozone experiment.

A.4.1 Bias

Figures A.2 and A.3 show the modelled bias averaged in time and space respectively. Bias is seen to be negative over polar regions especially the southern polar region and southern mid latitudes.

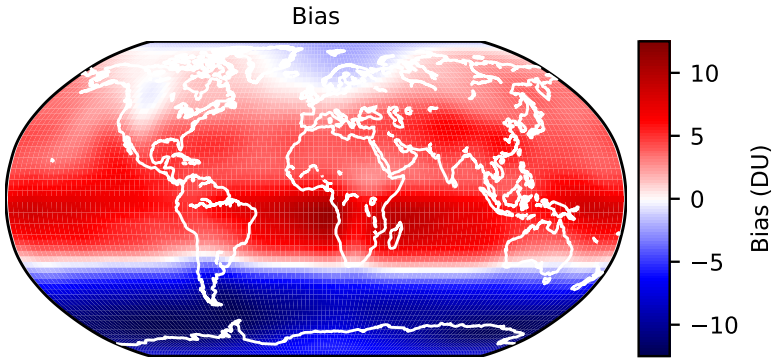

Figure A.2: Temporally averaged modelled bias.

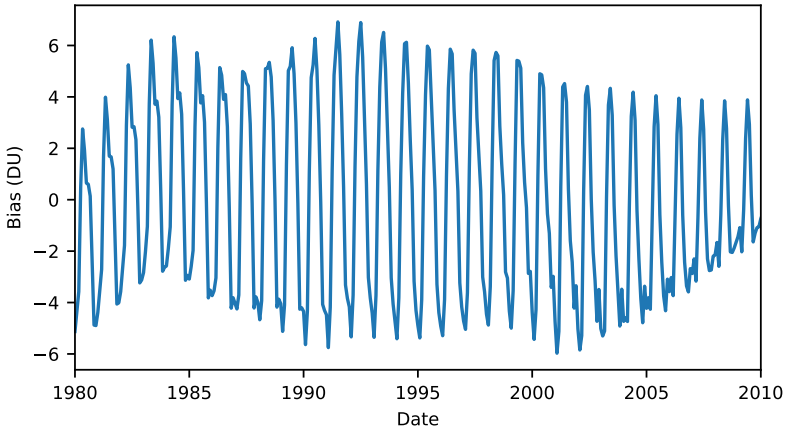

Figure A.3: Spatially averaged modelled bias.

A.4.2 Epistemic uncertainty

Figures A.4 and A.5 show the epistemic uncertainty averaged in time and space respectively. Epistemic uncertainty is highest at polar regions. Epistemic uncertainty increases for regions with sparse or no data including 2007–2010 (used to validate extrapolation) and 1993–1997 where there is a greater sparsity of data. This can be seen clearly in Figure A.5.

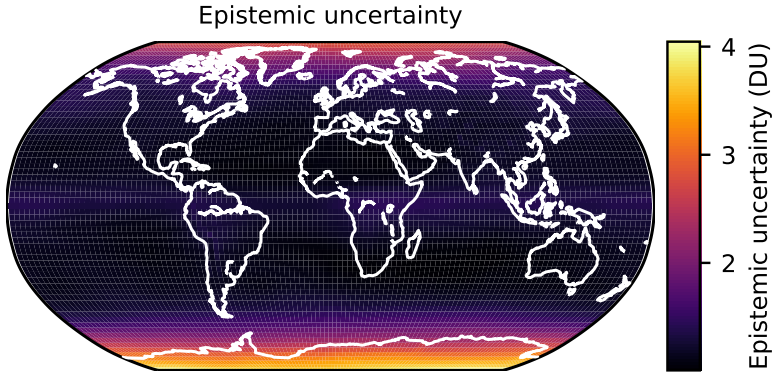

Figure A.4: Temporally averaged epistemic uncertainty.

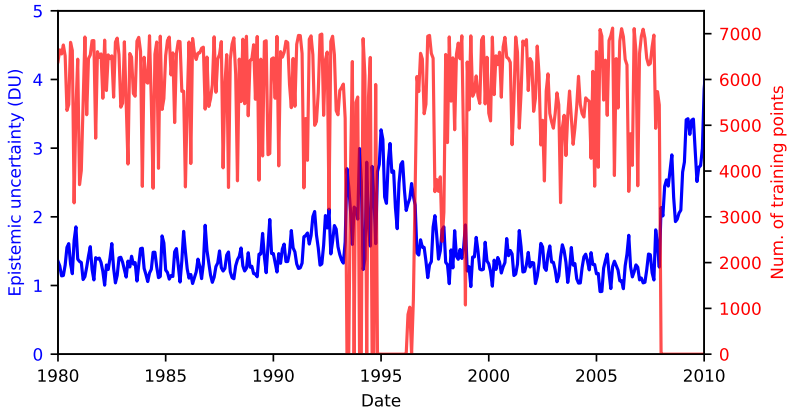

Figure A.5: Spatially averaged epistemic uncertainty and the number of training points per month.

A.4.3 Average model weight

Figure A.6 shows the average model weight for all 15 chemistry-climate models used in the ozone experiment.

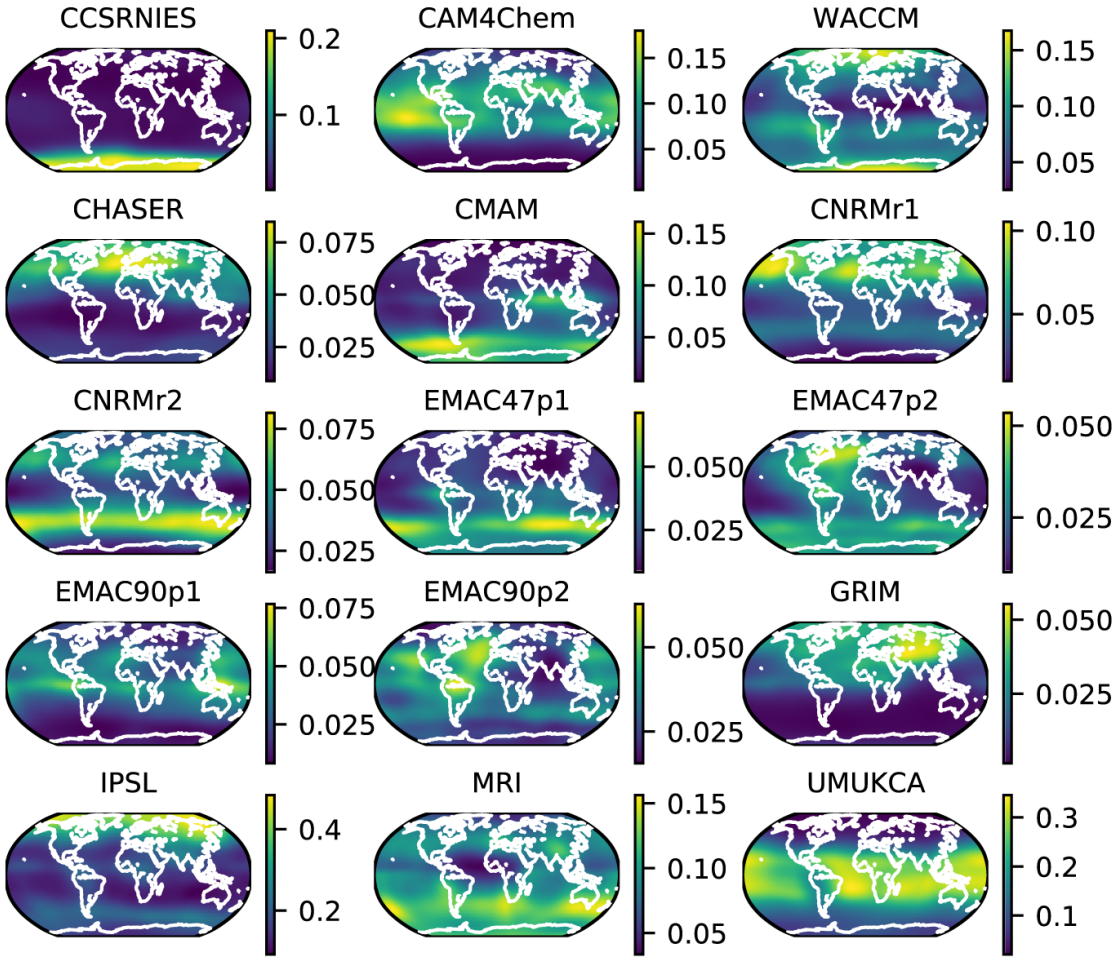

Figure A.6: Temporally averaged model weights for all 15 chemistry-climate models.

Appendix B

Supplementary material for chapter 5: A continuous vertically resolved ozone dataset from the fusion of chemistry climate models with observations using a Bayesian neural network

B.1 Training details

Table B.1: Hyperparameters used in the vertical ozone BNN training

Hyperparameter	Value
No. of NNs in BNN ensemble	48
Hidden layer size	500
Learning rate	0.0001
Optimiser	Adam
No. epochs	100000
Batch size	7500

References

Abalos, M., L. Polvani, N. Calvo, D. Kinnison, F. Ploeger, W. Randel, and S. Solomon. "New insights on the impact of ozone-depleting substances on the Brewer-Dobson circulation". *Journal of Geophysical Research: Atmospheres* 124.5 (2019), pp. 2435–2451. DOI: 10.1029/2018JD029301.

Abatzoglou, J. T., S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015". *Scientific data* 5 (2018), p. 170191. DOI: 10.1038/sdata.2017.191.

Abramowitz, G., N. Herger, E. Gutmann, D. Hammerling, R. Knutti, M. Leduc, R. Lorenz, R. Pincus, and G. A. Schmidt. "ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing". *Earth System Dynamics* 10.1 (2019), pp. 91–105. DOI: 10.5194/esd-10-91-2019.

Ahmed, K., D. A. Sachindra, S. Shahid, Z. Iqbal, N. Nawaz, and N. Khan. "Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms". *Atmospheric Research* 236 (2020), p. 104806. DOI: 10.1016/j.atmosres.2019.104806.

Akiyoshi, H., T. Nakamura, T. Miyasaka, M. Shiotani, and M. Suzuki. "A nudged chemistry-climate model simulation of chemical constituent distribution at northern high-latitude stratosphere observed by SMILES and MLS during the 2009/2010 stratospheric sudden warming". *Journal of Geophysical Research: Atmospheres* 121.3 (2016), pp. 1361–1380. DOI: 10.1002/2015JD023334.

Alsing, J. and W. Ball. "BASIC composite ozone time-series data V2". *Mendeley Data* 2 (2017). DOI: 10.17632/2mgx2xzzpk.

Alsing, J. "dlmmc: Dynamical linear model regression for atmospheric time-series analysis". *Journal of Open Source Software* 4.37 (2019), p. 1157. DOI: 10.21105/joss.01157.

Amos, M., P. J. Young, J. S. Hosking, J.-F. Lamarque, N. L. Abraham, H. Akiyoshi, A. T. Archibald, S. Bekki, M. Deushi, P. Jöckel, D. Kinnison, O. Kirner, M. Kunze, M. Marchand, D. A. Plummer, D. Saint-Martin, K. Sudo, S. Tilmes, and Y. Yamashita. "Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence". *Atmospheric Chemistry and Physics* 20.16 (2020), pp. 9961–9977. DOI: 10.5194/acp-20-9961-2020.

Andrews, D. G. and M. E. Mcintyre. "Planetary waves in horizontal and vertical shear: The generalized Eliassen-Palm relation and the mean zonal acceleration". *Journal of the Atmospheric Sciences* 33.11 (1976), pp. 2031–2048. DOI: https://doi.org/10.1175/1520-0469(1976)033\%3C2031:PWIHAV\%3E2.0.CO;2.

Annan, J. D. and J. C. Hargreaves. "On the generation and interpretation of probabilistic estimates of climate sensitivity". *Climatic Change* 104.3 (2011), pp. 423–436. DOI: doi.org/10.1007/s10584-009-9715-y.

Annan, J. D. and J. C. Hargreaves. "Understanding the CMIP3 multimodel ensemble". *Journal of Climate* 24.16 (2011), pp. 4529–4538. DOI: doi.org/10.1175/2011JCLI3873.1.

Anzai, Y. *Pattern recognition and machine learning*. Elsevier, 2012.

Austin, J., D. Shindell, S. R. Beagley, C. Brühl, M. Dameris, E. Manzini, T. Nagashima, P. Newman, S. Pawson, G. Pitari, E. Rozanov, C. Schnadt, and T. G. Shepherd. "Uncertainties and assessments of chemistry-climate models of the stratosphere". *Atmospheric Chemistry and Physics* 3.1 (2003), pp. 1–27. DOI: 10.5194/acp-3-1-2003.

Baker, N. C. and P. C. Taylor. "A framework for evaluating climate model performance metrics". *Journal of Climate* 29.5 (2016), pp. 1773–1782. DOI: 10.1175/JCLI-D-15-0114.1.

Baldwin, M. P., L. J. Gray, T. J. Dunkerton, K. Hamilton, P. H. Haynes, W. J. Randel, J. R. Holton, M. J. Alexander, I. Hirota, T. Horinouchi, D. B. A. Jones, J. S. Kinnersley, C. Marquardt, K. Sato, and M. Takahashi. "The quasi-biennial

oscillation". *Reviews of Geophysics* 39.2 (2001), pp. 179–229. DOI: 10.1029/1999RG000073.

Ball, W. T., J. Alsing, D. J. Mortlock, J. Staehelin, J. D. Haigh, T. Peter, F. Tummon, R. Stübi, A. Stenke, J. Anderson, A. Bourassa, S. M. Davis, D. Degenstein, S. Frith, L. Froidevaux, C. Roth, V. Sofieva, R. Wang, J. Wild, P. Yu, J. R. Ziemke, and E. V. Rozanov. "Evidence for a continuous decline in lower stratospheric ozone offsetting ozone layer recovery". *Atmospheric Chemistry and Physics* 18.2 (2018), pp. 1379–1394. DOI: 10.5194/acp-18-1379-2018.

Ball, W. T., J. Alsing, D. J. Mortlock, E. V. Rozanov, F. Tummon, and J. D. Haigh. "Reconciling differences in stratospheric ozone composites". *Atmospheric Chemistry and Physics* 17.20 (2017), pp. 12269–12302. DOI: 10.5194/acp-17-12269-2017.

Ball, W. T., J. Alsing, J. Staehelin, S. M. Davis, L. Froidevaux, and T. Peter. "Stratospheric ozone trends for 1985–2018: sensitivity to recent large variability". *Atmospheric Chemistry and Physics* 19.19 (2019), pp. 12731–12748. DOI: 10.5194/acp-19-12731-2019.

Barnes, E. A., J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson. "Viewing Forced Climate Patterns Through an AI Lens". *Geophysical Research Letters* 46.22 (2019), pp. 13389–13398. DOI: 10.1029/2019GL084944.

Barrett, J. W., P. M. Solomon, R. L. De Zafra, M. Jaramillo, L. Emmons, and A. Parrish. "Formation of the Antarctic ozone hole by the CIO dimer mechanism". *Nature* 336.6198 (1988), pp. 455–458. DOI: 10.1038/336455a0.

Bates, D. R. and M. Nicolet. "The photochemistry of atmospheric water vapor". *Journal of Geophysical Research* 55.3 (1950), pp. 301–327. DOI: 10.1029/JZ055i003p00301.

Bednarz, E. M., A. C. Maycock, N. L. Abraham, P. Braesicke, O. Dessens, and J. A. Pyle. "Future Arctic ozone recovery: the importance of chemistry and dynamics". *Atmospheric Chemistry and Physics* 16.18 (2016), pp. 12159–12176. DOI: 10.5194/acp-16-12159-2016.

Bellenger, H., E. Guilyardi, J. Leloup, M. Lengaigne, and J. Vialard. "ENSO representation in climate models: from CMIP3 to CMIP5". *Climate Dynamics* 42.7 (2014), pp. 1999–2018. DOI: 10.1007/s00382-013-1783-z.

Berlin, F. U. *The Quasi-Biennial-Oscillation (QBO) Data Serie last accessed: 1 September 2021)*. URL: https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html.

Berrisford, P., D. P. Dee, P. Poli, R. Brugge, M. Fielding, M. Fuentes, P. W. Kållberg, S. Kobayashi, S. Uppala, and A. Simmons. "The ERA-Interim archive Version 2.0". 1 (2011). URL: https://www.ecmwf.int/node/8174.

Blair, G. S., P. Henrys, A. Leeson, J. Watkins, E. Eastoe, S. Jarvis, and P. J. Young. "Data science of the natural environment: a research roadmap". *Frontiers in Environmental Science* 7 (2019), p. 121. DOI: 10.3389/fenvs.2019.00121.

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra. "Weight Uncertainty in Neural Network". In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. 2015, pp. 1613–1622. URL: https://proceedings.mlr.press/v37/blundell15.html.

Blunden, J. and D. S. Arndt. "State of the Climate in 2015". *Bulletin of the American Meteorological Society* 97.8 (2016). DOI: 10.1175/2016BAMSStateoftheClimate.1.

Bodeker, G. E., I. S. Boyd, and W. A. Matthews. "Trends and variability in vertical ozone and temperature profiles measured by ozonesondes at Lauder, New Zealand: 1986–1996". *Journal of Geophysical Research: Atmospheres* 103.D22 (1998), pp. 28661–28681. DOI: 10.1029/98JD02581.

Bodeker, G. E., B. Hassler, P. J. Young, and R. W. Portmann. "A vertically resolved, global, gap-free ozone database for assessing or constraining global climate model simulations". *Earth System Science Data* 5.1 (2013), pp. 31–43. DOI: 10.5194/essd-5-31-2013.

Bodeker, G. E., J. Nitzbon, J. S. Tradowsky, S. Kremser, A. Schwertheim, and J. Lewis. "A global total column ozone climate data record". *Earth System Science Data* 13.8 (2021), pp. 3885–3906. DOI: 10.5194/essd-13-3885-2021.

Bodeker, G. E. and S. Kremser. "Indicators of Antarctic ozone depletion: 1979 to 2019". *Atmospheric Chemistry and Physics* 21.7 (2021), pp. 5289–5300. DOI: 10.5194/acp-21-5289-2021.

Bodeker, G. E., J. Nitzbon, J. Lewis, A. Schwertheim, and J. S. Tradowsky. *NIWA-BS Total Column Ozone Database*. Version 3.4. Zenodo, 2018. DOI: 10.5281/zenodo.1346424.

Boé, J. "Interdependency in multimodel climate projections: component replication and result similarity". *Geophysical Research Letters* 45.6 (2018), pp. 2771–2779. DOI: 10.1002/2017GL076829.

Bowman, K. P. and A. J. Krueger. "A global climatology of total ozone from the Nimbus 7 total ozone mapping spectrometer". *Journal of Geophysical Research: Atmospheres* 90.D5 (1985), pp. 7967–7976. DOI: 10.1007/978-94-009-5313-0_74.

Boyd, J. P. "The noninteraction of waves with the zonally averaged flow on a spherical earth and the interrelationships on eddy fluxes of energy, heat and momentum". *Journal of Atmospheric Sciences* 33.12 (1976), pp. 2285–2291. DOI: 10.1175/1520-0469(1976)033<2285:TNOWWT>2.0.CO;2.

Brack, D. "Monitoring the Montreal protocol". In: *Verification yearbook*. VERTIC, 2003, pp. 209–226.

Braesicke, P., J. Neu, V. Fioletov, S. Godin-Beekmann, D. Hubert, I. Petropavlovskikh, M. Shiotani, and B.-M. Sinnhuber. "Update on Global ozone: past, present, and Future, Chapter 3". In: *Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project–Report No. 58*. WMO (World Meteorological Organization), Geneva, Switzerland, 2018. URL: http://ozone.unep.org/science/assessment/sap.

Brewer, A. W. "Evidence for a world circulation provided by the measurements of helium and water vapour distribution in the stratosphere". *Quarterly Journal of the Royal Meteorological Society* 75.326 (1949), pp. 351–363. DOI: 10.1002/qj.49707532603.

Brunner, L., R. Lorenz, M. Zumwald, and R. Knutti. "Quantifying uncertainty in European climate projections using combined performance-independence weighting". *Environmental Research Letters* 14.12 (2019), p. 124010. DOI: 10.1088/1748-9326/ab492f.

Butchart, N. "The Brewer-Dobson circulation". *Reviews of geophysics* 52.2 (2014), pp. 157–184. DOI: 10.1002/2013RG000448.

Butchart, N., I. Cionni, V. Eyring, T. G. Shepherd, D. W. Waugh, H. Akiyoshi, J. Austin, C. Brühl, M. P. Chipperfield, E. Cordero, M. Dameris, R. Deckert, S. Dhomse, S. M. Frith, R. R. Garcia, A. Gettelman, M. A. Giorgetta, D. E. Kinnison, F. Li, E. Mancini, C. McLandress, S. Pawson, G. Pitari, D. A. Plummer, E. Rozanov, F. Sassi, J. F. Scinocca, K. Shibata, B. Steil,

and W. Tian. "Chemistry–Climate Model Simulations of Twenty-First Century Stratospheric Climate and Circulation Changes". *Journal of Climate* 23.20 (2010), pp. 5349–5374. DOI: 10.1175/2010JCLI3404.1.

Butler, A., J. S. Daniel, R. W. Portmann, A. Ravishankara, P. J. Young, D. W. Fahey, and K. H. Rosenlof. "Diverse policy implications for future ozone and surface UV in a changing climate". *Environmental Research Letters* 11.6 (2016), p. 064017. DOI: 10.1088/1748-9326/11/6/064017.

CCMVal, S. *SPARC CCMVal Report on the Evaluation of Chemistry-Climate Models*. Tech. rep. WCRP/WMO, 2010. URL: http://www.atmosp.physics.utoronto.ca/SPARC.

Chapman, S. "XXXV. On ozone and atomic oxygen in the upper atmosphere". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 10.64 (1930), pp. 369–383. DOI: 10.1080/14786443009461588.

Chapman, W. E., A. C. Subramanian, L. Delle Monache, S. P. Xie, and F. M. Ralph. "Improving atmospheric river forecasts with machine learning". *Geophysical Research Letters* 46.17-18 (2019), pp. 10627–10635. DOI: 10.1029/2019GL083662.

Chattopadhyay, A., P. Hassanzadeh, and D. Subramanian. "Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network". *Nonlinear Processes in Geophysics* 27.3 (2020), pp. 373–389. DOI: 10.5194/npg-27-373-2020.

Chen, T., E. Fox, and C. Guestrin. "Stochastic Gradient Hamiltonian Monte Carlo". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. 2. 2014, pp. 1683–1691. URL: http://proceedings.mlr.press/v32/cheni14.pdf.

Chipperfield, M. P., S. Bekki, S. Dhomse, N. R. Harris, B. Hassler, R. Hossaini, W. Steinbrecht, R. Thiéblemont, and M. Weber. "Detecting recovery of the stratospheric ozone layer". *Nature* 549.7671 (2017), p. 211. DOI: 10.1038/nature23681.

Chipperfield, M. P., S. S. Dhomse, W. Feng, R. McKenzie, G. J. Velders, and J. A. Pyle. "Quantifying the ozone and ultraviolet benefits already achieved by the Montreal Protocol". *Nature communications* 6.1 (2015), pp. 1–8. DOI: 10.1038/ncomms8233.

Christensen, J. H., E. Kjellström, F. Giorgi, G. Lenderink, and M. Rummukainen. "Weight assignment in regional climate models". *Climate Research* 44.2-3 (2010), pp. 179–194. DOI: 10.3354/cr00916.

Christopoulos, C. D., S. Garimella, M. A. Zawadowicz, O. Möhler, and D. J. Cziczo. "A machine learning approach to aerosol classification for single-particle mass spectrometry". *Atmospheric Measurement Techniques* 11.10 (2018), pp. 5687–5699. DOI: 10.5194/amt-11-5687-2018.

Chrysanthou, A., A. C. Maycock, M. P. Chipperfield, S. Dhomse, H. Garny, D. Kinnison, H. Akiyoshi, M. Deushi, R. R. Garcia, P. Jöckel, O. Kirner, G. Pitari, D. A. Plummer, L. Revell, E. Rozanov, A. Stenke, T. Y. Tanaka, D. Visioni, and Y. Yamashita. "The effect of atmospheric nudging on the stratospheric residual circulation in chemistry–climate models". *Atmospheric Chemistry and Physics* 19.17 (2019), pp. 11559–11586. DOI: 10.5194/acp-19-11559-2019.

Cionni, I., V. Eyring, J.-F. Lamarque, W. Randel, D. Stevenson, F. Wu, G. Bodeker, T. Shepherd, D. Shindell, and D. Waugh. "Ozone database in support of CMIP5 simulations: results and corresponding radiative forcing". *Atmospheric Chemistry and Physics* 11.21 (2011), pp. 11267–11292. DOI: 10.5194/acp-11-11267-2011.

Claxton, T., R. Hossaini, O. Wild, M. P. Chipperfield, and C. Wilson. "On the Regional and Seasonal Ozone Depletion Potential of Chlorinated Very Short-Lived Substances". *Geophysical Research Letters* 46.10 (2019), pp. 5489–5498. DOI: 10.1029/2018GL081455.

Collins, W. J., J.-F. Lamarque, M. Schulz, O. Boucher, V. Eyring, M. I. Hegglin, A. Maycock, G. Myhre, M. Prather, D. Shindell, and S. J. Smith. "AerChemMIP: quantifying the effects of chemistry and aerosols in CMIP6". *Geoscientific Model Development* 10.2 (2017), pp. 585–607. DOI: 10.5194/gmd-10-585-2017.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth. "A strategy for operational implementation of 4D-Var, using an incremental approach". *Quarterly Journal of the Royal Meteorological Society* 120.519 (1994), pp. 1367–1387. DOI: 10.1002/qj.49712051912.

Crutzen, P. J. "The influence of nitrogen oxides on the atmospheric ozone content". *Quarterly Journal of the Royal Meteorological Society* 96.408 (1970), pp. 320–325. DOI: 10.1002/qj.49709640815.

Crutzen, P. J. "Ozone production rates in an oxygen-hydrogen-nitrogen oxide atmosphere". *Journal of Geophysical Research* 76.30 (1971), pp. 7311–7327. DOI: 10.1029/JC076i030p07311.

Crutzen, P. J. and F. Arnold. "Nitric acid cloud formation in the cold Antarctic stratosphere: A major cause for the springtime 'ozone hole'". *Nature* 324.6098 (1986), pp. 651–655. DOI: 10.1038/324651a0.

Curry, J. A. and P. J. Webster. "Climate science and the uncertainty monster". *Bulletin of the American Meteorological Society* 92.12 (2011), pp. 1667–1682. DOI: 10.1175/2011BAMS3139.1.

Davis, S. M., K. H. Rosenlof, B. Hassler, D. F. Hurst, W. G. Read, H. Vömel, H. Selkirk, M. Fujiwara, and R. Damadeo. "The Stratospheric Water and Ozone Satellite Homogenized (SWOOSH) database: a long-term database for climate studies". *Earth system science data* 8.2 (2016), pp. 461–490. DOI: 10.5194/essd-8-461-2016.

Dennison, F., J. Keeble, O. Morgenstern, G. Zeng, N. L. Abraham, and X. Yang. "Improvements to stratospheric chemistry scheme in the UM-UKCA (v10. 7) model: solar cycle and heterogeneous reactions". *Geoscientific Model Development* 12.3 (2019), pp. 1227–1239. URL: 10.5194/gmd-12-1227-2019.

Deushi, M. and K. Shibata. "Development of a Meteorological Research Institute chemistry-climate model version 2 for the study of tropospheric and stratospheric chemistry". *Papers in Meteorology and Geophysics* 62 (2011), pp. 1–46. DOI: 10.2467/mripapers.62.1.

deWolff, T., H. Carrillo, L. Martí, and N. Sanchez-Pi. "Assessing physics informed neural networks in ocean modelling and climate change applications". In: *AI: Modeling Oceans and Climate Change Workshop at ICLR 2021*. 2021. URL: https://hal.inria.fr/hal-03262684.

Dhomse, S. S., M. P. Chipperfield, R. P. Damadeo, J. M. Zawodny, W. T. Ball, W. Feng, R. Hossaini, G. Mann, and J. D. Haigh. "On the ambiguous nature of the 11 year solar cycle signal in upper stratospheric ozone". *Geophysical Research Letters* 43.13 (2016), pp. 7241–7249. DOI: 10.1002/2016GL069958.

Dhomse, S. S., D. Kinnison, M. P. Chipperfield, R. J. Salawitch, I. Cionni, M. I. Hegglin, N. L. Abraham, H. Akiyoshi, A. T. Archibald, E. M. Bednarz, S. Bekki,

P. Braesicke, N. Butchart, M. Dameris, M. Deushi, S. Frith, S. C. Hardiman, B. Hassler, L. W. Horowitz, R.-M. Hu, et al. "Estimates of ozone return dates from Chemistry-Climate Model Initiative simulations". *Atmospheric Chemistry and Physics* 18.11 (2018), pp. 8409–8438. DOI: 10.5194/acp-18-8409-2018.

Dhomse, S. S., C. Arosio, W. Feng, A. Rozanov, M. Weber, and M. P. Chipperfield. "ML-TOMCAT: Machine-Learning-Based Satellite-Corrected Global Stratospheric Ozone Profile Dataset from a Chemical Transport Model". *Earth System Science Data Discussions* (2021), pp. 1–29. DOI: 10.5194/essd-2021-225.

Dobson, G. M. B. "Origin and distribution of the polyatomic molecules in the atmosphere". *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 236.1205 (1956), pp. 187–193. DOI: 10.1098/rspa.1956.0127.

Dufresne, J.-L., M.-A. Foujols, S. Denvil, A. Caubel, O. Marti, O. Aumont, Y. Balkanski, S. Bekki, H. Bellenger, R. Benshila, et al. "Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5". *Climate Dynamics* 40.9-10 (2013), pp. 2123–2165. DOI: 10.1007/s00382-012-1636-1.

Dunkerton, T. "On the mean meridional mass motions of the stratosphere and mesosphere". *Journal of Atmospheric Sciences* 35.12 (1978), pp. 2325–2333. DOI: 10.1175/1520-0469(1978)035<2325:OTMMMM>2.0.CO;2.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". *Geoscientific Model Development* 9.5 (2016), pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016.

Eyring, V., N. Butchart, D. W. Waugh, H. Akiyoshi, J. Austin, S. Bekki, G. E. Bodeker, B. A. Boville, C. Brühl, M. P. Chipperfield, E. Cordero, M. Dameris, M. Deushi, V. E. Fioletov, S. M. Frith, R. R. Garcia, A. Gettelman, M. A. Giorgetta, V. Grewe, L. Jourdain, et al. "Assessment of temperature, trace species, and ozone in chemistry-climate model simulations of the recent past". *Journal of Geophysical Research: Atmospheres* 111.D22 (2006). DOI: 10.1029/2006JD007327.

Eyring, V., I. Cionni, G. E. Bodeker, A. J. Charlton-Perez, D. E. Kinnison, J. F. Scinocca, D. W. Waugh, H. Akiyoshi, S. Bekki, M. P. Chipperfield, M. Dameris, S. Dhomse, S. M. Frith, H. Garny, A. Gettelman, A. Kubin, U. Langematz, E. Mancini,

M. Marchand, T. Nakamura, et al. "Multi-model assessment of stratospheric ozone return dates and ozone recovery in CCMVal-2 models". *Atmospheric Chemistry and Physics* 10.19 (2010), pp. 9451–9472. DOI: 10.5194/acp-10-9451-2010.

Eyring, V., J.-F. Lamarque, P. Hess, A. F., K. Bowman, M. P. Chipperfield, B. Duncan, A. Fiore, A. Gettelman, M. A. Giorgetta, C. Granier, M. Hegglin, D. Kinnison, M. Kunze, U. Langematz, B. Luo, R. Martin, K. Matthes, P. A. Newman, T. Peter, et al. "Overview of IGAC/SPARC Chemistry-Climate Model Initiative (CCMI) community simulations in support of upcoming ozone and climate assessments". *SPARC Newsletter* 40 (2013), pp. 48–66. URL: http://oceanrep.geomar.de/20227/.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". *Geoscientific Model Development* 9 (2016). DOI: 10.5194/gmd-9-1937-2016.

Eyring, V., M. P. Chipperfield, M. A. Giorgetta, D. E. Kinnison, E. Manzini, K. Matthes, P. A. Newman, S. Pawson, T. G. Shepherd, and D. W. Waugh. "Overview of the new CCMVal reference and sensitivity simulations in support of upcoming ozone and climate assessments and the planned SPARC CCMVal report". *SPARC Newsletter* 30 (2008), pp. 20–26. URL: http://oceanrep.geomar.de/15163/.

Eyring, V., P. M. Cox, G. M. Flato, P. J. Gleckler, G. Abramowitz, P. Caldwell, W. D. Collins, B. K. Gier, A. D. Hall, F. M. Hoffman, et al. "Taking climate model evaluation to the next level". *Nature Climate Change* 9.2 (2019), pp. 102–110. DOI: 10.1038/s41558-018-0355-y.

Faghmous, J. H. and V. Kumar. "A big data guide to understanding climate change: The case for theory-guided data science". *Big data* 2.3 (2014), pp. 155–163. DOI: 10.1089/big.2014.0026.

Farman, J. C., B. G. Gardiner, and J. D. Shanklin. "Large losses of total ozone in Antarctica reveal seasonal ClO x/NO x interaction". *Nature* 315.6016 (1985), pp. 207–210. DOI: 10.1038/315207a0.

Fente, D. N. and D. K. Singh. "Weather forecasting using artificial neural network". In: *2018 second international conference on inventive communication and computa-*

tional technologies (ICICCT). IEEE. 2018, pp. 1757–1761. DOI: 10.1109/ICICCT.2018.8473167.

Fioletov, V. E., G. Labow, R. Evans, E. W. Hare, U. Köhler, C. T. McElroy, K. Miyagawa, A. Redondas, V. Savastiouk, A. M. Shalamyansky, J. Staehelin, K. Vanicek, and M. Weber. "Performance of the ground-based total ozone network assessed using satellite data". *Journal of Geophysical Research: Atmospheres* 113.D14 (2008). DOI: 10.1029/2008JD009809.

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. "Evaluation of Climate Models, Chapter 9". In: *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, 2014, pp. 741–866. DOI: 10.1017/CBO9781107415324.020.

Foong, A. Y. K., D. R. Burt, Y. Li, and R. E. Turner. "On the Expressiveness of Approximate Inference in Bayesian Neural Networks". In: vol. 33. 2020. URL: https://proceedings.neurips.cc/paper/2020/file/b6dfd41875bc090bd31d0b1740eb5b1b-Paper.pdf.

Ford, J. D., S. E. Tilleard, L. Berrang-Ford, M. Araos, R. Biesbroek, A. C. Lesnikowski, G. K. MacDonald, A. Hsu, C. Chen, and L. Bizikova. "Opinion: Big data has big potential for applications to climate change adaptation". *Proceedings of the National Academy of Sciences* 113.39 (2016), pp. 10729–10732. DOI: 10.1073/pnas.1614023113.

Froidevaux, L., J. Anderson, H.-J. Wang, R. A. Fuller, M. J. Schwartz, M. L. Santee, N. J. Livesey, H. C. Pumphrey, P. F. Bernath, J. M. Russell III, and M. P. McCormick. "Global OZone Chemistry And Related trace gas Data records for the Stratosphere (GOZCARDS): methodology and sample results with a focus on HCl, H_2O, and O_3". *Atmospheric Chemistry and Physics* 15.18 (2015), pp. 10471–10507. DOI: 10.5194/acp-15-10471-2015.

Froidevaux, L., Y. B. Jiang, A. Lambert, N. J. Livesey, W. G. Read, J. W. Waters, E. V. Browell, J. W. Hair, M. A. Avery, T. J. McGee, L. W. Twigg, G. K. Sumnicht, K. W. Jucks, J. J. Margitan, B. Sen, R. A. Stachnik, G. C. Toon, P. F. Bernath,

C. D. Boone, K. A. Walker, et al. "Validation of aura microwave limb sounder stratospheric ozone measurements". *Journal of Geophysical Research: Atmospheres* 113.D15 (2008). DOI: 10.1029/2007JD008771.

Gal, Y. and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. 2016, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

Garcia, R. R., A. K. Smith, D. E. Kinnison, Á. d. l. Cámara, and D. J. Murphy. "Modification of the gravity wave parameterization in the Whole Atmosphere Community Climate Model: Motivation and results". *Journal of the Atmospheric Sciences* 74.1 (2017), pp. 275–291. DOI: 10.1175/JAS-D-16-0104.1.

Gareau, B. J. "Lessons from the Montreal Protocol delay in phasing out methyl bromide". *Journal of Environmental Studies and Sciences* 5.2 (2015), pp. 163–168. DOI: 10.1007/s13412-014-0212-x.

Gaudel, A., O. R. Cooper, G. Ancellet, B. Barret, A. Boynard, J. P. Burrows, C. Clerbaux, P.-F. Coheur, J. Cuesta, E. Cuevas, S. Doniki, G. Dufour, F. Ebojie, G. Foret, O. Garcia, M. J. Granados-Muñoz, J. W. Hannigan, F. Hase, B. Hassler, G. Huang, et al. "Tropospheric Ozone Assessment Report: Present-day distribution and trends of tropospheric ozone relevant to climate and global atmospheric chemistry model evaluation". *Elementa: Science of the Anthropocene* 6 (2018). DOI: 10.1525/elementa.291.

Gillett, N. P. "Weighting climate model projections using observational constraints". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2054 (2015), p. 20140425. DOI: 10.1098/rsta.2014.0425.

Gillett, Z. E., J. M. Arblaster, A. J. Dittus, M. Deushi, P. Jöckel, D. E. Kinnison, O. Morgenstern, D. A. Plummer, L. E. Revell, E. Rozanov, R. Schofield, A. Stenke, K. A. Stone, and S. Tilmes. "Evaluating the relationship between interannual variations in the Antarctic ozone hole and southern hemisphere surface climate in chemistry–climate models". *Journal of Climate* 32.11 (2019), pp. 3131–3151. DOI: 10.1175/JCLI-D-18-0273.1.

Giorgi, F. and L. O. Mearns. "Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method". *Journal of Climate* 15.10 (2002), pp. 1141–1158. DOI: 10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux. "Performance metrics for climate models". *Journal of Geophysical Research: Atmospheres* 113.D6 (2008). DOI: 10.1029/2007JD008972.

Grazzini, F., G. C. Craig, C. Keil, G. Antolini, and V. Pavan. "Extreme precipitation events over northern Italy. Part I: A systematic classification with machine-learning techniques". *Quarterly Journal of the Royal Meteorological Society* 146.726 (2020), pp. 69–85.

Groenke, B., L. Madaus, and C. Monteleoni. "ClimAlign: Unsupervised statistical downscaling of climate variables via normalizing flows". In: *Proceedings of the 10th International Conference on Climate Informatics*. 2020, pp. 60–66. DOI: 10.1145/3429309.3429318.

Guo, H.-D., L. Zhang, and L.-W. Zhu. "Earth observation big data for climate change research". *Advances in Climate Change Research* 6.2 (2015), pp. 108–117. DOI: 10.1016/j.accre.2015.09.007.

Gurney, K. *An introduction to neural networks*. CRC press, 2018. DOI: 10.1201/9781315273570.

Harm, W. *Biological effects of ultraviolet radiation*. Cambridge University Press, 1980. DOI: 10.1002/jobm.19810210917.

Harrison, S. P., P. Bartlein, K. Izumi, G. Li, J. Annan, J. Hargreaves, P. Braconnot, and M. Kageyama. "Evaluation of CMIP5 palaeo-simulations to improve climate projections". *Nature Climate Change* 5.8 (2015), p. 735. DOI: 10.1038/nclimate2649.

Haughton, N., G. Abramowitz, A. Pitman, and S. J. Phipps. "Weighting climate model ensembles for mean and variance estimates". *Climate dynamics* 45.11-12 (2015), pp. 3169–3181. DOI: doi.org/10.1007/s00382-015-2531-3.

Haupt, S. E., J. Cowie, S. Linden, T. McCandless, B. Kosovic, and S. Alessandrini. "Machine Learning for Applied Weather Prediction". In: *2018 IEEE 14th Inter-*

national Conference on e-Science (e-Science). 2018, pp. 276–277. DOI: 10.1109/eScience.2018.00047.

Heath, D., A. J. Krueger, H. Roeder, and B. Henderson. "The solar backscatter ultraviolet and total ozone mapping spectrometer (SBUV/TOMS) for Nimbus G". *Optical Engineering* 14.4 (1975), p. 144323. DOI: 10.1117/12.7971839.

Herger, N., G. Abramowitz, R. Knutti, O. Angélil, K. Lehmann, and B. M. Sanderson. "Selecting a climate model subset to optimise key ensemble properties". *Earth System Dynamics* 9.1 (2018), pp. 135–151. DOI: 10.5194/esd-9-135-2018.

Herger, N., G. Abramowitz, S. Sherwood, R. Knutti, O. Angélil, and S. A. Sisson. "Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change". *Climate Dynamics* 53.3 (2019), pp. 1581–1596. DOI: 10.1007/s00382-019-04690-8.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, et al. "The ERA5 global reanalysis". 146.730 (2020), pp. 1999–2049. DOI: 10.1002/qj.3803.

Hoegh-Guldberg, O., D. Jacob, M. Bindi, S. Brown, I. Camilloni, A. Diedhiou, R. Djalante, K. Ebi, F. Engelbrecht, J. Guiot, Y. Hijioka, S. Mehrotra, A. Payne, S. I. Senevirante, A. Thomas, R. Warren, and G. Zhou. "Impacts of 1.5 C global warming on natural and human systems, Chapter 3". In: 2018.

Hossaini, R., E. Atlas, S. S. Dhomse, M. P. Chipperfield, P. F. Bernath, A. M. Fernando, J. Mühle, A. A. Leeson, S. A. Montzka, W. Feng, J. J. Harrison, P. Krummel, M. K. Vollmer, S. Reimann, S. O'Doherty, D. Young, M. Maione, J. Arduini, and C. R. Lunder. "Recent trends in stratospheric chlorine from very short-lived substances". *Journal of Geophysical Research: Atmospheres* 124.4 (2019), pp. 2318–2335. DOI: 10.1029/2018JD029400.

Hossaini, R., M. P. Chipperfield, S. A. Montzka, A. A. Leeson, S. S. Dhomse, and J. A. Pyle. "The increasing threat to stratospheric ozone from dichloromethane". *Nature Communications* 8.1 (2017), pp. 1–9. DOI: 10.1038/ncomms15962.

Hourdin, F., T. Mauritsen, A. Gettelman, J. C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, F. Rauser, C. Rio, L. Tomassini, M. Watanabe, and

D. Williamson. "The art and science of climate model tuning". *Bulletin of the American Meteorological Society* 98.3 (2017), pp. 589–602. DOI: 10.1175/BAMS-D-15-00135.1.

Iglesias-Suarez, F., P. J. Young, and O. Wild. "Stratospheric ozone change and related climate impacts over 1850–2100 as modelled by the ACCMIP ensemble". *Atmospheric Chemistry and Physics* 16.1 (2016), pp. 343–363. DOI: 10.5194/acp-16-343-2016.

Imai, K., N. Manago, C. Mitsuda, Y. Naito, E. Nishimoto, T. Sakazaki, M. Fujiwara, L. Froidevaux, T. von Clarmann, G. P. Stiller, D. P. Murtagh, P.-p. Rong, M. G. Mlynczak, K. A. Walker, D. E. Kinnison, H. Akiyoshi, T. Nakamura, T. Miyasaka, T. Nishibori, S. Mizobuchi, K.-i. Kikuchi, H. Ozeki, C. Takahashi, H. Hayashi, T. Sano, M. Suzuki, M. Takayanagi, and M. Shiotani. "Validation of ozone data from the Superconducting Submillimeter-Wave Limb-Emission Sounder (SMILES)". *Journal of Geophysical Research: Atmospheres* 118.11 (2013), pp. 5750–5769. DOI: 10.1002/jgrd.50434.

IPCC. "Meeting report of the Intergovernmental Panel on Climate Change expert meeting on assessing and combining multi model climate projections". *Intergovernal Panel Climate Change* (2010). Ed. by R. Knutti, G. Abramowitz, M. Collins, V. Eyring, P. J. Gleckler, B. Hewitson, L. Mearns, T. Stocker, Q. Dahe, et al.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. 2013.

IPCC. *IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. 2013.

Ivy, D. J., S. Solomon, N. Calvo, and D. W. Thompson. "Observed connections of Arctic stratospheric ozone extremes to Northern Hemisphere surface climate". *Environmental Research Letters* 12.2 (2017), p. 024004. DOI: 10.1088/1748-9326/aa57a4.

Jöckel, P., H. Tost, A. Pozzer, M. Kunze, O. Kirner, C. A. M. Brenninkmeijer, S. Brinkop, D. S. Cai, C. Dyroff, J. Eckstein, F. Frank, H. Garny, K.-D. Gottschaldt, P. Graf, V. Grewe, A. Kerkweg, B. Kern, S. Matthes, M. Mertens, S. Meul,

M. Neumaier, M. Nützel, S. Oberländer-Hayn, R. Ruhnke, T. Runde, R. Sander, D. Scharffe, and A. Zahn. "Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51". *Geoscientific Model Development* 9.3 (2016), pp. 1153–1200. DOI: 10.5194/gmd-9-1153-2016.

Jöckel, P., A. Kerkweg, A. Pozzer, R. Sander, H. Tost, H. Riede, A. Baumgaertner, S. Gromov, and B. Kern. "Development cycle 2 of the modular earth submodel system (MESSy2)". *Geoscientific Model Development* 3.2 (2010), pp. 717–752. DOI: 10.5194/gmd-3-717-2010.

Jonsson, A. I., J. De Grandpre, V. I. Fomichev, J. C. McConnell, and S. R. Beagley. "Doubled CO_2-induced cooling in the middle atmosphere: Photochemical analysis of the ozone radiative feedback". *Journal of Geophysical Research: Atmospheres* 109.D24 (2004). DOI: 10.1029/2004JD005093.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph. "The NCEP/NCAR 40-year reanalysis project". *Bulletin of the American meteorological Society* 77.3 (1996), pp. 437–472. DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kanevski, M. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press, 2009.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar. "Machine learning for the geosciences: Challenges and opportunities". *IEEE Transactions on Knowledge and Data Engineering* 31.8 (2018), pp. 1544–1554. DOI: 10.1109/TKDE.2018.2861006.

Katzav, J., H. A. Dijkstra, and A. J. de Laat. "Assessing climate model projections: State of the art and philosophical reflections". *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 43.4 (2012), pp. 258–276. DOI: 10.1016/j.shpsb.2012.07.002.

Kay, J. E., C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. M. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, et al. "The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate

change in the presence of internal climate variability". *Bulletin of the American Meteorological Society* 96.8 (2015), pp. 1333–1349.

Keller, C. A. and M. J. Evans. "Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10". *Geoscientific Model Development* 12.3 (2019), pp. 1209–1225. DOI: 10.5194/gmd-12-1209-2019.

Kiesewetter, G., B.-M. Sinnhuber, M. Vountas, M. Weber, and J. P. Burrows. "A long-term stratospheric ozone data set from assimilation of satellite observations: High-latitude ozone anomalies". *Journal of Geophysical Research: Atmospheres* 115.D10 (2010). DOI: 10.1029/2009JD013362.

Knutti, R. "The end of model democracy?" *Climatic Change* 102.3 (2010), pp. 395–404. DOI: 10.1007/s10584-010-9800-2.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl. "Challenges in Combining Projections from Multiple Climate Models". *Journal of Climate* 23.10 (2010), pp. 2739–2758. DOI: 10.1175/2009JCLI3361.1.

Knutti, R., D. Masson, and A. Gettelman. "Climate model genealogy: Generation CMIP5 and how we got there". *Geophysical Research Letters* 40.6 (2013), pp. 1194–1199. DOI: 10.1002/grl.50256.

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring. "A climate model projection weighting scheme accounting for performance and interdependence". *Geophysical Research Letters* 44.4 (2017), pp. 1909–1918. DOI: 10.1002/2016GL072012.

Knutti, R., T. F. Stocker, F. Joos, and G.-K. Plattner. "Probabilistic climate change projections using neural networks". *Climate Dynamics* 21.3-4 (2003), pp. 257–272. DOI: 10.1007/s00382-003-0345-1.

Komhyr, W. D., R. D. Grass, and R. K. Leonard. "Dobson spectrophotometer 83: A standard for total ozone measurements, 1962–1987". *Journal of Geophysical Research: Atmospheres* 94.D7 (1989), pp. 9847–9861. DOI: 10.1029/JD094iD07p09847.

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. "Machine learning applications in cancer prognosis and prediction". *Computational and structural biotechnology journal* 13 (2015), pp. 8–17. DOI: 10.1016/j.csbj.2014.11.005.

Kretschmer, M., J. Runge, and D. Coumou. "Early prediction of extreme stratospheric polar vortex states based on causal precursors". *Geophysical research letters* 44.16 (2017), pp. 8592–8600. DOI: 10.1002/2017GL074696.

Labe, Z. M. and E. A. Barnes. "Detecting climate signals using explainable AI with single-forcing large ensembles". *Journal of Advances in Modeling Earth Systems* (2021). DOI: 10.1029/2021MS002464.

Lakshminarayanan, B., A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: vol. 30. 2017. URL: https://papers.nips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Lamarque, J.-F., T. C. Bond, V. Eyring, C. Granier, A. Heil, Z. Klimont, D. Lee, C. Liousse, A. Mieville, B. Owen, M. G. Schultz, D. Shindell, S. J. Smith, E. Stehfest, J. Van Aardenne, O. R. Cooper, M. Kainuma, N. Mahowald, J. R. McConnell, V. Naik, K. Riahi, and D. P. van Vuuren. "Historical (1850-2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application". *Atmospheric Chemistry and Physics* 10.15 (2010), pp. 7017–7039. DOI: 10.5194/acp-10-7017-2010.

Lamarque, J.-F., F. Dentener, J. McConnell, C.-U. Ro, M. Shaw, R. Vet, D. Bergmann, P. Cameron-Smith, S. Dalsoren, R. Doherty, G. Faluvegi, S. J. Ghan, B. Josse, Y. H. Lee, I. A. MacKenzie, D. Plummer, D. T. Shindell, R. B. Skeie, D. S. Stevenson, S. Strode, G. Zeng, M. Curran, D. Dahl-Jensen, S. Das, D. Fritzsche, and M. Nolan. "Multi-model mean nitrogen and sulfur deposition from the atmospheric chemistry and climate model intercomparison project (ACCMIP): evaluation of historical and projected future". *Atmospheric Chemistry and Physics* 13 (2013). DOI: 10.5194/acp-13-7997-2013.

Lamarque, J.-F., D. T. Shindell, B. Josse, P. J. Young, I. Cionni, V. Eyring, D. Bergmann, P. Cameron-Smith, W. J. Collins, R. Doherty, S. Dalsoren, G. Faluvegi, G. Folberth, S. J. Ghan, L. W. Horowitz, Y. H. Lee, I. A. MacKenzie, T. Nagashima, V. Naik, D. Plummer, M. Righi, S. T. Rumbold, M. Schulz, R. B. Skeie, D. S. Stevenson, S. Strode, K. Sudo, S. Szopa, A. Voulgarakis, and G. Zeng. "The Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP): overview and description of models, simulations and climate

diagnostics". *Geoscientific Model Development* 6.1 (2013), pp. 179–206. DOI: 10.5194/gmd-6-179-2013.

Langematz, U., M. Tully, N. Calvo, M. Dameris, de Laat A.T.J, A. Klekociuk, R. Muller, and P. Young. "Polar stratospheric ozone: past, present, and future, Chapter 4". In: *Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project–Report No. 58*. WMO (World Meteorological Organization), Geneva, Switzerland, 2018. URL: http://ozone.unep.org/science/assessment/sap.

Lee, J., Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. "Deep neural networks as gaussian processes". *arXiv preprint arXiv:1711.00165* (2017). URL: https://arxiv.org/abs/1711.00165.

Lee, M., M. Jun, and M. G. Genton. "Validation of CMIP5 multimodel ensembles through the smoothness of climate variables". *Tellus, Series A: Dynamic Meteorology and Oceanography* 67.1 (2015). DOI: 10.3402/tellusa.v67.23880.

Lin, P., D. Paynter, L. Polvani, G. J. P. Correa, Y. Ming, and V. Ramaswamy. "Dependence of model-simulated response to ozone depletion on stratospheric polar vortex climatology". *Geophysical Research Letters* 44.12 (2017), pp. 6391–6398. DOI: 10.1002/2017GL073862.

Lorenz, R., N. Herger, J. Sedláček, V. Eyring, E. M. Fischer, and R. Knutti. "Prospects and caveats of weighting climate models for summer maximum temperature projections over North America". *Journal of Geophysical Research: Atmospheres* 123.9 (2018), pp. 4509–4526. DOI: 10.1029/2017JD027992.

Maganathan, T., S. Senthilkumar, and V. Balakrishnan. "Machine learning and data analytics for environmental science: a review, prospects and challenges". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 955. 1. IOP Publishing. 2020, p. 012107. DOI: 10.1088/1757-899x/955/1/012107.

Marchand, M., P. Keckhut, S. Lefebvre, C. Claud, D. Cugnet, A. Hauchecorne, F. Lefèvre, M.-P. Lefebvre, J. Jumelet, F. Lott, et al. "Dynamical amplification of the stratospheric solar response simulated with the Chemistry-Climate model LMDz-Reprobus". *Journal of Atmospheric and Solar-Terrestrial Physics* 75 (2012), pp. 147–160. DOI: 10.1016/j.jastp.2011.11.008.

Marsh, D. R., M. J. Mills, D. E. Kinnison, J.-F. Lamarque, N. Calvo, and L. M. Polvani. "Climate change from 1850 to 2005 simulated in CESM1 (WACCM)". *Journal of climate* 26.19 (2013), pp. 7372–7391. DOI: 10.1175/JCLI-D-12-00558.1.

Masson, D. and R. Knutti. "Climate model genealogy". *Geophysical Research Letters* 38.8 (2011). DOI: 10.1029/2011GL046864.

Matthews, A. G. d. G., M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. " GPflow: A Gaussian process library using TensorFlow". *Journal of Machine Learning Research* 18.40 (2017). URL: http://jmlr.org/papers/v18/16-537.html.

McGovern, A., D. J. Gagne, J. K. Williams, R. A. Brown, and J. B. Basara. "Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning". *Machine learning* 95.1 (2014), pp. 27–50. DOI: 10.1007/s10994-013-5343-x.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith. "Making the black box more transparent: Understanding the physical implications of machine learning". *Bulletin of the American Meteorological Society* 100.11 (2019), pp. 2175–2199. DOI: 10.1175/BAMS-D-18-0195.1.

McKenzie, R. L., P. J. Aucamp, A. F. Bais, L. O. Björn, M. Ilyas, and S. Madronich. "Ozone depletion and climate change: impacts on UV radiation". *Photochemical & Photobiological Sciences* 10.2 (2011), pp. 182–198. DOI: 10.1039/c0pp90034f.

McPeters, R. D. and G. J. Labow. "Climatology 2011: An MLS and sonde derived ozone climatology for satellite retrieval algorithms". *Journal of Geophysical Research: Atmospheres* 117.D10 (2012). DOI: 10.1029/2011JD017006.

Mears, C. A. and F. J. Wentz. "Construction of the Remote Sensing Systems V3.2 Atmospheric Temperature Records from the MSU and AMSU Microwave Sounders". *Journal of Atmospheric and Oceanic Technology* 26.6 (2009), pp. 1040–1056. DOI: 10.1175/2008JTECHA1176.1.

Meher, J. K. and L. Das. "Gridded data as a source of missing data replacement in station records". *Journal of Earth System Science* 128.3 (2019), p. 58. DOI: 10.1007/s12040-019-1079-8.

Merrifield, A. L., L. Brunner, R. Lorenz, I. Medhaug, and R. Knutti. "An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles". *Earth System Dynamics* 11.3 (2020), pp. 807–834. DOI: 10.5194/esd-11-807-2020.

Michou, M., D. Saint-Martin, H. Teyssèdre, A. Alias, F. Karcher, D. Olivié, A. Voldoire, B. Josse, V.-H. Peuch, H. Clark, J. N. Lee, and F. Chéroux. "A new version of the CNRM Chemistry-Climate Model, CNRM-CCM: description and improvements from the CCMVal-2 simulations". *Geoscientific Model Development* 4.4 (2011), pp. 873–900. DOI: 10.5194/gmd-4-873-2011.

Miller, A. J., R. M. Nagatani, L. E. Flynn, S. Kondragunta, E. Beach, R. Stolarski, R. D. McPeters, P. K. Bhartia, M. T. DeLand, C. H. Jackman, D. J. Wuebbles, K. O. Patten, and R. P. Cebula. "A cohesive total ozone data set from the SBUV(/2) satellite system". *Journal of Geophysical Research: Atmospheres* 107.D23 (2002), ACH 11-1-ACH 11–8. DOI: 10.1029/2001JD000853.

Molina, L. and M. Molina. "Production of chlorine oxide (Cl2O2) from the self-reaction of the chlorine oxide (ClO) radical". *Journal of Physical Chemistry* 91.2 (1987), pp. 433–436. DOI: 10.1021/j100286a035.

Molina, M. J. and F. S. Rowland. "Stratospheric sink for chlorofluoromethanes: chlorine atom-catalysed destruction of ozone". *Nature* 249.5460 (1974), pp. 810–812. DOI: 10.1038/249810a0.

Monteleoni, C., G. A. Schmidt, S. Saroha, and E. Asplund. "Tracking climate models". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 4.4 (2011), pp. 372–392. DOI: 10.1002/sam.10126.

Montzka, S. A., G. S. Dutton, P. Yu, E. Ray, R. W. Portmann, J. S. Daniel, L. Kuijpers, B. D. Hall, D. Mondeel, C. Siso, et al. "An unexpected and persistent increase in global emissions of ozone-depleting CFC-11". *Nature* 557.7705 (2018), p. 413. DOI: 10.1038/s41586-018-0106-2.

Morgenstern, O., P. Braesicke, F. O'Connor, A. Bushell, C. Johnson, S. Osprey, and J. Pyle. "Evaluation of the new UKCA climate-composition model-Part 1: The stratosphere". *Geoscientific Model Development* 2.1 (2009). DOI: 10.5194/gmd-2-43-2009.

Morgenstern, O., M. A. Giorgetta, K. Shibata, V. Eyring, D. W. Waugh, T. G. Shepherd, H. Akiyoshi, J. Austin, A. J. G. Baumgaertner, S. Bekki, P. Braesicke, C. Brühl, M. P. Chipperfield, D. Cugnet, M. Dameris, S. Dhomse, S. M. Frith, H. Garny, A. Gettelman, S. C. Hardiman, et al. "Review of the formulation of present-generation stratospheric chemistry-climate models and associated external forcings". *Journal of Geophysical Research: Atmospheres* 115.D3 (2010). DOI: 10.1029/2009JD013728.

Morgenstern, O., M. I. Hegglin, E. Rozanov, F. M. O'Connor, N. L. Abraham, H. Akiyoshi, A. T. Archibald, S. Bekki, N. Butchart, M. P. Chipperfield, M. Deushi, S. S. Dhomse, R. R. Garcia, S. C. Hardiman, L. W. Horowitz, P. Jöckel, B. Josse, D. Kinnison, M. Lin, E. Mancini, et al. "Review of the global models used within phase 1 of the Chemistry–Climate Model Initiative (CCMI)". *Geoscientific Model Development* 10.2 (2017), pp. 639–671. DOI: 10.5194/gmd-10-639-2017.

Morgenstern, O., K. A. Stone, R. Schofield, H. Akiyoshi, Y. Yamashita, D. E. Kinnison, R. R. Garcia, K. Sudo, D. A. Plummer, J. Scinocca, L. D. Oman, M. E. Manyin, G. Zeng, E. Rozanov, A. Stenke, L. E. Revell, G. Pitari, E. Mancini, G. Di Genova, D. Visioni, S. S. Dhomse, and M. P. Chipperfield. "Ozone sensitivity to varying greenhouse gases and ozone-depleting substances in CCMI-1 simulations". *Atmospheric Chemistry and Physics* 18.2 (2018), pp. 1091–1114. DOI: 10.5194/acp-18-1091-2018.

Muller, R. "A brief history of stratospheric ozone research". *Meteorologische Zeitschrift* 18.1 (2009), p. 3. DOI: 10.1127/0941-2948/2009/353.

Murgatroyd, R. and F. Singleton. "Possible meridional circulations in the stratosphere and mesosphere". *Quarterly Journal of the Royal Meteorological Society* 87.372 (1961), pp. 125–135. DOI: 10.1002/qj.49708737202.

NCAR. *The Climate Data Guide: Multivariate ENSO Index last accessed: 1 September 2021)*. 2019. URL: https://climatedataguide.ucar.edu/climate-data/multivariate-enso-index.

Neal, R. M. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012. DOI: 10.1007/978-1-4612-0745-0.

Newman, P., J. Daniel, D. Waugh, and E. Nash. "A new formulation of equivalent effective stratospheric chlorine (EESC)". *Atmospheric Chemistry and Physics* 7.17 (2007), pp. 4537–4552. DOI: 10.5194/acp-7-4537-2007.

Newton, R., G. Vaughan, E. Hintsa, M. T. Filus, L. L. Pan, S. Honomichl, E. Atlas, S. J. Andrews, and L. J. Carpenter. "Observations of ozone-poor air in the tropical tropopause layer". *Atmospheric Chemistry and Physics* 18.7 (2018), pp. 5157–5171. DOI: 10.5194/acp-18-5157-2018.

Nicely, J. M., B. N. Duncan, T. F. Hanisco, G. M. Wolfe, R. J. Salawitch, M. Deushi, A. S. Haslerud, P. Jöckel, B. Josse, D. E. Kinnison, A. Klekociuk, M. E. Manyin, V. Marécal, O. Morgenstern, L. T. Murray, G. Myhre, L. D. Oman, G. Pitari, A. Pozzer, I. Quaglia, L. E. Revell, E. Rozanov, A. Stenke, K. Stone, S. Strahan, S. Tilmes, H. Tost, D. M. Westervelt, and G. Zeng. "A machine learning examination of hydroxyl radical differences among model simulations for CCMI-1". *Atmospheric Chemistry and Physics* 20.3 (2020), pp. 1341–1361. DOI: 10.5194/acp-20-1341-2020.

Nowack, P., J. Runge, V. Eyring, and J. D. Haigh. "Causal networks for climate model evaluation and constrained projections". *Nature communications* 11.1 (2020), pp. 1–11. DOI: 10.1038/s41467-020-15195-y.

O'Gorman, P. A. and J. G. Dwyer. "Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events". *Journal of Advances in Modeling Earth Systems* 10.10 (2018), pp. 2548–2563. DOI: 10.1029/2018MS001351.

Oman, L. D., A. R. Douglass, J. R. Ziemke, J. M. Rodriguez, D. W. Waugh, and J. E. Nielsen. "The ozone response to ENSO in Aura satellite measurements and a chemistry-climate simulation". *Journal of Geophysical Research: Atmospheres* 118.2 (2013), pp. 965–976. DOI: 10.1029/2012JD018546.

Orbe, C., D. A. Plummer, D. W. Waugh, H. Yang, P. Jöckel, D. E. Kinnison, B. Josse, V. Marecal, M. Deushi, N. L. Abraham, A. T. Archibald, M. P. Chipperfield, S. Dhomse, W. Feng, and S. Bekki. "Description and Evaluation of the specified-dynamics experiment in the Chemistry-Climate Model Initiative". *Atmospheric Chemistry and Physics* 20.6 (2020), pp. 3809–3840. DOI: 10.5194/acp-20-3809-2020.

Orbe, C., H. Yang, D. W. Waugh, G. Zeng, O. Morgenstern, D. E. Kinnison, J.-F. Lamarque, S. Tilmes, D. A. Plummer, J. F. Scinocca, B. Josse, V. Marecal, P. Jöckel, L. D. Oman, S. E. Strahan, M. Deushi, T. Y. Tanaka, K. Yoshida, H. Akiyoshi, Y. Yamashita, A. Stenke, L. Revell, T. Sukhodolov, E. Rozanov, G. Pitari, D. Visioni, K. A. Stone, R. Schofield, and A. Banerjee. "Large-scale tropospheric transport in the Chemistry–Climate Model Initiative (CCMI) simulations". *Atmospheric Chemistry and Physics* 18.10 (2018), pp. 7217–7235. DOI: 10.5194/acp-18-7217-2018.

Owens, A., J. Steed, C. Miller, D. Filkin, and J. Jesson. "The atmospheric lifetimes of CFC 11 and CFC 12". *Geophysical Research Letters* 9.6 (1982), pp. 700–703. DOI: 10.1029/GL009i006p00700.

Patel, J., S. Shah, P. Thakkar, and K. Kotecha. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". *Expert systems with applications* 42.1 (2015), pp. 259–268. DOI: 10.1016/j.eswa.2014.07.040.

Pathak, J., A. Wikner, R. Fussell, S. Chandra, B. R. Hunt, M. Girvan, and E. Ott. "Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.4 (2018). DOI: doi.org/10.1063/1.5028373.

Pearce, T., F. Leibfried, and A. Brintrup. "Uncertainty in neural networks: Approximately bayesian ensembling". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 234–244. URL: https://proceedings.mlr.press/v108/pearce20a.html.

Pearce, T., R. Tsuchida, M. Zaki, A. Brintrup, and A. Neely. "Expressive priors in Bayesian neural networks: Kernel combinations and periodic functions". In: *Proceedings of the Thirty-fifth Conference on Uncertainty in Artificial Intelligence*. PMLR. 2019. URL: https://proceedings.mlr.press/v115/pearce20a.html.

Pearce, T., M. Zaki, A. Brintrup, N. Anastassacos, and A. Neely. "Uncertainty in neural networks: Bayesian ensembling". *stat* 1050 (2018), p. 12. URL: https://arxiv.org/abs/1810.05546v2.

Perlwitz, J., S. Pawson, R. L. Fogt, J. E. Nielsen, and W. D. Neff. "Impact of stratospheric ozone hole recovery on Antarctic climate". *Geophysical Research Letters* 35.8 (2008). DOI: 10.1029/2008GL033317.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker. "Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models". *Journal of Geophysical Research: Atmospheres* 113.D14 (2008). DOI: 10.1029/2007JD009334.

Polvani, L. M., M. Abalos, R. Garcia, D. Kinnison, and W. J. Randel. "Significant weakening of Brewer-Dobson circulation trends over the 21st century as a consequence of the Montreal Protocol". *Geophysical Research Letters* 45.1 (2018), pp. 401–409. DOI: 10.1002/2017GL075345.

Polvani, L. M., D. W. Waugh, G. J. Correa, and S.-W. Son. "Stratospheric ozone depletion: The main driver of twentieth-century atmospheric circulation changes in the Southern Hemisphere". *Journal of Climate* 24.3 (2011), pp. 795–812. DOI: 10.1175/2010JCLI3772.1.

Portmann, R., J. Daniel, and A. Ravishankara. "Stratospheric ozone depletion due to nitrous oxide: influences of other gases". *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1593 (2012), pp. 1256–1264. DOI: 10.1098/rstb.2011.0377.

Räisänen, J., L. Ruokolainen, and J. Ylhäisi. "Weighting of model results for improving best estimates of climate change". *Climate Dynamics* 35.2 (2010), pp. 407–422. DOI: 10.1007/s00382-009-0659-8.

Randel, W. J., R. R. Garcia, N. Calvo, and D. Marsh. "ENSO influence on zonal mean temperature and ozone in the tropical lower stratosphere". *Geophysical Research Letters* 36.15 (2009). DOI: 10.1029/2009GL039343.

Randel, W. J. and F. Wu. "A stratospheric ozone trends data set for global modeling studies". *Geophysical Research Letters* 26.20 (1999), pp. 3089–3092. DOI: 10.1029/1999GL900615.

Randel, W. J. and F. Wu. "Cooling of the Arctic and Antarctic polar stratospheres due to ozone depletion". *Journal of Climate* 12.5 (1999), pp. 1467–1479. DOI: 10.1175/1520-0442(1999)012<1467:COTAAA>2.0.CO;2.

Randel, W. J. and F. Wu. "A stratospheric ozone profile data set for 1979–2005: Variability, trends, and comparisons with column ozone data". *Journal of Geophysical Research: Atmospheres* 112.D6 (2007). DOI: 10.1029/2006JD007339.

Ratner, B. *Statistical and machine-learning data mining:: Techniques for better predictive modeling and analysis of big data*. CRC Press, 2017. DOI: 10.1201/b11508.

Reichler, T. and J. Kim. "How Well Do Coupled Models Simulate Today's Climate?" *Bulletin of the American Meteorological Society* 89.3 (2008), pp. 303–312. DOI: 10.1175/BAMS-89-3-303.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. "Deep learning and process understanding for data-driven Earth system science". *Nature* 566.7743 (2019), pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

Richardson, L. F. *Weather prediction by numerical process*. Cambridge university press, 1922. DOI: 10.1017/CBO9780511618291.

Rienecker, M. M., M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R. D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C. R. Redder, R. Reichle, F. R. Robertson, A. G. Ruddick, M. Sienkiewicz, and J. Woollen. "MERRA: NASA's modern-era retrospective analysis for research and applications". *Journal of climate* 24.14 (2011), pp. 3624–3648. DOI: 10.1175/JCLI-D-11-00015.1.

Rigby, M., S. Park, T. Saito, L. M. Western, A. L. Redington, X. Fang, S. Henne, A. J. Manning, R. G. Prinn, G. S. Dutton, P. J. Fraser, A. L. Ganesan, B. D. Hall, C. M. Harth, J. Kim, K.-R. Kim, P. B. Krummel, T. Lee, S. Li, Q. Liang, et al. "Increase in CFC-11 emissions from eastern China based on atmospheric observations". *Nature* 569.7757 (2019), p. 546. DOI: 10.1038/s41586-019-1193-4.

Rigby, M., R. G. Prinn, S. O'Doherty, S. A. Montzka, A. McCulloch, C. M. Harth, J. Mühle, P. K. Salameh, R. F. Weiss, D. Young, P. G. Simmonds, B. D. Hall, G. S. Dutton, D. Nance, D. J. Mondeel, J. W. Elkins, P. B. Krummel, L. P. Steele, and P. J. Fraser. "Re-evaluation of the lifetimes of the major CFCs and CH 3 CCl 3 using atmospheric trends". *Atmospheric Chemistry and Physics* 13.5 (2013), pp. 2691–2702. DOI: 10.5194/acp-13-2691-2013.

Rowland, F. S. "Stratospheric ozone depletion". *Twenty Years of Ozone Decline* (2009), pp. 23–66. DOI: 10.1098/rstb.2005.1783.

Ryan, E., O. Wild, A. Voulgarakis, and L. Lee. "Fast sensitivity analysis methods for computationally expensive models with multi-dimensional output". *Geoscientific Model Development* 11.8 (2018), pp. 3131–3146. DOI: 10.5194/gmd-11-3131-2018.

Rybka, H. and H. Tost. "Uncertainties in future climate predictions due to convection parameterisations". *Atmospheric Chemistry and Physics* 14.11 (2014), pp. 5561–5576. DOI: 10.5194/acp-14-5561-2014.

Salawitch, R., G. Gobbi, S. Wofsy, and M. McElroy. "Denitrification in the Antarctic stratosphere". *Nature* 339.6225 (1989), pp. 525–527. DOI: 10.1038/339525a0.

Samek, W., G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019. DOI: 10.1007/978-3-030-28954-6.

Sanderson, B. M., M. Wehner, and R. Knutti. "Skill and independence weighting for multi-model assessments". *Geoscientific Model Development* 10.6 (2017), pp. 2379–2395. DOI: 10.5194/gmd-10-2379-2017.

Sanderson, B. M., R. Knutti, and P. Caldwell. "A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble". *Journal of Climate* 28.13 (2015), pp. 5171–5194. DOI: 10.1175/JCLI-D-14-00362.1.

Sanderson, B. M., R. Knutti, and P. Caldwell. "Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties". *Journal of Climate* 28.13 (2015), pp. 5150–5170. DOI: 10.1175/JCLI-D-14-00361.1.

Sankey, D. and T. G. Shepherd. "Correlations of long-lived chemical species in a middle atmosphere general circulation model". *Journal of Geophysical Research: Atmospheres* 108.D16 (2003). DOI: 10.1029/2002JD002799.

Schneider, T., S. Lan, A. Stuart, and J. Teixeira. "Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations". *Geophysical Research Letters* 44.24 (2017), pp. 12–396. DOI: 10.1002/2017GL076101.

Schoeberl, M. R. and D. L. Hartmann. "The dynamics of the stratospheric polar vortex and its relation to springtime ozone depletions". *Science* 251.4989 (1991), pp. 46–52. DOI: 10.1126/science.251.4989.46.

Schultz, M. G., S. Schröder, O. Lyapina, O. R. Cooper, I. Galbally, I. Petropavlovskikh, E. Von Schneidemesser, H. Tanimoto, Y. Elshorbany, M. Naja, et al. "Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations". *Elementa: Science of the Anthropocene* 5 (2017). DOI: 10.1525/elementa.244.

Schwarz, T. "Mechanisms of UV-induced immunosuppression". *The Keio journal of medicine* 54.4 (2005), pp. 165–171. DOI: 10.2302/kjm.54.165.

Scinocca, J., N. McFarlane, M. Lazare, J. Li, and D. Plummer. "The CCCma third generation AGCM and its extension into the middle atmosphere". *Atmospheric Chemistry and Physics* 8.23 (2008), pp. 7055–7074. DOI: 10.5194/acp-8-7055-2008.

Scinocca, J. F., D. B. Stephenson, T. C. Bailey, and J. Austin. "Estimates of past and future ozone trends from multimodel simulations using a flexible smoothing spline methodology". *Journal of Geophysical Research: Atmospheres* 115.D3 (2010). DOI: 10.1029/2009JD013622.

Sekiya, T. and K. Sudo. "Roles of transport and chemistry processes in global ozone change on interannual and multidecadal time scales". *Journal of Geophysical Research: Atmospheres* 119.8 (2014), pp. 4903–4921. DOI: 10.1002/2013JD020838.

Sekiya, T. and K. Sudo. "Role of meteorological variability in global tropospheric ozone during 1970–2008". *Journal of Geophysical Research: Atmospheres* 117.D18 (2012). DOI: 10.1029/2012JD018054.

Sengupta, U., M. Amos, J. S. Hosking, C. E. Rasmussen, M. Juniper, and P. J. Young. "Ensembling geophysical models with Bayesian neural networks". *Advances in Neural Information Processing Systems* 33 (2020). URL: https://papers.nips.cc/paper/2020/file/0d5501edb21a59a43435efa67f200828-Paper.pdf.

Shepherd, T. G. "Dynamics, stratospheric ozone, and climate change". *Atmosphere-Ocean* 46.1 (2008), pp. 117–138. DOI: 10.3137/ao.460106.

"Signatures of the Antarctic ozone hole in Southern Hemisphere surface climate change". *Nature geoscience* 4.11 (2011), p. 741. DOI: 10.1038/ngeo1296.

Solomon, S. "Stratospheric ozone depletion: A review of concepts and history". *Reviews of Geophysics* 37.3 (1999), pp. 275–316. DOI: 10.1029/1999RG900008.

Solomon, S., R. R. Garcia, F. S. Rowland, and D. J. Wuebbles. "On the depletion of Antarctic ozone". *Nature* 321.6072 (1986), pp. 755–758. DOI: 10.1038/321755a0.

Solomon, S., D. J. Ivy, D. Kinnison, M. J. Mills, R. R. Neely, and A. Schmidt. "Emergence of healing in the Antarctic ozone layer". *Science* 353.6296 (2016), pp. 269–274. DOI: 10.1126/science.aae0061.

Solomon, S., D. Kinnison, J. Bandoro, and R. Garcia. "Simulation of polar ozone depletion: An update". *Journal of Geophysical Research: Atmospheres* 120.15 (2015), pp. 7958–7974. DOI: 10.1002/2015JD023365.

Solomon, S., R. W. Portmann, and D. W. Thompson. "Contrasts between Antarctic and Arctic ozone depletion". *Proceedings of the National Academy of Sciences* 104.2 (2007), pp. 445–449. DOI: 10.1073/pnas.0604895104.

Son, S.-W., L. M. Polvani, D. W. Waugh, H. Akiyoshi, R. Garcia, D. Kinnison, S. Pawson, E. Rozanov, T. G. Shepherd, and K. Shibata. "The Impact of Stratospheric Ozone Recovery on the Southern Hemisphere Westerly Jet". *Science* 320.5882 (2008), pp. 1486–1489. DOI: 10.1126/science.1155939.

SPARC/IO3C/GAW. *SPARC/IO3C/GAW Report on Long-term Ozone Trends and Uncertainties in the Stratosphere*. Ed. by I. Petropavlovskikh, S. Godin-Beekmann, D. Hubert, R. Damadeo, B. Hassler, and V. Sofieva. SPARC/IO3C/GAW, 2018. DOI: 10.17874/f899e57a20b.

Stolarski, R. S. and R. J. Cicerone. "Stratospheric chlorine: a possible sink for ozone". *Canadian journal of Chemistry* 52.8 (1974), pp. 1610–1615. DOI: 10.1139/v74-233.

Struthers, H., G. Bodeker, J. Austin, S. Bekki, I. Cionni, M. Dameris, M. Giorgetta, V. Grewe, F. Lefèvre, F. Lott, E. Manzini, T. Peter, E. Rozanov, and M. Schraner. "The simulation of the Antarctic ozone hole by chemistry-climate models". *Atmospheric Chemistry and Physics* 9.17 (2009), pp. 6363–6376. DOI: 10.5194/acp-9-6363-2009.

Sudo, K. and H. Akimoto. "Global source attribution of tropospheric ozone: Long-range transport from various source regions". *Journal of Geophysical Research: Atmospheres* 112.D12 (2007). DOI: 10.1029/2006JD007992.

Sudo, K., M. Takahashi, J.-i. Kurokawa, and H. Akimoto. "CHASER: A global chemical model of the troposphere 1. Model description". *Journal of Geophysical Research: Atmospheres* 107.D17 (2002), ACH–7. DOI: 10.1029/2001JD001113.

Szopa, S., Y. Balkanski, M. Schulz, S. Bekki, D. Cugnet, A. Fortems-Cheiney, S. Turquety, A. Cozic, C. Déandreis, D. Hauglustaine, A. Idelkadi, J. Lathière, F. Lefevre, M. Marchand, R. Vuolo, N. Yan, and J.-L. Dufresne. "Aerosol and ozone changes as forcing for climate evolution between 1850 and 2100". *Climate dynamics* 40.9-10 (2013), pp. 2223–2250. DOI: 10.1007/s00382-012-1408-y.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl. "An overview of CMIP5 and the experiment design". *Bulletin of the American Meteorological Society* 93.4 (2012), pp. 485–498. DOI: 10.1175/BAMS-D-11-00094.1.

Tebaldi, C. and R. Knutti. "The use of the multi-model ensemble in probabilistic climate projections". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1857 (2007), pp. 2053–2075. DOI: 10.1098/rsta.2007.2076.

Tebaldi, C., R. L. Smith, D. Nychka, and L. O. Mearns. "Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles". *Journal of Climate* 18.10 (2005), pp. 1524–1540. DOI: 10.1175/JCLI3363.1.

Tegtmeier, S., M. I. Hegglin, J. Anderson, A. Bourassa, S. Brohede, D. Degenstein, L. Froidevaux, R. Fuller, B. Funke, J. Gille, A. Jones, Y. Kasai, K. Krüger, E. Kyrölä, G. Lingenfelser, J. Lumpe, B. Nardi, J. Neu, D. Pendlebury, E. Remsberg, A. Rozanov, L. Smith, M. Toohey, J. Urban, T. von Clarmann, K. A. Walker, and R. H. J. Wang. "SPARC Data Initiative: A comparison of ozone climatologies from international satellite limb sounders". *Journal of Geophysical Research: Atmospheres* 118.21 (2013). DOI: 10.1002/2013JD019877.

Thomason, L. W., N. Ernest, L. Millán, L. Rieger, A. Bourassa, J.-P. Vernier, G. Manney, B. Luo, F. Arfeuille, and T. Peter. "A global space-based stratospheric aerosol climatology: 1979–2016". *Earth System Science Data* 10.1 (2018), pp. 469–492. DOI: 10.5194/essd-10-469-2018.

Thompson, D. W. J. and S. Solomon. "Understanding recent stratospheric climate change". *Journal of Climate* 22.8 (2008), pp. 1934–1943. DOI: 10 . 1175 / 2008JCLI2482.1.

Thompson, D. W. and S. Solomon. "Interpretation of recent Southern Hemisphere climate change". *Science* 296.5569 (2002), pp. 895–899. DOI: 10.1126/science.1069270.

Tilmes, S., J.-F. Lamarque, L. K. Emmons, D. E. Kinnison, P.-L. Ma, X. Liu, S. Ghan, C. Bardeen, S. Arnold, M. Deeter, F. Vitt, T. Ryerson, J. W. Elkins, F. Moore, J. R. Spackman, and M. Val Martin. "Description and evaluation of tropospheric chemistry and aerosols in the Community Earth System Model (CESM1. 2)". *Geoscientific Model Development* 8 (2015), pp. 1395–1426. DOI: 10.5194/gmd-8-1395-2015.

Toon, O. B., P. Hamill, R. P. Turco, and J. Pinto. "Condensation of HNO3 and HCl in the winter polar stratospheres". *Geophysical Research Letters* 13.12 (1986), pp. 1284–1287. DOI: 10.1029/GL013i012p01284.

Velders, G. J., S. O. Andersen, J. S. Daniel, D. W. Fahey, and M. McFarland. "The importance of the Montreal Protocol in protecting climate". *Proceedings of the National Academy of Sciences* 104.12 (2007), pp. 4814–4819. DOI: 10.1073/pnas.0610328104.

Virts, K., A. Shirey, G. Priftis, K. Ankur, M. Ramasubramanian, H. Muhammad, A. Acharya, and R. Ramachandran. "A quantitative analysis on the use of supervised machine learning in Earth science". In: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2020, pp. 2252–2255. DOI: 10.1109/IGARSS39084.2020.9323770.

Viste, E., D. Korecha, and A. Sorteberg. "Recent drought and precipitation tendencies in Ethiopia". *Theoretical and Applied Climatology* 112.3 (2013), pp. 535–551. DOI: 10.1007/s00704-012-0746-3.

Voldoire, A., E. Sanchez-Gomez, D. Salas y Mélia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, M. Déqué, J. Deshayes, H. Douville, E. Fernandez, G. Madec, E. Maisonnave, M.-P. Moine, S. Planton, D. Saint-Martin, S. Szopa, S. Tyteca, R. Alkama, S. Belamari, A. Braun, L. Coquart, and F. Chauvin. "The CNRM-CM5. 1 global climate model: description and basic evaluation". *Climate Dynamics* 40.9-10 (2013), pp. 2091–2121. DOI: 10.1007/s00382-011-1259-y.

Vuuren, D. P. van, J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. J. Smith, and S. K. Rose. "The representative concentration pathways: an overview". *Climatic Change* 109.1 (2011), p. 5. DOI: 10.1007/s10584-011-0148-z.

Wang, S. S.-C., Y. Qian, L. R. Leung, and Y. Zhang. "Identifying key drivers of wildfires in the contiguous US using machine learning and game theory interpretation". *Earth's Future* (2021). DOI: 10.1029/2020EF001910.

Wardah, T., S. Y. Sharifah Nurul Huda, S. M. Deni, and B. Nur Azwa. "Radar rainfall estimates comparison with kriging interpolation of gauged rain". In: *2011 IEEE Colloquium on Humanities, Science and Engineering*. 2011, pp. 93–97. DOI: 10.1109/CHUSER.2011.6163877.

Watanabe, S., T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, T. Ise, H. Sato, E. Kato, K. Takata, S. Emori, and M. Kawamiya. "MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments". *Geoscientific Model Development* 4.4 (2011), p. 845. DOI: 10.5194/gmd-4-845-2011.

Waters, J., L. Froidevaux, R. Harwood, R. Jarnot, H. Pickett, W. Read, P. Siegel, R. Cofield, M. Filipiak, D. Flower, J. Holden, G. Lau, N. Livesey, G. Manney, H. Pumphrey, M. Santee, D. Wu, D. Cuddy, R. Lay, M. Loo, et al. "The Earth observing system microwave limb sounder (EOS MLS) on the aura Satellite". *IEEE Transactions on Geoscience and Remote Sensing* 44.5 (2006), pp. 1075–1092. DOI: 10.1109/TGRS.2006.873771.

Waugh, D. W. and V. Eyring. "Quantitative performance metrics for stratospheric-resolving chemistry-climate models". *Atmospheric Chemistry and Physics* 8.18 (2008), pp. 5699–5713. DOI: 10.5194/acp-8-5699-2008.

Weber, J., S. Archer-Nicholls, N. L. Abraham, Y. M. Shin, T. J. Bannan, C. J. Percival, A. Bacak, P. Artaxo, M. Jenkin, M. A. H. Khan, D. E. Shallcross, R. H. Schwantes, J. Williams, and A. T. Archibald. "Improvements to the representation of BVOC chemistry–climate interactions in UKCA (v11. 5) with the CRI-Strat 2 mechanism: incorporation and evaluation". *Geoscientific Model Development* 14.8 (2021), pp. 5239–5268. DOI: 10.5194/gmd-14-5239-2021.

Weber, M., M. Coldewey-Egbers, V. E. Fioletov, S. M. Frith, J. D. Wild, J. P. Burrows, C. S. Long, and D. Loyola. "Total ozone trends from 1979 to 2016 derived from five merged observational datasets–the emergence into ozone recovery". *Atmospheric Chemistry and Physics* 18.3 (2018), pp. 2097–2117. DOI: 10.5194/acp-18-2097-2018.

Weber, M., S. Dikty, J. P. Burrows, H. Garny, M. Dameris, A. Kubin, J. Abalichin, and U. Langematz. "The Brewer-Dobson circulation and total ozone from seasonal to decadal time scales". *Atmospheric Chemistry and Physics* 11.21 (2011), pp. 11221–11235. DOI: 10.5194/acp-11-11221-2011.

Wit, T. D. de, S. Bruinsma, and K. Shibasaki. "Synoptic radio observations as proxies for upper atmosphere modelling". *Journal of Space Weather and Space Climate* 4 (2014), A06. DOI: 10.1051/swsc/2014003.

Witte, J. C., A. M. Thompson, H. G. Smit, M. Fujiwara, F. Posny, G. J. Coetzee, E. T. Northam, B. J. Johnson, C. W. Sterling, M. Mohamad, S.-Y. Ogino, A. Jordan, and F. R. da Silva. "First reprocessing of Southern Hemisphere ADditional OZonesondes (SHADOZ) profile records (1998–2015): 1. Methodology and evaluation". *Journal of Geophysical Research: Atmospheres* 122.12 (2017), pp. 6611–6636. DOI: 10.1002/2016JD026403.

WMO. *Scientific Assessment of Ozone Depletion: 2010, Global Ozone Research and Monitoring Project-Report No. 52*. Geneva, Switzerland: WMO, 2011.

WMO. *Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project-Report No. 58*. Geneva, Switzerland: WMO, 2018.

Xu, L., N. Chen, X. Zhang, Z. Chen, C. Hu, and C. Wang. "Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning". *Climate dynamics* 53.1-2 (2019), pp. 601–615. DOI: 10.1007/s00382-018-04605-z.

Yang, J. and M. Hu. "Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation". *Science of the Total Environment* 633 (2018), pp. 677–683. DOI: 10.1016/j.scitotenv.2018.03.202.

Young, P. J., V. Naik, A. M. Fiore, A. Gaudel, J. Guo, M. Y. Lin, J. L. Neu, D. D. Parrish, H. E. Rieder, J. L. Schnell, S. Tilmes, O. Wild, L. Zhang, J. Ziemke, J. Brandt, A. Delcloo, R. M. Doherty, C. Geels, M. I. Hegglin, L. Hu,

U. Im, R. Kumar, A. Luhar, L. Murray, D. Plummer, J. Rodriguez, A. Saiz-Lopez, M. G. Schultz, M. T. Woodhouse, and G. Zeng. "Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends". *Elementa: Science of the Anthropocene* 6.1 (2018). DOI: 10.1525/elementa.265.

Young, P. J., A. H. Butler, N. Calvo, L. Haimberger, P. J. Kushner, D. R. Marsh, W. J. Randel, and K. H. Rosenlof. "Agreement in late twentieth century Southern Hemisphere stratospheric temperature trends in observations and CCMVal-2, CMIP3, and CMIP5 models". *Journal of Geophysical Research: Atmospheres* 118.2 (2013), pp. 605–613. DOI: 10.1002/jgrd.50126.

Young, P. J., A. B. Harper, C. Huntingford, N. D. Paul, O. Morgenstern, P. A. Newman, L. D. Oman, S. Madronich, and R. R. Garcia. "The Montreal Protocol protects the terrestrial carbon sink". *Nature* 596.7872 (2021), pp. 384–388. DOI: doi.org/10.1038/s41586-021-03737-3.

Young, P. J., K. H. Rosenlof, S. Solomon, S. C. Sherwood, Q. Fu, and J.-F. Lamarque. "Changes in stratospheric temperatures and their implications for changes in the Brewer–Dobson circulation, 1979–2005". *Journal of Climate* 25.5 (2012), pp. 1759–1772. DOI: 10.1175/2011JCLI4048.1.

Yukimoto, S. *Meteorological research institute earth system model version 1 (MRI-ESM1): model description*. Technical reports of the Meteorological Research Institute, 2011. DOI: 10.11483/mritechrepo.64.

Yukimoto, S., Y. Adachi, M. Hosaka, T. Sakami, H. Yoshimura, M. Hirabara, T. Y. Tanaka, E. Shindo, H. Tsujino, M. Deushi, R. Mizuta, S. Yabu, A. Obata, H. Nakano, T. Koshiro, T. Ose, and A. Kitoh. "A new global climate model of the Meteorological Research Institute: MRI-CGCM3—model description and basic performance". *Journal of the Meteorological Society of Japan. Ser. II* 90 (2012), pp. 23–64. DOI: 10.2151/jmsj.2012-A02.