Norwegian University
of Life Sciences

Master's Thesis 2024   30 ECTS
Faculty of Science and Technology

A Data-Driven Exploration: Providing Early Feedback on Socioeconomic Mitigation Strategies for Climate Change

Johannes Larsen Fjeldså
Environmental Physics and Renewable Energy

Abstract

The current rate of climate change is unprecedented in millennia and represents one of humanity's most significant issues for the 21st century, with serious consequences for both civilization and the environment if no action is taken. To reduce the future impact of changes, coordinated international efforts such as those outlined in the Paris Agreement are required. However, mitigation and adaptation efforts should not only be extensive, but also be precise in order to align international socioeconomic development with temperature targets as well as other societal and environmental sustainability goals. In order to inform about the future climate, extensive efforts through institutions such as the Intergovernmental Panel on Climate Change (IPCC) have resulted in extensive scenario-based research with Earth system models (ESMs). This work forms the knowledge basis for information provided to the worlds decision makers.

Currently, established methods for near-term feedback on mitigation efforts are mainly based on the global surface air temperature (GSAT) variable alone. By these methods, a clear separation of socioeconomic pathways does not emerge before 20 to 30 years after emission separation due to the internal variability of the climate system. Here, we use a machine learning approach to create a separation of the climatic response from the socioeconomic development pathways based on ESM output data. This policy feedback strategy has not been described previously. Using 40 realizations of ACCESS-ESM1.5 under SSP1-2.6 and SSP5-8.5, where emission starts to differ in 2015, we estimate that a classification accuracy above 80 % is attainable by the appropriate feature-set/model combinations as early as 2026 based on the mean accuracy across 50 random states. However, the uncertainty of estimated accuracy is greatly reduced towards 2030–2040, indicating that real-world applications are not yet attainable. Our findings suggest that classification models trained on ESM-forecasts have the potential to become a powerful tool for providing early feedback on how the climate system responds to mitigation efforts.

Sammendrag

Den observerte hastigheten av klimaendringer er uten sidestykke de siste 2000 årene og representerer en av menneskehetens største utfordringer for det 21. århundre. Uten tiltak vil klimaendringene rask ha betydelige konsekvenser for både sivilisasjon og miljø globalt. Parisavtalen skisserer en helt nødvendig koordineringen av internasjonal innsats for å redusere konsekvensene for framtidige generasjoner. I tillegg til å være omfattende må avbøtende tiltak også være effektiv for å tilpasse internasjonal sosioøkonomisk utvikling til klimamål så vel som andre samfunnsmessige og miljømessige bærekraftsmål. For å forstå det fremtidige klimaet gjennomføres i dag omfattende scenariobasert forskning med jord-system modeller (ESMer) motivert at institusjoner som FNs klimapanel (IPCC). Dette danner grunnlaget for informasjon brukt av verdens beslutningstakere.

De tidligere etablerte metodene for rask tilbakemelding på avbøtende tiltak er hovedsakelig basert på framskrivninger av global gjennomsnittlig lufttemperatur (GSAT) alene. Ved disse metodene oppstår det ikke et klart skille mellom sosioøkonomiske utviklingsveier før 20 til 30 år etter utslippsseparasjon på grunn av klimasystemets interne variabilitet. I denne oppgaven bruker vi en maskinlæringstilnærming for å skape en separasjon av klimaresponsen fra de sosioøkonomiske utviklingsbanene basert på data fra ESM simuleringer. Denne tilbakemeldingsstrategien har ikke blitt beskrevet tidligere. Ved å bruke 40 realiseringer av ACCESS-ESM1.5 under SSP1-2.6 og SSP5-8.5, der utslippene blir differensiert i 2015, estimerer vi at en klassifiseringsnøyaktighet over 80% kan oppnås med passende variabelsett og modellkombinasjon allerede i 2026 basert på gjennomsnittlig nøyaktighet over 50 tilfeldige tilstander. Usikkerheten til den estimert nøyaktighet er imidlertid sterkt redusert mot 2030–2040, noe som indikerer at applikasjoner på virkelig data ennå ikke er oppnåelige. Våre funn tyder på at klassifiseringsmodeller trent på ESM-framskrivninger har potensial til å bli et kraftig verktøy for rask tilbakemelding på klimasystemets respons på avbøtende tiltak.

Acknowledgments

The purpose of a master's thesis is to prove that the student has gained the necessary skills and knowledge to practice in their chosen field at a professional level - ChatGPT 3.5 (09.05.24). Throughout the work on this thesis, as well as trough my earlier life, I have received a great deal of support and assistance which now allows me to present:

"*A Data-Driven Exploration: Providing Early Feedback on Socioeconomic Mitigation Strategies for Climate Change*"

Firstly, I would like to thank my supervisor Prof. Bjørn Hallvard Samset from NMBU who has helped and guided the whole thesis process. Thank you for providing a fun and insightful topic and thank you for providing an excellent stepping stone into the professional world of science. I am forever grateful for the opportunities to learn and develop that you have provided. I also thank Dr. Mareile Astrid Wolff of NMBU for providing the first introduction to meteorology and climate physics; it is your "fault" that soon (hopefully) refer to myself as a physicist and developed a true passion for science. Finally, I want to thank the other educators at NMBU and UiO. Your efforts have far-reaching consequences. For me and many others, you have created the groundwork for lifelong learning, exciting jobs, and the ability to deliver critical services to society.

In addition, I want to thank my friends from NMBU and elsewhere for fun times with food, laughter, and recreation in nature, which are all essentials of life, best enjoyed with you. To my family, thank you for all the knowledge and happiness you have provided, I am proud to have our love for nature as one of the strong common passions. Lastly, my Mathilde, who deserves endless gratitude: Thank you for being a constant source of motivation and strength. Writing this thesis while smiling would not have been possible without your unconditional love and support.

Sincerely, Johannes Larsen Fjeldså (12.05.24)

The difficulty lies in the very expression "relation to the world",
which presupposes two sorts of domains, that of nature and that of culture,
domains that are at once distinct and impossible to separate completely.

– Bruno Latour (06.10.15)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Research Question

The climate of the Earth is currently in a state of rapid change [1]. The mean state of climate, which includes temperature, precipitation, and many other factors we commonly think of as weather, is, at large, governed by solar irradiation, the rotational and orbital properties of the Earth, as well as complex interactions between the atmosphere, ocean, cryosphere, and biosphere [2]. A climate system driver represents a modification that shifts the mean state by changing the energy equilibrium, resulting in warming or cooling effects [1]. Perturbations from this mean state are also a matter of internally generated and externally forced climate variability. Thus, *climate change* is by no means new on a geological time scale. However, throughout the 20th and early 21st centuries, there has been a marked increase in global mean surface temperature (GMST), with the observed rate of GMST change over the last 50 years being unprecedented in at least the last 2000 years. This warming trend has resulted in each of the last four decades being successively warmer than the preceding decades since the preindustrial levels of 1850 [1]. As a response to this warming, we observed strong changes in the cryosphere and some, less certain, increments in precipitation. These changes in climate have already led to a change in living conditions as well as an increase in the density and severity of extreme events beyond natural variability [3]. The evident acceleration in climate change after 1970 has led to greater effort in research and, consequently, in understanding the earth system (ES). Perhaps the most fundamental question of whether human activities affect the climate (anthropogenic climate change (ACC)) has transitioned from being "suspected" in First IPCC Assessment Report (published 1990) (FAR) to being recognized as an "established fact" in Assessment Report 6 (published 2023) (AR6), thus marking a change in the climatic history of the Earth [4].

The Paris agreement, which entered into force in November 2016, aims to limit global warming to well below 2.0 °**C** above preindustrial levels, with efforts to pursue a target of 1.5 °**C** [5]. To achieve this, global greenhouse gas (GHG) emissions must peak by 2020 and then decline rapidly, reaching net zero emissions by the mid-century [6, 7]. In the absence of enhanced policies beyond those of 2020, it is expected that emissions will increase, resulting in an average global temperature increase of 3.2 °**C** by the year 2100. Transitioning quickly to green technologies and resources is imperative to align with the Paris Agreement's temperature targets, as projected cumulative emissions from existing, and planned, fossil fuel infrastructure exceed what is permissible for limiting warming to 1.5 °**C** and are approximately equal to the levels needed to limit warming to 2.0 °**C** [7].

As we are currently adapting to the impact of ACC on nature and society, it becomes increasingly evident that the current adaptations are made with respect to the historical climate of the past and present, and not the climate of the future [8]. Projections of future near-surface air temperature (tas) and precipitation (pr), based on a large number of climate model simulations (as shown in Figure 1.1) point toward future amplification of climate change. Due to the complexity of international society, there are a variety of potential pathways leading to markedly diverse living conditions. Therefore, understanding these pathways is crucial to formulating effective mitigation policies and adaptation strategies. To aid in this process of decision making, the science community seeks tools that confidently inform on the path to which international society accrues [1]. To form a common framework for investigations of future climate the shared socioeconomic pathway (SSP) scenario framework was established in the work on AR6 by Coupled Model Intercomparison Project (CMIP). Despite extensive investigations of projections of indicators of climate change, a clear response to mitigation efforts, such as a decrease in the level of emissions, takes between 20 and 30 years to observe due to internal variability in the climate system [9]. This delay complicates efforts to understand the immediate impact of mitigation policies and therefore requires long-term planning and monitoring.

This thesis aims to further develop the ability of science to provide fast feedback on the climate impact of global policies. Concerning this issue, we address the following research question:

- How fast can one, with easy-to-access response variables, provide an accurate classification of socioeconomic pathways?

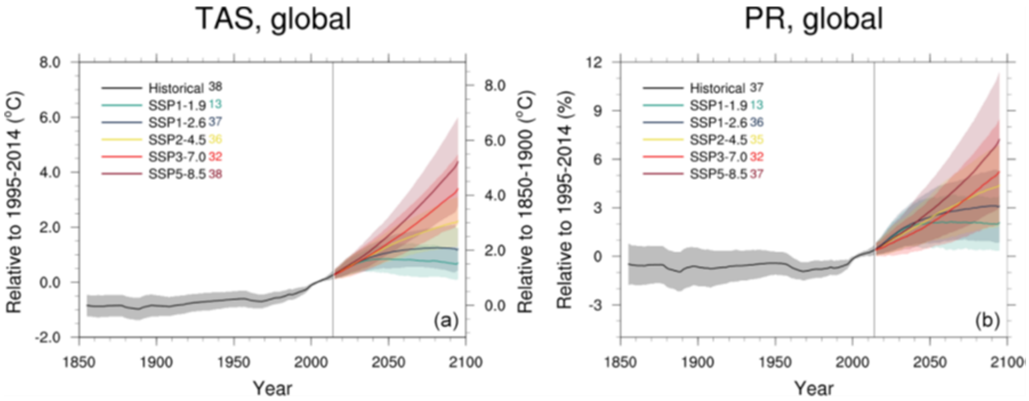


Figure 1.1: Time-series of projected changes in temperature and precipitation under ScenarioMIP scenarios. a) Global average temperature time series (11-year running averages) of changes from current baseline (1995–2014, left axis) and preindustrial baseline (1850–1900, right axis, obtained by adding a 0.84 °**C** offset) for SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5. (b) Global average precipitation time series (11-year running averages) of percent changes from the current baseline (1995-2014) for SSP1-1.9, SSP1-2.6, SSP2-4.5, SSP3-7.0 and SSP5-8.5. Thick lines are ensemble means (number of models shown in the legends). The shading represents the $\pm 1.64\sigma$ interval, where σ is the standard deviation of the smoothed trajectories computed year by year (thus approximating the 5%–95% confidence interval around the mean of a normal distribution). Note that the uncertainty bands are computed for the anomalies with respect to the historical baseline (1995–2014). Figure and description from Tebaldi et al. (2021) [9].

1.2 Approach, methodology, and thesis limitations

This thesis focuses on using machine learning (ML) methods and algorithms to evaluate the capability of SSPs separation. In order to obtain this, the appropriate architecture for this ML-pipeline is:

1. Detection of appropriate response variables for which to train models and to use for model testing.
2. Geographical masking to detect high signal to noise ratio (SNR)-areas within each variable.
3. Feature selection.
4. Training and evaluation of ML models on data from Earth system models (ESMs).

The goal is to provide a proof of concept for training ML classification models on ESM data. Due to time restrictions, some limitations will be applied:

1. We will only use output data from one ESM.
2. We will only use the simulations from two SSPs.
3. We will only apply pre-deep-learning algorithms for the classification.
4. We will not evaluate the generalization abilities based on in-situ observations.

1.3 Outline

In Chapter 2, the notion of Earth system science (ESS) and the usage of ESMs in climate science are introduced to better understand the validity of the data origin. Furthermore, the role of response variables as indicators of climate change is discussed. In Chapter 3, the method for data pre-processing and the theory for feature selection and ML algorithms are introduced. Hereunder are the general objectives of the classification task, the mathematical background of the models, and finally the model evaluation. In Chapter 4, the results are presented in parallel with the appropriate discussion. To round of, Chapter 5 gives a brief summary and conclusions before presenting some suggestions for future work. The appendices are available Appendix A.2.1-A.6. For a more detailed outline we refer to the table of contents on page iv-v.

Chapter 2

Background

2.1 The mean state and internal variability of climate

In the earth system (ES), atmospheric processes are the most influential in determining the properties of the Earth's climate [2]. Among the most important services are the atmosphere's regulation of solar irradiation, balancing of the energy and momentum budget, and distribution of freshwater in the hydrological cycle. However, as illustrated in Figure 2.1, the ES is a complex interrelated system consisting of biotic and abiotic factors that are in constant state of change. These components include, but are not limited to: The ocean, which covers 72 % of the Earth's surface, not only serves as a vital force in maintaining energy and momentum balance through wind-driven and thermohaline circulation but also interacts with the atmosphere, shaping the planet's wet and dry zones. This intricate relationship underscores the importance of the ocean as the main driver of most modes of variation in the climate system [1]. The croyosphere refers to Earth's frozen bodies of water, including but not limited to sea ice and glaciers/ice sheets. The main effects of the cryosphere on the ES include the regulation of the radiation budget through surface albedo and effects such as alteration of global sea level and thermohaline by melting of ice sheets and sea ice, respectively. Climate regulates the geographical distribution of biomes; accordingly, the terrestrial biosphere is a key component of the ES due to its significant influence on the living conditions of humans and other species. On the other hand, the impact on climate by the terrestrial biosphere includes regulation of the radiation budget through surface albedo, damping of winds in the boundary layer, and last but not least, strong regulation of the hydrological cycle.

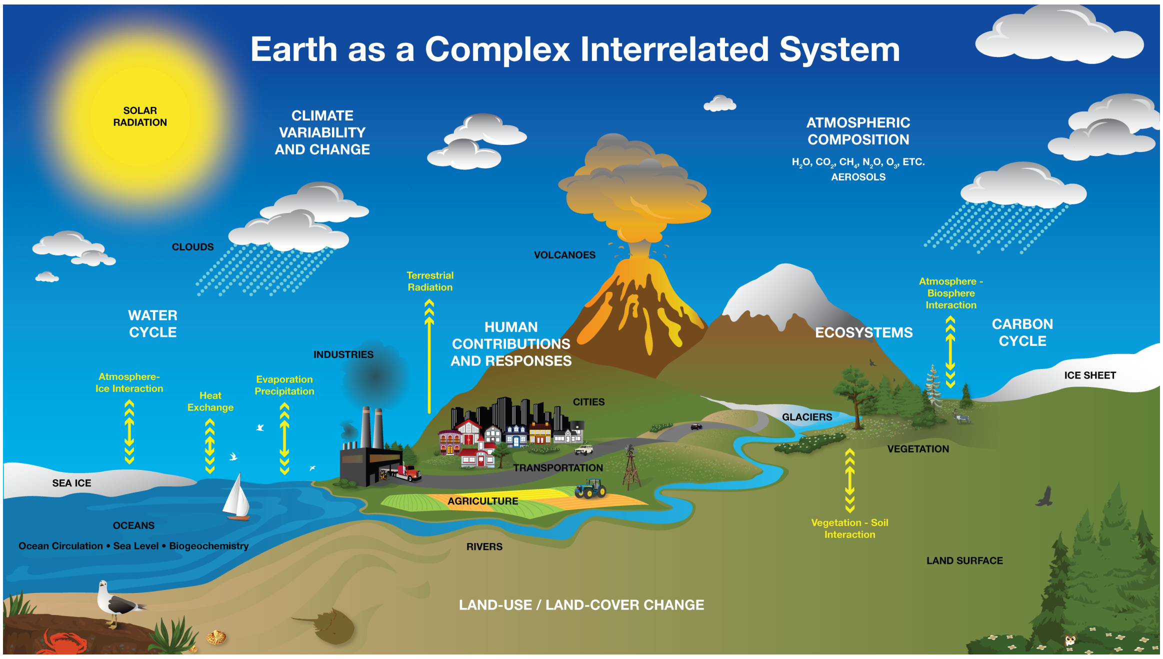

Earth as a Complex Interrelated System
SOLAR RADIATION
CLIMATE VARIABILITY AND CHANGE
ATMOSPHERIC COMPOSITION
H_2O, CO_2, CH_4, N_2O, O_3, ETC.
AEROSOLS
CLOUDS
VOLCANOES
Terrestrial Radiation
Atmosphere - Biosphere Interaction
CARBON CYCLE
WATER CYCLE
HUMAN CONTRIBUTIONS AND RESPONSES
ECOSYSTEMS
INDUSTRIES
ICE SHEET
Atmosphere-Ice Interaction
Heat Exchange
Evaporation Precipitation
CITIES
GLACIERS
VEGETATION
TRANSPORTATION
SEA ICE
AGRICULTURE
OCEANS
Ocean Circulation • Sea Level • Biogeochemistry
RIVERS
Vegetation - Soil Interaction
LAND SURFACE
LAND-USE / LAND-COVER CHANGE

Figure 2.1: The earth system is a complex interrelated system consisting of a vast array of biotic and abiotic factors. Together, they govern the state of the climate both locally and globally. In modern times, humans have started affecting the ES to a greater extent. Illustration from NASA [10].

Together, all the components of the ES are what balance the climate. In addition, all components hold properties that generate year-to-year perturbations from the mean state, either by themselves or by coupling with other components. This is referred to as *the internal variability of the climate system*. Principal modes of variability such as El Niño-Southern Oscillation (ENSO), the Indian Ocean Dipole (IDO), and the North Atlantic Oscillation (NAO) are examples that accentuate the complexity and high degree of interraledability in the ES [1, 2].

To understand the past and future evolution of the state of the ES, it is crucial to gain knowledge about the interactions between the various components. At the heart of these studies lies the concept of Earth system science (ESS), which has facilitated the transition from traditional examinations of individual elements within the ES to a more comprehensive understanding of our planet [4]. This knowledge expansion is a multifactorial process whose main tools include enhanced global cooperation through establishments such as the Intergovernmental Panel on Climate Change (IPCC) and CMIP, intensified in-situ surveillance through meteorological stations, remote sensing, and climate proxy analysis, and last but not least, a significant increase in the application of ESMs [1]. In fact, much of what we know about the causes of year-to-year climate variability is learned from ESM experiments [2].

2.2 Climate system drivers and human influence

Persistent changes in the mean state are induced by climate system drivers, which are mechanisms that modify the energy budget through alterations in the effective radiative forcing (ERF), thereby either mitigating or intensifying the greenhouse effect. These effects are referred to as *externally forced variability.* The ERF is determined by the change in the net downward flux of radiation at the top of the atmosphere, calculated after the ES has adjusted to perturbations caused by the climate system driver, excluding the radiative response due to changes in surface temperatures [1]. Climate system drivers are typically divided into a) natural climate drivers, such as orbital and solar activity changes, as well as volcanic stratospheric aerosols, and b) antropgenically mediated climate drivers such as GHGs (both well-mixed greenhouse gass (WMGHGs) and halogoneted GHGs, ozone (O_3), stratospheric water vapor, etc.), and changes in land use and tropospheric aerosols.

Figure 2.2 displays the historic development of the ERF from key climate system drivers. In combination with the Milankovitch cycles, solar activity governs the total solar irradiation (TSI). The solar activity follows an eleven-year solar sunspot cycle, whose dynamics can be seen as an oscillation in Figure 2.2 [2]. Strong volcanic eruptions, that is, eruptions whose plums transport sulfate aerosols into the stratosphere, are capable of significantly reducing the solar energy flux. The result is a warming of the stratosphere during the 1-2 year life of aerosols in the stratosphere, as well as a cooling of the surface for a period of 2-3 years [2]. This effect can be seen in Figure 2.2 as large short-term dips in the total ERF. Tropospheric aerosols originate from natural sources (both biotic and abiotic) and antropogenic sources that are generated primarily through the burning of fossil fuels and biomass. Since the industrial revolution, aerosols have reduced ERF, as shown in Figure 2.2. As all antropgenically mediated climate drivers, aerosols span a wide ERF range between 1750 and 2020. Figure 2.2 shows an increased cooling effect that developed in the 20th century in response to the increase in concentrations. Observations after 2000 exhibit predominantly negative trends in aerosol abundance, which implies a reduction in negative ERF as seen in Figure 2.2. The effects of tropospheric aerosols as a climate system driver are not fully understood; this is reflected in the wide uncertainty range.

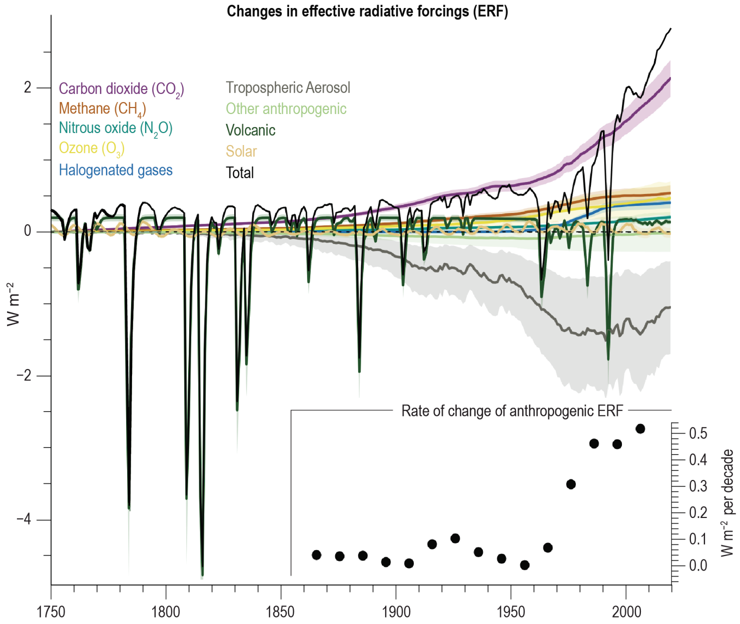


Figure 2.2: Historic development of ERF divided into key climate system drivers. Solid lines represent the global annual mean, and shaded ranges represent the 5%-95% uncertainty range. Sudden, large dips are following medium and large volcanic eruptions. The inserted plot shows the rate of change (linear trend) in antropogenic ERF for 30-year periods centered on each dot. Illustration from IPCC AR6 [1].

Well-mixed greenhouse gasses (WMGHGs) include chemical compounds that a) contribute to the greenhouse effect by absorbing infrared radiation and b) have a life in the atmosphere long enough for them to mix approximately uniformly. Key compounds include carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O). In 2019, the concentration of CO_2 reached 410 ppm (measurements by the National Oceanic and Atmospheric Administration (NOAA)), a level not seen for more than 2 Ma. Furthermore, the rate of increase in WMGHGs concentrations has been unmatched in at least the past 800,000 years [1]. As seen in Figure 2.2, the cumulative effect of GHGs on the ERF is by far the single category of climate system drivers that contribute to the positive trend.

2.3 Indicators of climate change

Sections 2.1 and 2.2 can be summarized by: Earth's climate system is perturbed from its mean state by internal variability, influenced by interactions between the atmosphere, ocean, cryosphere, biosphere, and lithosphere. This complex, dynamic system generates year-to-year variability. Trends in the mean state result from shifts in climate system drivers. Natural drivers like solar and orbital variations are long-term and somewhat predictable, while events like volcanic activity cause shorter-term disturbances. Antropogenically mediated climate drivers are, for humanity as a whole, controllable through emissions of GHGs and tropospheric aerosols. In this section, we will present some key *indicators of climate change* that are commonly assessed in scenario projections. Although the numerous observed changes in global climate change indicators have significant impacts on their own, changes in weather and climate extremes pose a more substantial risk to human and ecosystem well-being. The AR6 assessment indicates an increase in the intensity and frequency of future climate extremes, underscoring the significant incentives to monitor these trends. Some commonly assessed extremes include temperature extremes, heavy precipitation, floods, droughts, and storms [1].

The global surface air temperature (GSAT) is perhaps the most well-studied indicator of climate change, with reliable measurements dating back to 1850 in multiple locations around the world. The GSAT is calculated by taking the spatial mean of near-surface air temperature (tas), thus GSAT quantifies the temperature of the atmosphere close to the surface, a property that is a key indicator of climate *impact* [3]. The GSAT trend of the 20th century was positive, with an estimated anomaly between 1850-1900 and 2011-2020 climatologies of 1.09 $[0.91 - 1.23]$ °C [1]. However, the changes have not been evenly distributed spatially or temporally. The GSAT has had a positive trend since the industrial revolution; however, warming accelerated after the 1970s, following a rate that has been unprecedented in at least 2000 years. The observed heating is not spatially uniform, with more rapid heating above land and in the arctic and polar regions. Temperature is a key property of the climate system, as it is known to govern the spatial distribution and magnitude of precipitation, circulation in both the atmosphere and ocean, and the mass distribution of ice and water. All of which have great implications for living conditions. In the context of extreme events, the evaluation of the intensity and frequency of both hot and cold extremes is typically used as a standard measure. GHG forcing is the primary factor driving the warming of temperature extremes, influenced by regional factors such as changes in land use, aerosols, and natural variability, leading to varied regional impacts and associated uncertainties. Since 1950, there has been a global increase in warm days and nights and a decrease in cold days, with similar trends in temperature extremes and heatwave characteristics observed regionally on continents such as Europe, Australasia, and North America. Furthermore, annual minimum land temperatures have increased significantly since the 1960s, especially in the Arctic, indicating a stronger warming trend compared to global surface temperature [1].

The global total column of water vapor (TCWV) has been linked to the initial increase in temperature, displaying a rising pattern since the first representative observations were made in 1979. The magnitude of this change is reported to be around 7% per °C, a statistic that is nevertheless disputed due to modifications in observational methodologies [1]. Although there has been a rise in TCWV, global trends in precipitation have not shown a convincing increase, with anomalies from 1980 to 2019 exhibiting a high proportion of insignificant observations. This is in part due to high year-to-year variability in observations but also to low coverage of observations over oceans. Precipitation remains a crucial measure to consider, as a significant portion of the anticipated vulnerability to climate change is related to alterations in precipitation trends [3]. In the context of extreme events, the evaluation of the intensity and frequency of heavy precipitation is the main variable of evaluation. Heavy precipitation is a more complex metric than temperature extremes, with multiple driving factors that include (in addition to thermodynamics) dynamical modes such as ENSO as well as changes in aerosol forcing and land-use change. Through AR6, IPCC reported a likely increase in both frequency and intensity of heavy precipitation since 1950, particularly in North America, Europe and Asia, where observational data are abundant enough for a valid assessment [1].

Changes in the cryosphere, such as ice sheet melting and declining sea ice coverage, are clear indicators of the long-term effects of global warming. Arctic SIA is a key indicator of climate change. From 1979 to 2019, Arctic SIA has consistently decreased throughout all months, with the decadal average Arctic sea ice area (Arctic SIA) for September declining from 6.23 million km^2 in the first decade to 3.76 million km^2 in the last decade of the period [1]. Other indicators, such as the Antarctic and Greenland ice sheets, have shown similar responses, with an increase in the melting rate. In addition to runoff from the terrestrial cryosphere, the ocean undergoes significant heating, resulting in thermal expansion. Since 1901, global mean sea level (GMSL) has increased by 0.2 [0.15, 0.25] m at a growth rate of 1.3 [0.6, 2.1]. From 1901 to 1971, mmyr^{-1} increased to 1.9 [0.8-2.9]. mmyr^{-1} increased from 1971 to 2006, reaching 3.7 [3.2-4.2]. mmyr^{-1} from 2006 to 2018 [1].

2.4 Earth system modeling

2.4.1 Historical development of earth system modeling

In this and the following sections, we will shift the focus from the physical systems to the Earth system models (ESMs), one of the key tools in order to understand the dynamics of the past, present, and future climate. We will start the overview by briefly presenting the historical development.

In 1922, Lewis Fry Richardson published a framework for numerical analysis of the atmosphere in his book *Weather prediction by numerical process* [11]. This framework, combined with electronic computers, enabled the evolution of the general circulation models (GCMs) in the 1960s by pioneers like William Sellers [12]. The early atmospheric circulation models were highly motivated by weather forecasting and not necessarily for climate investigations, unlike modern ESMs. The common factor of Richardson's framework, the early GCMs, and the atmospheric components of a modern ESM is that they all solve the laws of thermodynamics and Newton's laws of motion to gain an understanding of the state of the atmosphere [12, 13, 14].

Throughout the late 20th and early 21st centuries, GCMs transitioned from atmosphere-only models through atmosphere-ocean coupled models and into modern, interdisciplinary ESMs with a strong representation of ES components such as the cryosphere, biogeochemical cycles, and ecosystems [4]. This evolution has been allowed by a strong increase in computing power. This has also allowed a better representation by increasing the model resolution; the typical horizontal model resolution has increased from about 500 km in global models in FAR to 100 km in AR6 [15]. Subgrid processes, that is, processes for which the resolution is too coarse, are solved using parameterization schemes. These are semi-empiric statistical representations of processes such as diabatic heating, friction, clouds, precipitation, turbulence, etc., all of which are indeed important processes [14]. As a result, modern ESMs represents the ES with many main components considered, although they are still, to this day, only reductionist representations.

2.4.2 The Coupled Model Intercomparison Project (phase 6) and the Shared Socioeconomic Pathways

The Coupled Model Intercomparison Project (CMIP) was formed in 1995 in an effort to coordinate the design and distribution of global ESMs. The project is currently in its sixth phase, called Coupled Model Intercomparison Project phase 6 (CMIP6). Present in all stages of CMIP is the Diagnostic, Evaluation and Characterization of Klima (DECK) as well as historical simulations. This is a set of benchmark simulations that ensure that a model is appropriately tuned for inclusion in CMIP [16]. CMIP is divided into 21 CMIP endorsed Model Intercomparison Projects (MIPs), where participants conduct experiments. All CMIP participants ESMs can simulate a variety of experiments. Here, we focus on experiments related to Scenario Model Intercomparison Project (ScenarioMIP). From O'Neill et al. (2016), we summarize the ScenarioMIP objectives (of primary concern for this thesis) [17]:

i ScenarioMIP should facilitate integrated research leading to,

   a) better understanding of the physical climate system, and

   b) better understanding of climate's effect on societies.

ii The results should lead to new climate information that facilitates further research within both climate science and human systems.

iii The main social impact is a more comprehensive understanding of future community vulnerability that leads to informed policy making for adaptation and mitigation efforts

To assert the validity of the experiment, it is essential to introduce the scenario framework employed in CMIP6 and IPCCs' AR6. The scenario framework consists of two main components: 1) the representative concentration pathways (RCPs) representing main scenarios for radiative forcing (Wm^{-2}) by 2100, and 2) the shared socioeconomic pathways (SSPs) representing main narratives for global development if new climate policies were adapted [6, 18].

The RCPs was first introduced as the scenario framework for long-term experiments in the fifth phase of CMIP (CMIP5) [19]. For CMIP5, the RCPs provided a target radiative forcing level of 2100 relative to preindustrial conditions. The evolution of the forcing levels is then governed by the climate system drivers. The RCPs include the forcing levels of 2.6, 4.5, 6.0, and 8.5 Wm^{-2} as van Vuuren et al. (2011) identified the range of 2.5Wm^{-2} to 9Wm^{-2} plausible through literature studies [20]. Table 2.1 from van Vuuren et al. (2014) identifies the main characteristics of the RCPs.

Table 2.1: Key outcome properties of the RCPs in 2100. Table from van Vuuren et al. (2014) [18].

	Radiative forcing	**CO_2 equivalent concentration**	**Rate of change in radiative forcing**	**Key reference**
RCP 8.5	8.5 Wm^{-2}	1350 ppm	Rising	[21]
RCP 6.0	6.0 Wm^{-2}	850 ppm	Stabilizing	[22]
RCP 4.5	4.5 Wm^{-2}	650 ppm	Stabilizing	[23]
RCP 2.6	2.6 Wm^{-2}	450 ppm	Declining	[24]

In CMIP6 and AR6, the full scenario framework was adapted by the development of the SSPs and the linking of the SSPs to the RCPs [16]. The SSPs is based on five distinctly different narratives of socioeconomic development, all balancing the challenges of mitigation *of*, or adaptation *to*, climate change [6, 25]. Figure 2.3 illustrates the placement and relations of the SSPs in the adaptation-mitigation matrix. The SSPs ranges from a sustainable world (SSP1) to a fragmented world with strong competitiveness between nations (SSP3). To present the development of SSPs, we divide the process into three main parts: 1) the conceptual SSP narratives that provide baseline descriptions; 2) the assessment of quantitative factors that create projections of global demand for resources and resulting projections of antropogenic radiative forcing for the SSP if no mitigation is performed; and 3) the mitigation scenarios that assess the mitigation cost from the projected forcing generated in 2) to the different forcing levels of the RCPs.

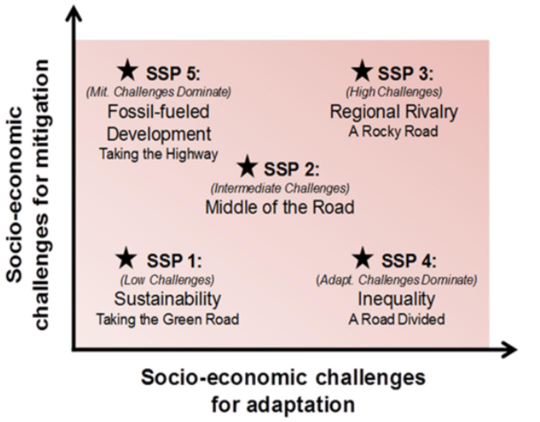


Figure 2.3: Overview of the SSPs with respect to the challenges of adaptation and mitigation to climate change. Illustration from O'Neill et al. (2017) [25].

1) The SSP baseline scenarios

The design of the narratives provides the underlying logic for each SSP, focusing on the socioeconomic change that requires a qualitative description as opposed to formal quantitative models [6]. We can summarize the narratives using formulations from on O'Neill et al. (2017) and Riahi et al. (2017) [6, 25]:

- *SSP1 Sustainability "Taking the Green Road"* - In SSP1, there are low challenges concerning both mitigation and adaptation due to:
  - Global transition towards sustainability and inclusivity with improvement in the management of global commons.
  - Shift in economic focus towards human well-being through large investment in education and health leading to a reduced inequality.
- *SSP2 Middle of the Road* - In SSP2, there are intermediate challenges concerning both mitigation and adaptation due to:
  - Uneven progress in development and income growth, as well as slow progress in achieving sustainable development goals by both national and global institutions.
  - Moderate global population growth is leveling off in the second half of the century, while income inequality persists or improves slowly.
- *SSP3 Regional Rivalry "A Rocky Road"* - In SSP3, there are high challenges concerning both mitigation and adaptation due to;

  - Strong nationalism, competitiveness, and security concerns drive the focus towards domestic or regional issues with a strengthening of policies toward national and regional security.
  - Emphasis on achieving energy and food security within regions at the expense of broader development.
  - Decline in investments in education and technological development, slow economic growth, material-intensive consumption, and persistent or worsening inequalities.

- *SSP4 Inequality "A Road Divided"* - In SSP4, there are great challenges concerning adaptation and low challenges regarding mitigation due to:
  - Highly unequal investments in human capital create increased inequalities between and within countries.
  - High technology development in knowledge- and capital-intensive sectors, diversification in the energy sector, and environmental policies that focus on local issues in middle- and high-income areas.

- *SSP5 Fossil-fueled Development "Taking the Highway"* - In SSP5, there are low challenges concerning adaptation and high challenges concerning mitigation due to:
  - Increasing faith in competitive markets and innovation for rapid technological progress and human capital.
  - The exploitation of fossil fuel resources and the adoption of resource-intensive lifestyles create rapid global economic growth.
  - Peak and decline of the global population in the 21st century, successful management of local environmental problems like air pollution, and faith in effective management of social and ecological systems, including geo-engineering if necessary.

2) Assessment of quantitative factors

The qualitative descriptions from the SSP baseline scenarios are then translated into quantitative projections of socioeconomic drivers (hereafter projections toward 2100) of population, education, urbanization, and economic development [6]. These projections provide quantitative information on socioeconomic development for use in integrated assessment models (IAMs). An IAM is a model that integrates scientific and socio-economic aspects of climate change to evaluate the impact of policies [26]. Here, policies are the SSP baseline scenarios, and the impact is the projected energy demand and land-use change. The impact of the respective SSP generates a wide range of GHG and pollutant emissions, with a resulting range for radiative forcing seen in Figure 2.4 [6]. The antropogenically mediated radiative forcing levels of Figure 2.4 are projected under the assumption of no mitigation. The SSP baseline scenarios project a range of radiative forcing between 5.0 – 8.7 Wm^{-2} by the end of the century. By comparing the end-of-century radiative forcing levels of Figure 2.4 with the SSP placements in the adaptation-mitigation matrix of Figure 2.3, we observe a high correlation between the mitigation challenges and the forcing levels for the individual SSP [6]. The lower range of radiative forcing is covered by the sustainable world of SSP1; however, the lower range is estimated at $\sim$ 5 Wm^{-2}, which demonstrates the need for mitigation efforts to reach the forcing levels of RCP4.5 and 2.6. The upper range is covered by the fossil-fueled development world; this is also the only SSP baseline scenario that reaches the forcing levels of RCP8.5.

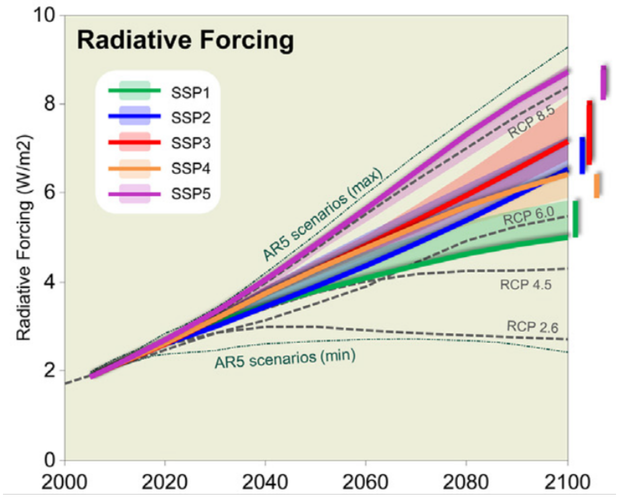


Figure 2.4: Time-series of the projected antropogenically mediated radiative forcing for ScenarioMIP baseline scenarios. The projections are simulated under the assumption of no mitigation of climate change. Illustration from Riahi et al. (2017) [6].

3) The mitigation scenarios

Mitigation scenarios are included to create plausible scenarios, as it is highly unlikely that no action will be taken throughout the 21st century. Figure 2.4 suggests that reducing climate impact is necessary to achieve lower levels of radiative forcing by 2100. The effectiveness of mitigation policies, as well as the relationship between the baseline radiative forcing levels of the target and the SSP [6], have a significant impact on the cost and attainability of the climate targets, including those of the Paris Agreement. For example, the development in the baseline scenarios of

SSP1 and SSP5 differs significantly in terms of carbon and energy intensity, reducing the required effort. For more information, see Riahi et al. (2017) [6].

The full scenarios and their predictions on indicators of climate change

Together, the forcing levels from the RCPs and the climate change scenarios of the SSPs create the RCP-SSP scenario matrix. In 2016, O'Neill et al. presented the scenarios and their rationale for ScenarioMIP. Figure A.1 found in Appendix A.2.1 shows the relation between the scenarios in the RCP-SSP matrix, with tier 1 (main scenarios) being SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5 (abbreviations use the form SSPx-y where y is the forcing of the respective RCP) [17]. Figures 2.4 and A.1 show that SSP3 and 5 remain close to the baseline forcing level, while SSP1 and 2 at the lower end require more robust mitigation to adapt.

Figure 1.1 shows the time-series of the GSAT generated from the near-surface air temperature variable of ESMs. The GSAT is reported relative to the climatology of 1995-2014, which is 0.84°C higher compared to the climatology of 1850-1900, a period often used as a proxy for preindustrial conditions. As shown, the multi-model mean of all scenarios projects a warming between 0.69 and 3.99°C, thus violating the 1.5°C goal of the Paris agreement [9]. The predicted GSAT anomalies in SSP1-2.6 compared to the reference period 1850-1900, represented by the multi-model mean (and 5-95 % multi-model ranges in square brackets) are 1.6 $[1.1 - 2.2]$ °C in the time period from 2021 to 2040 (near-term period) and 2.0 $[1.3 - 2.8]$ °C in the time period from 2081 to 2100 (long-term period). Thus, the mean prediction complies with the upper 2 °C goal of the Paris agreement. The other scenarios yields: SSP2-4.5 has a near-term period anomaly of 1.6 $[1.0, 2.3]$ °C and long-term period anomaly of 2.9 $[2.1, 4.0]$ °C. SSP3-6. has a near-term period anomaly of 1.6 $[1.0, 2.4]$ °C and long-term period anomaly of 3.9 $[2.8, 5.5]$ °C. Lastly, the high radiative forcing scenario SSP5-8.5 projects an anomaly of 1.7 $[1.2-2.4]$ °C in the near-term period and 4.8 $[3.6-6.5]$ °C in the long-term period [1]. The closeness of the near-term mean values and the dispersion of the long-term mean values substantiate the disparate rates of warming under the policy scenarios of the SSP framework. The progression is succinctly summarized in Table 2.2 delineates the best-estimation year at which a given scenario attains specific warming thresholds relative to the 1850-1900 preindustrial proxy. The uncertainty interval of the estimate, predicated on the 5% - 95% range of the multi-model ensemble following an 11-year running mean smoothing procedure, is displayed within square brackets [9]. The number of models that reach a level x and the total count of accessible models X are represented in parentheses as: (x/X).

Table 2.2: Times of reaching certain warming levels compared to 1850–1900 in the ScenarioMIP scenarios. SSP1-1.9 is a Tier 2 scenario specifically designed to reach the 1.5 °C goal of the Paris agreement (see Appendix A.2.1 for the position of SSP1-1.9 in the RCP-SSP matrix) [17]. The best estimate is shown as the main year, while uncertainty ranges from 5% to 95% of the multi-model ensamble post-smoothing by 11-year running means are shown in square brackets. The lower range of uncertainty is guided by the ensamble members with the highest warming projections, since this crosses the respective thresholds the fastest. The opposite is true for the upper range. The analysis is performed using a common subset of 31 models for Tier 1 scenarios, while all 13 available models are used for SSP1–1.9. The number of models reaching a level x and the number of available models X are shown in parentheses as: (x/X). Table from Tebaldi et al. (2021) [9].

	SSP1-1.9	**SSP1-2.6**	**SSP2-4.5**	**SSP3-7.0**	**SSP5-8.5**
1.5°C	2029 [2021, NA] (11/13)	2028 [2020, NA] (30/31)	2028 [2020, 2047] (31/31)	2028 [2020, 2045] (31/31)	2026 [2020, 2040] (31/31)
2.0°C	NA [2036, NA] (2/13)	2064 [2032, NA] (17/31)	2046 [2032, 2082] (31/31)	2043 [2031, 2064] (31/31)	2039 [2030, 2055] (31/31)
3.0°C	NA [NA, NA] (0/13)	NA [NA, NA] (0/31)	2094 [2058, NA] (16/31)	2069 [2052, NA] (31/31)	2060 [2048, 2083] (31/31)
4.0°C	NA [NA, NA] (0/13)	NA [NA, NA] (0/31)	NA [NA, NA] (1/31)	2091 [2071, NA] (17/31)	2078 [2062, NA] (27/31)
5.0°C	NA [NA, NA] (0/13)	NA [NA, NA] (0/31)	NA [NA, NA] (0/31)	NA [2088, NA] (3/31)	2094 [2075, NA] (15/31)

The anticipated increase in temperature over the course of the 21st century does not exhibit a uniform spatial distribution, with critical dynamics encompassing disparities between land and ocean, as well as between the tropics, arctic, and antarctic. The intensity of extreme weather events correlates with the degree of global warming. Consequently, the severity and occurrence of extreme heat events are inevitable and will escalate linearly with the level of global warming in the scenario. Furthermore, it is highly probable that the severity of these extremes will increase more dramatically than the overall global warming on land, which is crucial in evaluating temperature as a driver of climate impact [1].

The predicted changes in precipitation (pr), shown in the right panel of Figure 1.1, show less clear patterns with non-significant statistical changes due to model uncertainty and internal variability. As a result, global change figures are not reported in the near-term period. By the end of the century, the predicted changes are 2.9 $[1.0, 5.2]\%$, 4.0 $[2.3, 6.7]\%$, 4.7 $[2.3, 8.2]\%$, and 6.5 $[3.4 - 10.9]\%$ for SSP1-2.6, SSP2-4.5, SSP3-7.0, and SSP5-8.5, respectively [1]. Thus, it shows an overlap in 5-95 % model predictions and a less clear signal than what is observed for temperature. At a global warming threshold of 4°C above preindustrial levels, heavy precipitation events that were previously rare will become more frequent and intense globally and across all continents. This increase in frequency and intensity of heavy precipitation is extremely likely for most continents and very likely for most AR6 regions. However, the likelihood of extreme precipitation increases substantially at higher levels of global warming on all continents. Thus, as seen in Table 2.2, the need and scale of adaptations to extreme precipitation will occur at significantly different times throughout the 21st century, depending on the scenario, with only SSP3-7.0 and SSP5-8.5 crossing the threshold 4 °C.

2.4.3 Uncertainty in earth system modeling

Uncertainty pervades ES modeling, complicating efforts to accurately simulate and predict the behavior of the interconnected environmental processes on Earth. This uncertainty arises from various sources, including the complexity of the ES, the limitations of available data and computational methods, and the inherent unpredictability of natural phenomena, challenging our ability to reliably forecast future climate scenarios and inform effective policy responses [27]. Here we will account for the commonly discussed sources of model-based uncertainty.

Differences in structural properties and parameter properties (which refer to the mathematical framework constructed from theoretical and empirical understanding of the ES and the parameterization methods of subgrid and boundary (e.g., top of atmosphere) processes, respectively) represent the deciding factors for an ESM's climate sensitivity. Climate sensitivity is a measure of the magnitude of global warming (measured by the GSAT) that is expected as a response to a doubling of CO_2 compared to preindustrial levels [2]. The transient climate sensitivity is the GSAT changed at the instant of the doubling of atmospheric CO_2. In ESMs, the transient climate sensitivity (TCR) is estimated from the CMIP DECK experiments *1pctCO2* [28]. The benchmark TCR of AR6 is calculated (by a combined assessment of process understanding, instrumental records, paleoclimates, and emergent constraints) to have a central value of 1.8 °C and a likely range of $1.3 - 2.2$ °C [1]. The equilibrium climate sensitivity is the GSAT long-term response after the ES has reached an equilibrium stage (leaving out feedback's associated with ice sheets) [1]. In ESMs, the equilibrium climate sensitivity (ECS) is estimated from the CMIP DECK experiment *abrupt-4xCO2*. The benchmark ECS for AR6 is calculated with a central value of 3 °C and a likely range of $2.0 - 5.0$ °C [1]. Because of the latency of ocean heat uptake, the TCR will always be smaller than the ECS, implying that the full realization of the climate system takes time. Since the structural and parameter properties are generally different between ESMs, the climate sensitivity will also be different, allowing for coverage of models in a wider range of the benchmark range. Furthermore, the heterogeneity of the ESMs climate sensitivity will alter the responses to indicators of climate change. In general, the effect on GSAT is positively correlated with the climate sensitivity in CMIP ESMs; however, the effects are not uniformly distorted spatially [29]. A higher global temperature is, in turn, associated with a higher global precipitation through the Clausius-Clapeyron relation [1]. The effect is seen by comparing the panels in Figure 1.1.

In projections of climate change, the external forcing is prescribed by the person running the model. In that, assumptions are made about projections of socioeconomic drivers. The validity of these assumptions declines as the timescale of the projection is moved further into the future. This is known as the scenario uncertainty of predictive modeling. In order to reduce the uncertainty of the scenario, the scenario framework (presented in Section 2.4.2) aims to provide a more accurate prediction by considering various socioeconomic factors [17]. However, as shown in Figure 2.4, projections are still subject to uncertainty due to the inherent unpredictability of human systems.

The internal variability of the climate system (see Section 2.1) is displayed through the non-deterministic nature of ESMs. That is, two model realizations will not develop alike, even though they have the same starting conditions. As a result, the internal variability of the ES can complicate the interpretation of ESM results, as it can be challenging to distinguish between internal variability and forced variability in short-term predictions [27]. Multi-model ensembles are used in CMIP experiments in order to assess the robustness and uncertainty of the predictions. Uncertainties arising from the variable climate sensitivity of ESMs are also addressed using a multi-model mean, as this provides a wider coverage within the likely range of the benchmark ECS and TCR.

2.5 Previous work on SSP separation

In the process of situating this thesis within a broader scientific context, a literature review was conducted. Throughout the review, we found no studies employing ML techniques to predict the socioeconomic scenario of ESM data. Although this identifies a gap in the existing science toolbox, highlighting the potential significance of this method as a strong tool for informing policies. On the other hand, global-level climate change indicators have been extensively explored through initiatives such as ScenarioMIP and CMIP6. These studies provide a robust framework for understanding the divergence of scenarios based on individual indicators of climate change.

In its AR6, the IPCC provides an in-depth evaluation of changes in the global climate throughout the time period from 2021 to 2040. Key indicators such as the GSAT, Arctic sea ice cover, and precipitation were assessed. The findings suggest minimal scenario-dependent variations in GSAT when averaged over the near-term period relative to 1995–2014, projecting an increase of 0.4 °C to 1.0 °C. This projection is primarily influenced by uncertainties in the ECS and the TCR [1]. Regarding Arctic sea ice, CMIP6 model outputs indicate a likely decrease in the Arctic sea ice area in September in the near term, with a significant majority of model simulations supporting this trend. There are significant regional differences in precipitation, with some regions experiencing significant flux changes. However, when averaged globally, these changes largely cancel out. Hence, precipitation does not show robust significant changes globally in the near-term period and is therefore not evaluated in depth with regard to SSP separation. Consequently, GSAT emerges as the most robust indicator of global climate change. The separation times of GSAT trajectories under different SSPs are detailed in Table 2.3. Tebaldi et al. (2021) define a SSP separation as the year by which the multi-model ensamble means separate with a positive difference of 0.1 ° that persists for the remainder of the century.

Table 2.3: Time of separation between smoothed GSAT trajectories under pairs of scenarios. Shown are the years by which the ensemble means separate and, in square brackets, the years by which the last of the separation among individual models' trajectories takes place. Separation is defined as the emergence of a positive difference (we use 0.1 °C as the threshold) that persists for the remainder of the century. We first apply a 21-year running mean to the GSAT time series in order to characterize separation "of climates". Table and caption from Tebaldi et al. (2021) [9].

	SSP1-2.6	**SSP2-4.5**	**SSP3-7.0**	**SSP5-8.5**
SSP1-1.9	2042 [2050]	2034 [2043]	2031 [2041]	2027 [2036]
SSP1-2.6		2039 [2053]	2037 [2048]	2030 [2036]
SSP2-4.5			2046 [2058]	2031 [2044]
SSP3-7.0				2034 [2053]

In this thesis, the available data are limited to realizations spanning from 2015 to 2100, i.e. the main period evaluated in ScenarioMIP. Consequently, the climatologist approach as performed by Tebaldi et al. (2021) is not available for the full near-term period, limiting the investigation to the annual mean GSAT [9]. We will use the p-value of the ANOVA test statistics to relate the GSAT class separation to the findings of Tebaldi et al. (2021). This will in turn provide a rough benchmark for the feature-sets in order to assess whether or not including more indicators of climate change in a predictive model will increase the SSP separation or increase the noise ratio. However, it will limit the comparability and also the interpretability of the assessment metrics as being *climatological* responses as the uncertainty in near-term projections of the annual mean GSAT is shown to be equally large as the uncertainty arising from internal variability and structural properties of theESM [1].

Chapter 3

Theory and Method

3.1 Data Origin and Pre-Processing

3.1.1 Data Origin [1]

In this thesis, we present an analysis of the output data of Australian Community Climate and Earth System Simulator earth system model 1.5 (ACCESS-ESM1.5), which was one of the CMIP6 participant models, ran on experiments from ScenarioMIP. The key configurations of ACCESS-ESM1.5 are summarized in Table 3.1 and are provided in detail in Ziehn et al. (2020) [31]. The choice of using only one model as opposed to a multi-model approach is made under consideration that a) we aim for a proof of concept, and b) in multi-model ensembles we have different climate sensitivities, introducing more noise to the data-set. This is generally a favorable behavior when using multi-model ensembles, as discussed in Section 2.4.3; however, for our intended usage, it will drastically complicate the analysis, as it introduces an additional dimension of uncertainty in the results.

Table 3.1: Key configurations of ACCESS-ESM1.5. For further details see Ziehn et al. (2020) [31].

Component	**Comments**
Atmosphere	• **Resolution**: lon: 1.875°, lat: 1.25°, height: 38 levels (40 km model top). • Uses the UK Met Office Unified Model (UM)
Land	• Uses the Community Atmosphere Biosphere Land Exchange (CABLE) model with multiple tiles per grid cell. • CABLE is configured with 10 vegetated types.
Ocean	• **Resolution**: Horizontal grid 360 × 300 that generates a nominal resolution 1°. 50 vertical levels at nominally 10-200 m thickness. • uses the National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL) Modular Ocean Model (MOM) version 5.

We chose ACCESS-ESM1.5 for two main reasons: 1) ACCESS-ESM1.5 has a lot of accessible data: This makes ML easier to perform and the results more stable. 2) Appropriate climate sensitivity and projection abilities: The ECS is reported at 3.87 °C and the TCR is reported at 1.95 °C, which are near average among CMIP6 reported ESMs and within the *likely* range of combined assessment of AR6 [1, 16, 31].

In Sections 2.3 and 2.4, we presented the historical evolution of key indicators of climate change and the multi-model predictions presented by the IPCC in AR6 for ScenarioMIP scenarios, respectively. In most sciences, the system consists of *describing* variables and *response* variables. For the climate system, we have referred to the descriptive variables as "climate system drivers" and will continue this usage. The response variables include meteorological variables introduced as "indicators of climate change". For the remainder of the thesis, we refer to these indicators as *"response variables"* or when deployed in the ML as features. However, note that a *feature* can be a subset of the global-level indicator by masking (see Section 3.1.3).

In this thesis, we only use response variables for the model training, even though both descriptive and response variables are measurable in situ. This decision is based on the fact that the effects of climate change are primarily experienced *through* the response variables rather than directly through climate system drivers (with certain exceptions) [3]. Within the realm of response variables, we focus on atmospheric variables, since many key response variables of the ocean, cryosphere, and circulation either have a) week separation within the near-term period, or b) have a lower coverage of available in-situ measurements. The latter is important in the case of model application to real-world data. For the analysis, near-surface air temperature (tas) and precipitation (pr) are the main variables. The four remanding variables maximum value of daily maximum temperature (txx), maximum 5-day precipitation (rx5day), growing season length (gsl), and frost days (fd) are response variables of extreme events calculated from the main variables using the Expert Team on Climate Change Detection and Indices (ETCCDI) definitions [32]. Table 3.2 provides a summary and explanations of the variables employed in this thesis.

[1] Based on term paper submitted in the subject *FYS-STK4155 - Applied Data Analysis and Machine Learning* autumn 2023 [30]

Table 3.2: Overview and description of response variables. Descriptions from the respective source.

Acronym	Unit	Description	Sources
tas	K	Near-surface (2 meter elev.) air temperature.	[33]
txx	°C	Maximum of daily maximum temperature in year.	[32]
pr	kg $m^{-2}s^{-1}$	Precipitation flux including both liquid and solid phases.	[33]
rx5day	mm	Maximum precipitation in five consecutive days in year.	[32]
gsl	days	Annual (January 1st to December 31st in the Northern hemisphere and July 1st to June 30th in the Southern hemisphere) count of days between first period of at least 6 days with daily mean temperature above 5 °C and first span after July 1st (NH) or January 1st (SH) of at least 6 days with daily mean temperature $<$ 5 °C.	[32]
fd	days	Annual count of days when daily minimum temperature $<$ 0 °C.	[32]

Since these extreme variables are simple to compute using the original indicator, they provide a low-cost source of *new* dynamics that do not emerge in the original data. This could potentially reduce the demand for high-accuracy in-situ measurements of additional indicators.

The model output data are produced as part of the CMIP6 exercise and subsequent reprocessing by the CICERO Center for International Climate Research. Figure 3.1 displays the mean value of the response variables across 40 realizations and near-term period for the grid data. The main variables are provided as daily means generated from the original daily temporal resolution. The extreme variables are delivered as annual data, following their definitions. In order to more effectively analyze the evolution of the scenarios, we perform smoothing through temporal and spatial averaging. These calculations are presented in Section 3.1.2. As seen in Figure 3.1, all variables exhibit strong spatial variability. This will make the separation of scenarios harder due to the stochastic nature of the ESM, but it also represents an opportunity to detect areas where the ACC emerges earlier than in the global mean. We account for the masking process needed to take advantage of the spatial variability in Section 3.1.3.

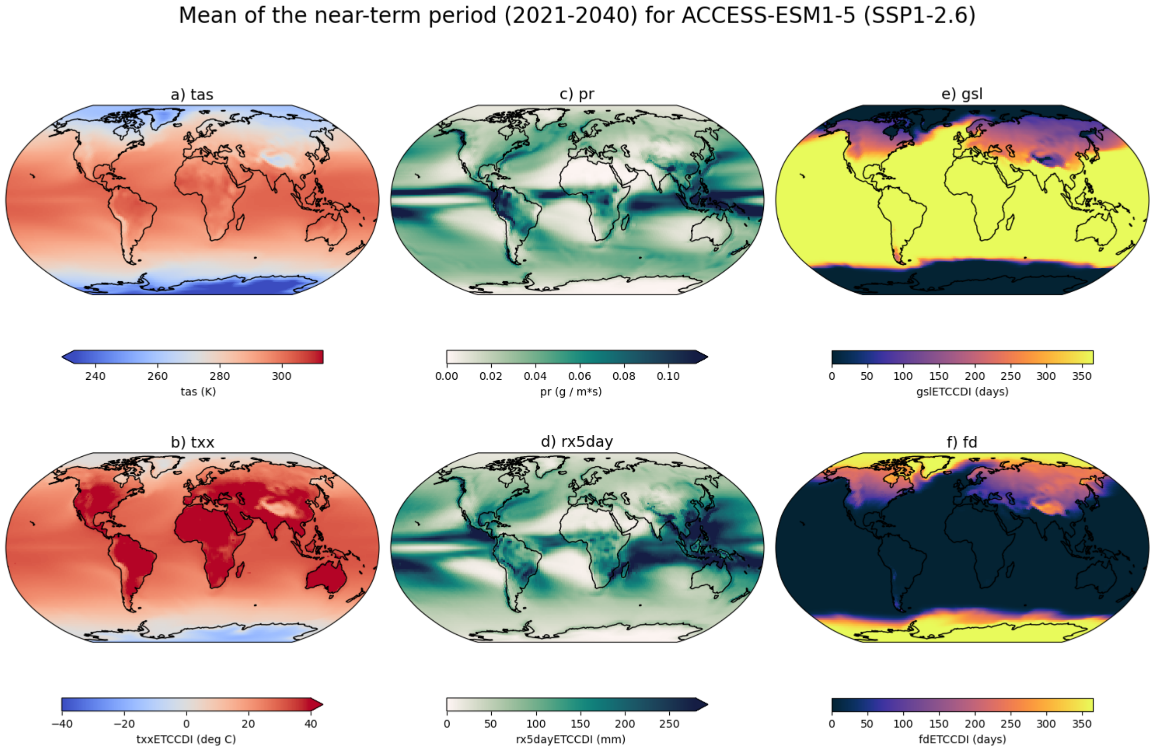


Figure 3.1: Grid data from ACCESS-ESM1.5 SSP1-2.6. The data shows the mean of 40 realizations across the near-term period. The subfigures hold data for: a) tas, b) txx, c) pr, d) rx5day, e) gsl, and f) fd, descriptions of abbreviations given in Table 3.2.

3.1.2 Global mean temporal evolution

The time-series of Figure 3.2 is the global mean temporal evolution generated from the grid data of Figure 3.1. This is done by downsampling from the original resolution presented in tab 3.1 to: a) a yearly temporal resolution; and b) a global spatial resolution.

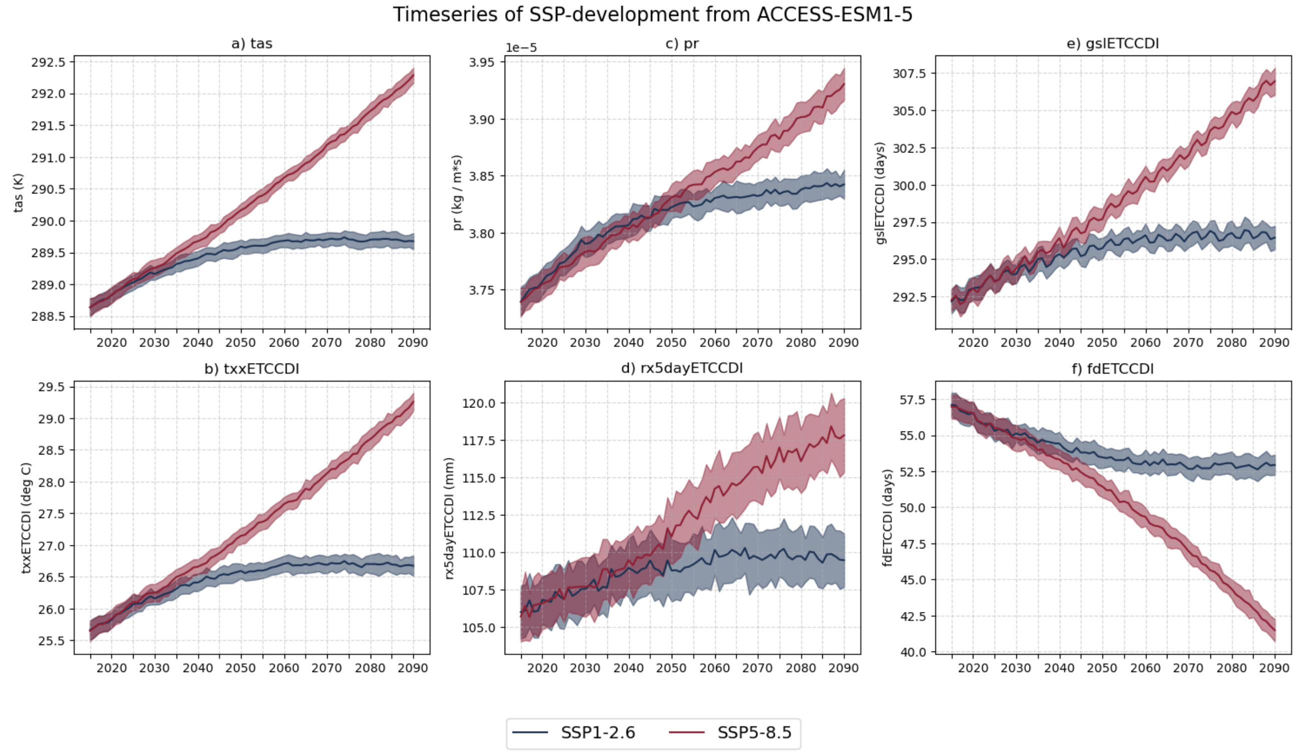


Figure 3.2: Time-series from ACCESS-ESM1.5. The data shown is ensamble means with shaded area being the standard deviation generated from the global (spatial) mean and yearly (temporal) mean of 40 realizations per scenario. The common factor for all variables is that univariate distributions do not show a clear separation within the near-term period. Note that "clear separation" here refers to no overlap between the standard deviations of each scenario. This dynamic is more evident if intermediate warming scenarios, such as SSP2-4.5 and SSP3-6.0, are included, as shown in Figure 1.1. The subfigures hold time-series of; a) tas, b) txx, c) pr, d) rx5day, e) gsl and f) fd, descriptions of abbreviations given in Table 3.2.

Thus, the temporal resolution is downsampled for the main variables tas and pr. For one realization x, this is done by the arithmetic mean of equation (3.1) for all observations within the year. In this context, the index i denotes the latitudinal component of the grid θ, while the index j signifies the longitudinal component ϕ. As a result, the temporal signal is smoothed, effectively removing the annual seasonality and revealing only the variability of longer time scales. Since ScenarioMIP experiments are simulating the future, the data include leap years; thus the length of the year holds the values $N \in [365, 366]$ in accordance with the four-year leap year cycle.

$$x_{i,j}^{(\text{year})} = \frac{1}{N} \sum_{\text{day}=1}^{N} x_{i,j}^{(\text{day})} \quad , \text{for year} \in [2015, 2100] \tag{3.1}$$

In addition, the spatial resolution is downsampled from the original latitude-longitude grid to one global observation. The downsampling is performed by the latitude-weighted mean of equation (3.2) across all grid boxes $x_{i,j}$.

$$x_{\text{global}}^{(\text{year})} = \frac{\sum_i \sum_j x_{i,j}^{(\text{year})} cos(\theta_{i,j})}{\sum_i \sum_j cos(\theta_{i,j})} \quad , \text{for year} \in [2015, 2100] \tag{3.2}$$

If masks are used, the global mean temporal evolution is computed across all non-NaN grid boxes, that is, Nan grid boxes do not affect the mean. We introduce the weight for the grid box (i, j); $cos\theta_{i,j}$ to account for the variable area of the grid boxes at different latitudes θ. $x_{\text{global}}^{(\text{year})}$ represents a random realization. For future notation, we introduce the notation $x_{s,r}^{(\text{year})} := x_{\text{global}}^{(\text{year})}$ for year$\in [2015, 2100]$, SSPs $s \in [\text{SSP1-2.6}, \text{SSP5-8.5}]$ and realization $r \in [1, 2, \ldots, 40]$.

3.1.3 Masking: Regional differences

We use masking to utilize the regional differences of response variables, a process in which the unmasked areas are set to NaN to remove their impact on the global means. Masking allows for the detection of early emerging signals, as different regions will experience different developments. Differences occur due to variability in dynamical relations, but are also the result of different responses to antropogenically mediated climate drivers. The resulting time-series will have a new and distinctively different development as shown in Figure 3.3, both between masks but more importantly between the mask and the global mean (calculations using *nomask*).

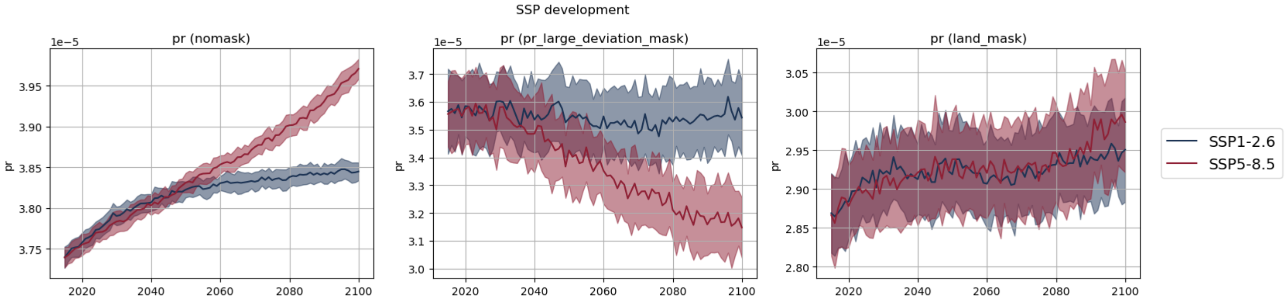


Figure 3.3: Example time-series of SSP development for precipitation using *nomask*, *pr_large_deviation_mask* and *land_mask*. All time-series contain the pr response variable, however the means are generated using subsets of the full grid. Mask descriptions are provided in Table 3.3 and visualized in figure 3.4.

The mask detection process is well documented in the analysis code, which is available in Appendix A.6. In total, we have evaluated 17 masks in the six response variables. An overview of the configurations is available in Table A.1 found in Appendix A.2.2. To evaluate the masks, we use visual analysis of the time-series. The main criteria include a) visual analysis of the signal to noise ratio (SNR) and SSP evolution in the near-term period, and b) if masks generate a similar evolution of SSPs, the mask containing a larger area is kept. After mask evaluation, seven masks are kept in addition to the baseline global mean *nomask*. These are displayed in Figure 3.4, where the red-shaded areas represent the included subset for the grid boxes. Table 3.3 holds the response variable-mask combinations that are kept for the calculations of the global annual means and used in ML analysis, as well as the quantitative mask descriptions.

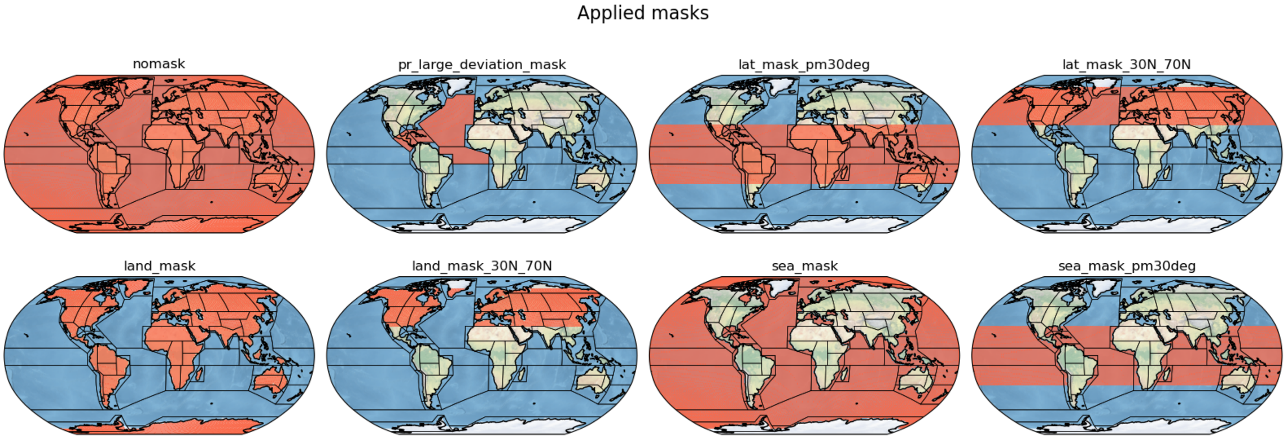


Figure 3.4: Masks used for data-set aggregation in this study. Areas shaded in red indicate the included subsets. Table 3.3 contains the combinations of variables and masks used in subsequent analyses, along with descriptions of the masks.

Table 3.3: The response variable-mask combinations that are kept fordata-set aggregation and ML analysis. The quantitative description is also provided bellow each mask while Figure 3.4 visualizes the masks.

Mask short name, **mask description: All cells $x_{i,j}$ ...**	**tas**	**pr**	**txx**	**rx5day**	**gsl**	**fd**
nomask -	X	X	X	X	X	X
land_mask, ... where landfrac$_{\mathrm{i,j}} \geq 0.8$	X	X	X	X		
sea_mask, ... where landfrac$_{\mathrm{i,j}} \leq 0.2$	X	X	X	X		
lat_mask_30N_70N, ... where 30N $\leq \phi_{i,j} \leq$ 70N					X	X
land_mask_30N_70N, ... where landfrac$_{\mathrm{i,j}} \geq 0.8$ and 30N $\leq \phi_{i,j} \leq$ 70N					X	X
lat_mask_pm30deg, ... where 30S $\leq \phi_{i,j} \leq$ 30N			X	X		
sea_mask_pm30deg, ... where landfrac$_{\mathrm{i,j}} \leq 0.2$ and 30S $\leq \phi_{i,j} \leq$ 30N		X	X			
pr_large_deviation_mask, ... in AR6 areas 'SCA', 'CAR', 'NAO', 'EAO'		X		X		

3.1.4 Cross-sections

Following the calculation of the global annual means, we can derive a cross-section for each year. In this thesis, we analyze six response variables; therefore, the typical cross section $\mathcal{S}$ is defined as a subspace $\in \mathbb{R}^6$. However, since masking can alter the properties of a response variable, rooms of higher dimensions can also be constructed. Each year's cross-section contains 40 realizations per SSP, that is, $x_{s,r}^{(\mathrm{year})} \in \mathcal{S}$. Figure 3.5 illustrates two examples $\mathcal{S}$ using a space $\in \mathbb{R}^2$ spanned by tas and pr.

For some ML algorithms, the scale differences of the dimensions are important. For others, scaling does not affect performance. Therefore, we use the standard scaled realization

$$x'^{(year)}_{s,r} = \frac{x^{(year)}_{s,r} - \overline{x}^{(year)}_{:,:}}{\sigma^{(year)}_{:,:}} \quad , \text{ for } \quad \begin{array}{l} \text{year} \in [2015, 2100] \\ s \in [\text{SSP1-2.6}, \text{SSP5-8.5}] \\ r \in [1, 2, \ldots, 40] \end{array} \quad , \tag{3.3}$$

to all ML and statistics. In equation (3.3) $\overline{x}^{(year)}_{:,:}$ and $\sigma^{(year)}_{:,:}$ is the mean and standard deviation across all SSPs and realizations respectively.

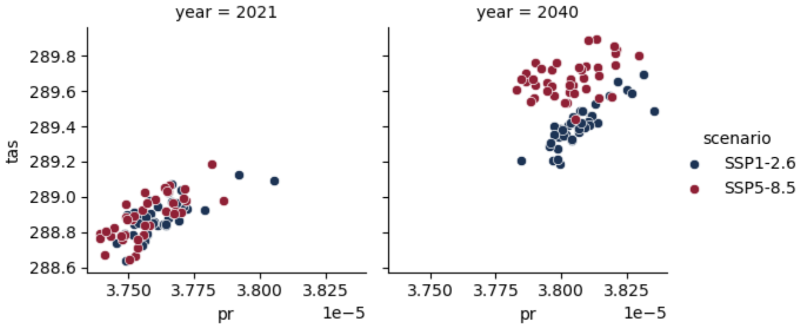


Figure 3.5: Example cross-sections in 2021 and 2040. These cross-sections illustrates the realizations position in two-dimensional room spanned by tas and pr. The two cross-sections also illustrates how the temporal evolution alters the cluster-overlap between scenarios.

3.2 Machine learning

Machine learning (ML) is a field of artificial intelligence that enables computer learning without explicit rule-based programming by the practitioner [34]. Consequently, the algorithm aims to *learn* a mapping from an input *training-set* to ultimately develop a model capable of making predictions on new, unseen data. In this thesis, we use realizations of the ACCESS-ESM1.5 ESM, both for the creation of the training-set and the *test-set*. The test-set is separated from the full set of realizations and serves as a proxy for real-world data, aiding in the estimation of the generalization ability of the model. The ESM data-set consists of both features, that is, the input used by the model to make predictions and the target variable. Thus, we are in a *supervised* learning setting where the target scenario is known. Figure 3.6 shows a flowchart of the supervised learning pipeline for this thesis. Some steps are already accounted for, such as the data collection and creation described in Section 2.4 and data pre-processing described in Section 3.1. The remaining steps are described throughout this Section 3.2 as well as executed in the results and discussion of Chapter 4.

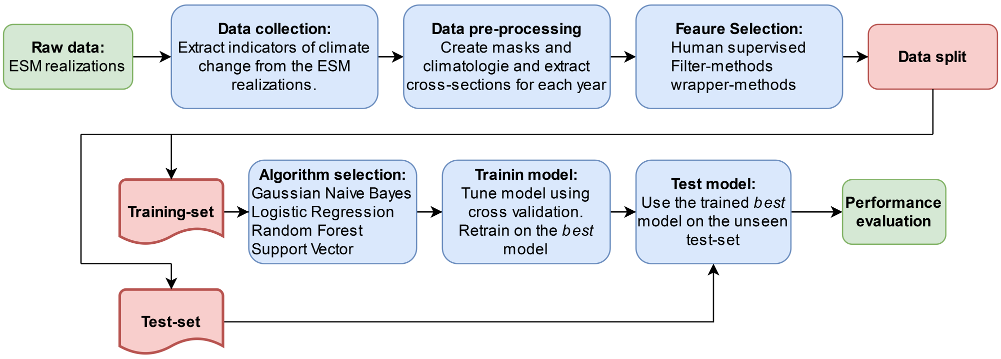


Figure 3.6: The supervised learning pipeline for this thesis. The data collection and creation is described in Section 2.4 and data pre-processing in Section 3.1. The remaining steps is accounted for throughout this section as well as the results and discussion.

3.2.1 About the classification problem [1]

One of the most central problems solved by ML-algorithms is the classification task, where the objective is to assign observations to discrete classes. Through the ML theory we refer to these as classes; however, they are in our specific case the scenarios SSP1-2.6 and SSP5-8.5. Given any observation x, and a classifier f, the following relation holds.

$$f(x) \mapsto \{0, 1\} \quad , \tag{3.4}$$

where $\mapsto$ denotes "maps to", and $\{0, 1\}$ will be classes [34]. For data that are assumed to be bimodal, that is, there are two class-clusters, equation (3.4) is adequate to describe the relation between classifier, observation, and class. Taking a step back and investigating the broader objective of ML we cite Goodfellow et al. (2017):

The factors determining how well a ML algorithm will perform are its ability to

1. *Make the training error small.*

2. *Make the gap between the training and test error small* [34].

The illustration in Figure 3.7 concisely represents how one, as practitioners, will balance goals 1 and 2. Moving from left to right, the figure refers to; under-fitting, "correctly"-fitting and over-fitting.

The balance between under-fitting and over-fitting is referred to as the bias-variance trade-off. A model with high bias (left pane of Figure 3.7) will miss the relations between features and the target variable, leading to under-fitting by *under-training*. This is often identified in the model evaluation as a high error in both the training- and test-set. On the other hand, a model with high variance will have a low loss in the training set due to "noise modeling", leading to over-fitting [36]. This is identified in the model evaluation as a large gap between the training and the test error. An over-fitted model will have a low generalization capacity, thus "breake-down" when tested on unseen data.

Balancing the bias-variance trade-off is done mainly through altering the model capacity, that is, the model's ability to learn complex relations. To reduce over-fitting the practitioner tunes hyperparameters [34]. All classifiers are

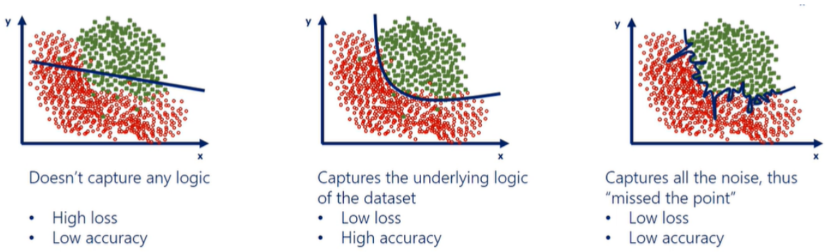


Figure 3.7: Example illustration of the bias-variance trade-off. The decision boundary represented by the blue line can be fitted in a vast array of configurations. A linear boundary (left) is not accurate enough and a perfect line (right) will not perform well on unseen data. Illustration from 365datascience.com [35].

tuned on the training set using 10-fold cross-validation to account for the low sample size. To reduce the size of the hyperparameter space, we perform an initial analysis using a randomized search that evaluates 100 randomly generated configurations within the defined search space [2]. Secondly, the tuning for the analysis presented in Chapter 4 is performed with a grid search over the smaller hyperparameterspace identified by the tune distributions of the random search [3]. Hyperparameters are different between each classifier and will therefore be presented in more detail through Section 3.2.3 alongside the model theory. In addition to employing well-tuned classifiers, some key concepts to ensure good generalization performance include data standardization, presented in Section 3.1, and the choice of evaluation metrics that fit the goal of the classifier and the data, discussed in Section 3.2.4. The model evaluation is performed on the test-set generated from the full data-set. In this thesis, we use a train-test split of 70/30 % with a random assignment of realizations to the two sets [4]. Given the small sample size for each cross-section, the distribution of the test-set across the data-set will differ, resulting in some random states being more challenging for classification than others. In addition, the splitting is done without knowing the number of realizations from each class. To mitigate this uncertainty, we run the full tune-train-test-evaluate pipeline for all algorithms through 50 random states for each cross-section. However, before any training is performed, we must reduce the dimension of the feature space through feature selection.

3.2.2 Feature selection

Model training, commonly known as *learning*, entails the systematic extraction of patterns from data-sets. When such learning processes are employed on high-dimensional data, challenges may manifest as most data-sets contain features that are either irrelevant, redundant, and/or noisy. Some key issues include the curse of dimensionality, under- and over-fitting, and considerable computational cost. Dimesionality reduction through feature selection is a key process of the ML-pipeline in order to mitigate these issues [37]. Since we are using ESM data for training, all realizations are labeled with classes; thus, we will use supervised algorithms. In Section 3.2.2, we will present the theory and strategies for feature selection used in this thesis' analysis. The first part includes metrics used for the selection of human-supervised features before moving on to algorithms, filter methods, wrapper methods, and traditional statistics. Embedded feature selection, commonly referred to as regularization, is included in Section 3.2.3 for the respective classifiers. All feature-sets contain k selected features and will be bench marked against the global means of all variables, that is,

```
nomask_baseline = ['fd: nomask', 'gsl: nomask', 'pr: nomask', 'tas: nomask', 'txx: nomask']
```

thus $k_{\text{nomask}} = 5$.

Since cross-sections develop through the century, different features will create different class separations at different times, as illustrated in Figure 3.5. Thus, the feature's importance and, consequently, its rank will be shifted from cross-section to cross-section. The ensamble time-series and earlier analysis show that we will have to assume a successful classification toward the later part of the near-term period, i.e. 2030–2040. Since it is somewhat artificial to choose features that perform the best on each individual cross-section (dynamic feature-sets, one selection per cross-section) in the context of using the model for real data classification, we must use static feature-sets. We will therefore, for

[2] In order to perform the random search, we use the scikit-learn implementation *sklearn.model_selection.RandomizedSearchCV()*, which allows for a random search and cross-validation simultaneously.

[3] In order to perform the grid search, we use the scikit-learn implementation *sklearn.model_selection.GridSearchCV()*, which allows for a random search and cross-validation simultaneously.

[4] In order to perform the train-test split, we use the split implementation from the scikit-learn library; *sklearn.model_selection.train_test_split()*, which allows for a fast and randomized distribution of realizations.

all methods except the embedded selectors, optimize the feature selection to support the best possible classification in 2030–2040.

3.2.2.1 Metrics used for human feature selection

In order to assess the feature usefulness, we will use a combination of visual analysis of data distributions, knowledge of the climate system, and inter-feature correlation. The distribution analysis is based of asserting the features of temporal evolution through time-series, such as those in Figure 3.2 and 3.3, as well as density plots of cross-sections in the near-term period, examples include box and violin plots, kernel density plots, and pairwise feature scatter plots. The goal of the visual analysis is to assert a) which features have a strong trend and small amount of noise, that is, they have a high SNR, and b) which features interact beneficially to create class separation.

The correlations between features x_1 and x_2 are calculated using the Pearson correlation of equation (3.5) where $\mathbb{E}$ denotes the expectation value, μ_j and σ_j is the mean of the class and the standard deviation, respectively. Correlation is an important metrics in ML because the learning is more effective if the inter-feature correlation is small, while the correlation to the class/scenario is large. And because some classifiers assume the independence between features, which is indicated by $\rho_{x_1,x_2} = 0$ [38].

$$\rho_{x_1,x_2} = \frac{\text{cov}(x_1, x_2)}{\sigma_1\sigma_2} = \frac{\mathbb{E}[(x_1 - \mu_1)(x_2 - \mu_2]}{\sigma_1\sigma_2} \tag{3.5}$$

3.2.2.2 Filter methods

Filter methods are performed independently of the model learning algorithm and therefore rely solely on the data characteristics. These methods involve a two-step process: 1) ranking the p features of the original data-set $f_1, f_2, ..., f_p$ according to an evaluation criterion that can be univariate or multivariate, and 2) filtering out low-ranked features or selecting a set of features by the practitioner, either as a defined number of features k or as a percentage of features. However, without the guidance of a specific learning algorithm, the k selected features $f_1, f_2, ...f_k$ might not be optimal for the target algorithms [37].

F-score values

The F score values are a filter method based on univariate statistics, that is, for each feature f_i the F-value is calculated. The F-value is the test statistics of a one-way (univariate) analysis of variance (ANOVA) which employees the null hypothesis H_0: The mean of the population class is the same across all classes. In order for the results of the ANOVA to be valid, we assume. 1) independent samples, 2) each sample is from a Gaussian distributed population, and 3) the population standard deviation is equal across all classes [38]. If violations are made, the test statistics have some loss of power. The F-value is calculated as the ratio between the mean square of the class MS$_{class}$ and the mean square of the error (residuals) MS$_{error}$ as

$$F_{\text{value}}(f_i) = \frac{MS_{class}(f_i)}{MS_{error}(f_i)} \quad . \tag{3.6}$$

Here, MS_{class} is the proportion of the total data-set variance between classes (SS_{class}) divided by its degrees of freedom. Likewise, MS_{error} is the proportion of total total data-set variance associated with variability within classes (SS_{error}) divided by its degrees of freedom. For more mathematical details, we refer to a derivation of one-way ANOVA and the decomposition of the total sum of squares [38]. The resulting feature selection is therefore based on the assumption that the feature f_i with a high statistical probability of class separation is the best feature. We implement the F-value feature selection by calculating the univariate F-values of the cross-section and returning the ranked features and their F-value [5].

Mutual information scores

The mutual information score is a univariate filter method based on the information gain from a feature, by maximizing the feature relevance and minimizing the redundancy of the feature [37]. This method is at it's core based on information theoretical concepts which are discrete. In order to estimate the mutual information between features we have to first consider the concept of entropy as a measure of uncertainty in a discrete random variable X. For an arbitrary random variable X, the entropy is defined by

[5] In order to perform the f-value feature selection, we use the scikit-learn implementation *sklearn.feature_selection.f_classif()*.

$$H(X) = -\sum_{x_t \in X} P(x_t) \log(P(x_t)) \quad , \tag{3.7}$$

where $P(x_t)$ denotes the probability of observing x_t over all values of X. As a result, the entropy measures the expected information conveyed by realizing a random trial, alternatively; how much uncertainty remains after the realization. For dependent random variables X and Y, the conditional entropy of X is

$$H(X \mid Y) = -\sum_{y_j \in Y} P(y_j) \sum_{x_t \in X} P(x_t|y_j) \log(P(x_t|y_j)) \quad . \tag{3.8}$$

That is, the remaining uncertainty of X given Y. Letting Y represent the target class, the information gain or mutual information maximization (MIM) is calculated by

$$J_{MIM}(X_i) = I(X_i; Y) = H(X_i) - H(X_i|Y) \quad . \tag{3.9}$$

That is, the MIM of an unselected feature X_i is calculated by the difference between the entropy and the conditional entropy of X_i given the class target [37]. As such, equation (3.9) uncovers the amount of information shared by the feature and the class.

The above theory is the derivation of the information for a discrete feature X_i; therefore, our continuous features have to be discretized by the algorithm. The entropy estimation of continuous features is less straightforward; for details, we refer to Kraskov et al. (2004) [39]. The resulting feature selection is therefore based on the assumption that the feature f_i with a strong correlation to the class is the best feature. We implement the MIM feature selection by estimating the information gain the feature in each cross-section and returning the ranked features and their MIM score [6].

Minimum Redundancy Maximum Relevance

The information gain of MIM algorithm is univariate and therefore does not consider feature interactions. To address this, the minimum redundancy maximum relevance (mRMR) algorithm penalizes features that share high mutual information with previously selected features $X_j \in \mathcal{S}$, where $\mathcal{S}$ is the set of selected features. The information gain of a discrete feature is calculated in terms of entropy as

$$J_{mRMR}(X_i) = I(X_i; Y) - \frac{1}{|\mathcal{S}|} \sum_{X_j \in \mathcal{S}} I(X_i; X_j) \quad . \tag{3.10}$$

Here, as more features are selected, the effect of feature redundancy $I(X_i; X_j)$ is gradually reduced [37, 40]. For continuous features, we approximate the information gain using the F-value of the variable. For the redundancy estimation, the Pearson correlation between the feature and the previously selected features are assessed assuming a high absolute value indicates redundancy. Letting y denote the class vector, the information gain of mRMR is estimated by the continuous score function for f_i,

$$\text{score}(f_i) = \frac{\text{F}_{\text{value}}(f_i, y)}{\sum_{f_j \in \mathcal{S}} |\rho(f_i, f_j)| / (|\mathcal{S}| - 1)} \quad , \tag{3.11}$$

where f_j represents the j-th feature of the previously chosen set of features $\mathcal{S}$ and $|\mathcal{S}|$ is the size of $\mathcal{S}$. This method is known as the F-test correlation quotient. We implement the mRMR feature selection by estimating the information gain *with* redundancy penalization of the cross-section and returning the ranking as well as the feature importance itself [7].

[6] In order to perform the MIM feature selection, we use the scikit-learn implementation *sklearn.feature_selection.mutual_info_classif()*. In accordance with the method of Kraskov et al. the only tunable hyperparameter of *mutual_info_classif()* is the number of neighbors used in the estimation; it is, however, left at the default value $h = 3$.

[7] In order to perform the mRMR feature selection, we use the implementatio from the mrmr-selection package, *mrmr_selection.mrmr_classif()*. There are no tunable hyperparameters if you don't want to return a subset of k features.

3.2.2.3 Wrapper methods

The previously defined filter methods determine the usefulness of a feature based on how relevant they are to the target, following the algorithm-specific criterion. As a result, the feature ranking is uniform whether we employ logistic regression, neural networks, or any other classifier. On the other hand, wrapper methods depend on the classifier's predictive ability to determine feature usefulness. There are several evaluation criteria. The Z-values (ratio of mean to standard deviation) from the decline in classification accuracy were used in the first iteration of Boruta. SHapley Additive exPlanations (SHAP) is used in Boruta (SHAP) to determine the features' usefulness. However, for wrapper methods, the focus is less on the evaluation criterion and more on the algorithm itself. For further information on how importance is assessed, see Kursa (2010) for the original method or Lundberg (2017) for the SHAP approach [41, 42].

Boruta(SHAP)

In traditional feature selection the feature ranking is returned from the algorithm presenting the relationship between features under the respective assumptions. However, the practitioner decides the threshold for which a feature is viewed of high usefulness or not. Boruta assesses this threshold, thus removing the uncertainty of a practitioner choice and presenting a *all-relevant* set of features [41].

Let $X \in \mathbb{R}^{r \times p}$ denote the design matrix which contains all p features as columns and r realizations as rows. Before passing X to the classifier, all columns (features) are copied and shuffled randomly (in terms of realizations within the copied column) and added as additional columns to X, forming $X_{boruta} \in \mathbb{R}^{r \times 2p}$. These added columns are termed *shadow features*. The classifier is then fitted and the importance is assessed. To establish the threshold at which the importance of a feature is considered sufficient for retention, it must exceed the highest score among the shadow features. Such features are then marked with a "hit" [41].

The trial is repeated N times since the random shuffle of the algorithm introduces some "luck" into whether a feature is significant or not. A trial may yield "hit" or "no-hit" for each feature, resulting in a probability of a hit of $p(\text{hit}) = 0.5$. Each feature f_i has h_i hits after N iterations, and for p features, N repetition will create a hit distribution that follows the binomial distribution. From the cumulative mass function of the bionomial distribution, the two-sided symmetric confidence interval $[m_q(N), M_q(N)]$ is constructed, for which three hypotheses can be formed: H_0 we do not know if f_i is useful, H_1: f_i is useful and H_2 f_i is not useful. If $m_q(N) < h_i < M_q(N)$ H_0 cannot be rejected and the feature is marked as tentative. Furthermore, if $h_i > M_q(N)$ we reject H_0 and assume H_1 and conversely if $h_i < m_q(N)$ we reject H_0 and assume H_2 leading to the loss of w features. Then the shadow features are removed and the new design matrix $X_{boruta_N} \in \mathbb{R}^{r \times 2(p-w)}$ is formed to repeat the process [41]. The number of iterations can be set to an arbitrary number of iterations or until all features are either accepted or tossed. If N is a fixed number, tentative features (not accepted or rejected) can be rejected or accepted according to the practitioner's choice.

In this thesis, we run boruta(SHAP) with the implementation from the BorutaShap package [8]. To mitigate the stochastic nature of the algorithm, we run it with *'n_trials'*=100 iterations for each cross-section. All features that are tentative at the end of the iteration are rejected. We built the boruta algorithm around the random forest (RF) classifier as was done in the original implementation [41]. Since the forest is only meant to aid feature selection, it is roughly tuned using a 3 fold cross-validated random search over 50 configurations within,

```
param_grid = {
    'n_estimators': [int(x) for x in np.linspace(start = 200, stop = 2000, num = 5)],
    'max_depth': [None] + [int(x) for x in np.linspace(10, 110, num = 5)],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
    }
```

For details on the RF and tuning see Section 3.2.3.

3.2.2.4 The MANOVA and ANOVA f-scores (classical statistics)

After feature selection is performed, we will have multiple feature-sets $f_{\text{set}} = f_1, f_2, ...f_k$. To confirm that the set allows for an early separation between classes, we run ANOVA. For these cross-sections, it is appropriate to use a one-factorial experimental design where the class is the describing factor. Further, we let $f_1, f_2, ...f_k$ describe the class; therefore, we perform a multivariate analysis of variance (MANOVA) as opposed to the univariate test statistics of equation (3.6). This choice is appropriate because the relations between the features span the room in which we perform our classification; hence, the interactions are important.

[8] In order to perform the Boruta(SHAP) feature selection we use the BorutaShap implementation *BorutaShap.BorutaShap()*.

The null hypothesis H_0 of MANOVA is that there is no significant difference in the means of the features across classes. Reversibly, the alternative hypothesis is that we will observe a difference in classes based on the feature combination. To assess the hypothesis, test statistics are needed; however, for MANOVA the test statistics are no longer the simple F_{value} of equation (3.6). There are multiple test statistics commonly used for MANOVA; in this thesis, we will use the Wilks Lambda calculated by,

$$\Lambda^* = \frac{|E|}{|H+E|} \quad , \tag{3.12}$$

where the E matrix represents the *within* class variance analog to SS_{error}, and the H matrix represents the *between* class variance, analog to SS_{class} [43]. For further details on the Wilks Lambda and the approximation of its F-value, we refer to these lecture notes [43]. To assess the significance of the F-value, we set the p-value at $\alpha = 0.05$ to reject the null hypothesis. Since we only use two classes, no post hoc analyses are needed; we already know which classes are different. Since we have new cross-sections for each year we will perform the MANOVA for each cross-section to assess when the p-value converges for the different feature combinations. We perform the MANOVA test using the implementation provided by the statsmodel package [9]. The MANOVA assumptions are the multivariate analog to those of the univariate ANOVA.

To relate our results to previous work on SSP separation from Tebaldi et al. (2021), we also calculate the p-value of univariate ANOVA for GSAT (*'tas: nomask'*). This is done by calculating equation (3.6) with the statsmodel implementation of ANOVA [10].

3.2.3 Classification models [1]

Generative and discriminative classifiers represent two fundamentally different approaches to the classification task. Generative classifiers model the joint probability distribution of the input features and the target labels, allowing them to generate new instances. On the other hand a discriminative classifier models the conditional probability of the target label given the input features, focusing solely on the boundary between classes. In this section, we will present four commonly used classifiers, the Gaussian naïve Bayes classifier, logistic regression, the random forest classifier and the support vector classifier.

Gaussian naïve Bayes Classifier

The naïve Bayes classifier is a supervised learning algorithm, conducting estimations of class probabilities grounded in the principles of Bayes' theorem. In contrast to the tree other classifiers, the naïve Bayes classifier uses a generative approach. In the context of continuous variables, exemplified by indicators of climate change, the application of the Gaussian naïve Bayes classifier (GNB classifier) is employed.

In probability theory, Bayes' theorem describes the probability of events based on known prior probabilities [38]. Let y_j be the class variable from the sample space of the scenarios $\mathcal{S}$. If $\mathcal{S}$ holds k scenarios, we assume that the probabilities of the scenario are $P(y_j) > 0$ and $\sum_j P(y_j) = 1$ for $j = 1, ..., k$. For binary classes of the *scenario* sample space, we have $\mathcal{S} = \{\text{SSP1-2.6}, \text{SSP5-8.5}\}$. Furthermore, $\mathbf{x} = x_1, ..., x_p$ is a realization capable of taking into account the values of $\mathcal{S}$. Then the posterior probability of y_j given a random arbitrary realization of an unknown class $\mathbf{x}$ is described by Bayes' theorem as

$$\begin{aligned} P(y_j \mid x_1, ..., x_p) &= \frac{P(y_j \cap x_1, ..., x_p)}{P(x_1, ..., x_p)} \\ &= \frac{P(x_1, ..., x_p \mid y_j) \cdot P(y_j)}{P(x_1, ..., x_p)} \quad . \end{aligned} \tag{3.13}$$

Based on equation (3.13), the posterior probability $P(y_j \mid x_1, ..., x_p); j \in \{1, ..., k\}$ is calculated using: The prior probability $P(y_j)$, that is, the probability of a realization $\mathbf{x}$ belonging to y_j, and the conditional probabilities $P(x_1, ..., x_p \mid y_j)$. Since we conduct the experiment in a supervised learning setting, $P(x_1, ..., x_p \mid y_j)$ is a known set of probabilities for $y_j \in \mathcal{S}$. In addition, the probability of observing the realization $x_1, ..., x_p$ is given by,

[9] In order to perform the MANOVA analysis, we use the statsmodel implementation *statsmodels.multivariate.manova.from_formula()* to calculate the test statistics, including the p-value used to check the statistical significance of the class mean separation.

[10] In order to perform the ANOVA analysis, we use the statsmodel implementation of ordinary least squares (OLS), *statmodels.formula.api.ols()*, to fit the linear model of the ANOVA analysis. Furthermore, to calculate the test statistics (including the p-value used to check the statistical significance of the class mean separation), we use *statsmodels.api.stats.anova_lm()*.

$$P(\mathbf{x}) = \sum_{j=1}^{k} P(x_1, ..., x_p \mid y_j)P(y_j) \quad , \tag{3.14}$$

if we write it in terms of prior probabilities [38, 44]. In the naïve bayes classifier, we assume conditional independence between every pair of features within each scenario of $\mathcal{S}$; this is referred to as the *naïve* assumption as real-world data rarely are independent [45]. Using the naïve assumption, $P\left(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right) = P\left(x_i \mid y\right)$ for $i \in [1, p]$ yielding the simplified expression,

$$P(y_j \mid x_1, ..., x_p) = \frac{P(y_j) \prod\limits_{i=1}^{p} P(x_i \mid y_j)}{\sum\limits_{j=1}^{k} P(x_1, ..., x_p \mid y_j)P(y_j)} \quad , \text{ for } j \in [1, k]. \tag{3.15}$$

The classification is now performed by choosing the class j with the highest probability. For binary classification, the maximum a posterior (MAP) scenario, and thus the estimated scenario is given by

$$y_{MAP} = \text{arg}_{\mathrm{y_j}} \max P(y_j) \prod_{i=1}^{p} P(x_i \mid y_j) \quad , \tag{3.16}$$

where the prior $P(\mathbf{x}) = \sum\limits_{j=1}^{k} P(x_1, ..., x_p \mid y_j)P(y_j)$ is removed as it is constant between scenarios and thus does not alter the proportionality of probabilities [44]. To apply the classifier to continuous features, we assume that the distribution $P(x_i \mid y_j)$ follows the Gaussian density function of equation 3.17.

$$P\left(x_i \mid y_j\right) = \frac{1}{\sqrt{2\pi\hat{\sigma}_{y_j}^2}} \exp\left(-\frac{\left(x_i - \hat{\mu}_{y_j}\right)^2}{2\hat{\sigma}_{y_j}^2}\right) \tag{3.17}$$

The classifier obtained by inserting (3.17) into (3.16) is known as the GNB classifier. To summarize we use Figure 3.8, which shows a schematic overview of the inner workings of the univariate GNB classifier using classes $y1, y2$ and one realization x (ignoring the subscript i for the features as we are in the univariate distribution). Firstly, the density functions of equation (3.17) are parameterized from the class mean and standard deviation, which are estimated by equation (3.18) and (3.19) respectively. Here we let n be the number of realizations r in the class.

$$\hat{\mu}_{y_j} = \frac{1}{n} \sum_{r=1}^{n} \mathbf{x}_r \quad , \tag{3.18}$$

$$\hat{\sigma}_{y_j} = \sqrt{\frac{1}{n} \sum_{r=1}^{n} (\mathbf{x}_r - \hat{\mu}_{y_j})^2} \quad . \tag{3.19}$$

For an unseen realization x the unknown probability $P(x \mid y_j)$ is estimated by equation (3.17) where the distance from the mean represented by the squared z-scores scales the probability. The probability of observing x within y_j is then estimated by equation (3.16).

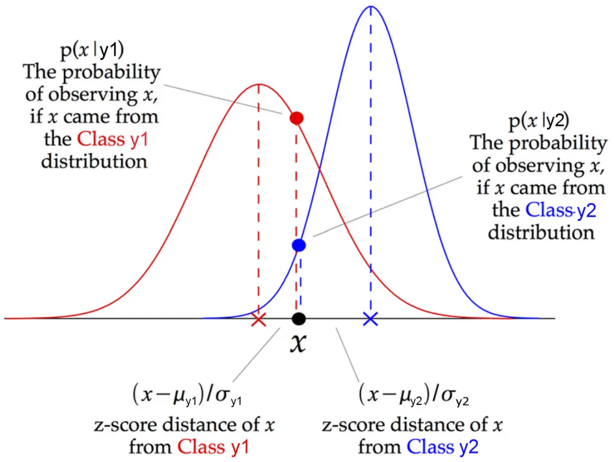

p(x | y1)
The probability of observing x, if x came from the Class y1 distribution
p(x | y2)
The probability of observing x, if x came from the Class-y2 distribution
x
$(x - \mu_{y1})/\sigma_{y1}$
z-score distance of x from Class y1
$(x - \mu_{y2})/\sigma_{y2}$
z-score distance of x from Class y2

Figure 3.8: Schematic overview of the parameters in the GNB classifier algorithm. Illustration adapted from medium.com/@Kashishdafe [46].

The GNB classifier is employed utilizing the estimator provided by scikit-learn[11]. The only tunable hyperparameter, *'var_smoothing'*, is subject to optimization through a cross-validated grid search within [np.logspace$(0, -9, \text{num} = 100)$], encompassing 100 logarithmically spaced values ranging from 10^{-9} to 1. To present the usage of *'var_smoothing'* we recall the naïve assumption of the algorithm, "all pairs of features are conditionally independent". If this assumption

[11] In order to create the GNB classification model we use the *sklearn.naive_bayes.GaussianNB()* estimator from scikit-learn.

is violated, the variance ratio between features is too small, causing numerical errors [47]. This is addressed by *'var_smoothing'*, as this is the proportion of the standard deviation of the largest feature which is added to all features to make them less dependent. The effect is a widening of the Gaussian density function, effectively increasing probabilities for realizations further from the mean, hence smoothing. For details on mathematical calculations, we refer to the source code of github.com/scikit-learn [47]. In summary, the GNB model for each cross-section and random state is found using a cross-validated grid search of

```
param_grid = {
      'var_smoothing': [np.logspace(0, -9, num=100)]
      }
```

The *best* model, i.e. the highest performing model on the training-set, is re-fitted with the full training-set and evaluated using the test-set.

Logistic Regression

Logistic regression is a descriminative supervised ML classifier. At its core, logistic regression (LR) is a specification of the general linear model and thus uses the features of a realization $\mathbf{x}$ to predict a class likelihood. The learning part of the algorithm is done by fitting a logistic function to each feature i of the p-dimensional vector $\mathbf{x} \in \mathbb{R}^p$. This is done through maximum likelihood estimations as opposed to the least squares fitting used in *linear* regression. In this section, we will explore the underlying mathematics of LR as well as accounting for assumptions made during the derivation.

The general linear model of equation 3.20 expresses how the target $\mathbf{y} \in \mathbb{R}^r$ responds to the input values contained in the design matrix $\mathbf{X} \in \mathbb{R}^{r \times p}$.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ &= \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p + \epsilon \quad . \end{aligned} \tag{3.20}$$

This model assumes a mapping from the continuous *real* axis of the p features by the estimators β to the *real* continuous axis of $\mathbf{y}$. However, in the binary classification problem, the target $\mathbf{y}$ is discrete, as shown in Figure 3.9. The class variable y_j is now in the sample space $\mathcal{S} = \{y_1, y_2\}$ with $P(y_j) > 0$ for $j = 1, 2$ and $P(y_1) + P(y_2) = 1$. In order to use the general linear model on $\mathcal{S}$, we have to perform the transformation of $\mathcal{S} \to \mathcal{S}^* = \mathbb{R}$ with the goal of expressing the class probability $p \in [0, 1]$ as a function of features $\in \mathbb{R}$. Transformation is performed using the *logit function*, which maps the values of p to the real space. Some key values are mapped as,

$$\begin{cases} p = 1 & \to \text{logit}(1) = \log 1 - \log 0 = \infty \\ p = 0.5 & \to \text{logit}(0.5) = \log 1 = 0 \\ p = 0 & \to \text{logit}(0) = \log 0 = -\infty \end{cases} \quad ,$$

thus ensuring that y covers the full real space $[-\infty, \infty]$. Since all realizations were originally found at $p = 0$ or $p = 1$, they now reside at $-\infty$ or ∞, respectively. The second-order linear model is then fitted, which yields the relation $\text{logit}(p) = \beta_0 + \beta_1 x_r$. Before transforming back with the sigmoid function equation, yielding the probabilities,

$$\begin{aligned} P(y_r = 1 \mid x_r) &= \frac{e^{\beta_0 + \beta_1 x_r}}{1 + e^{\beta_0 + \beta_1 x_r}} \\ P(y_r = 0 \mid x_r) &= 1 - P(y_r = 1 \mid x_r) \end{aligned} \quad . \tag{3.21}$$

$\mathbf{y}$ is now in $\mathcal{S}$. Thus, the probabilities of equation (3.21) are also the likelihoods that we aim to maximize using maximum likelihood estimation. The total likelihood of observing the data-set given the estimators $P(\{(y_r, x_r\} \mid \beta)$ is estimated by the product of all individual likelihoods of equation (3.21), thus yielding equation (3.22),

$$P(\{(y_r, x_r\} \mid \beta) = \prod_{r=1}^{n} [P(y_r = 1 \mid x_r, \beta)]^{y_r} [1 - P(y_r = 0 \mid x_r, \beta)]^{1 - y_r} \quad . \tag{3.22}$$

Taking the logarithm of the expression, the product is reduced to a sum. The error function $\mathcal{C}$ of the LR is the negative log-likelihood, called *cross-entropy*,

$$\mathcal{C}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left(y_i\left(\beta_0 + \beta_1 x_i\right) - \log\left(1 + \exp\left(\beta_0 + \beta_1 x_i\right)\right)\right) + r(\beta) \quad , \tag{3.23}$$

where $r(\beta)$ is a penalty term added for regularization purposes. In this thesis, we use the l2 regularization $r(\beta) = \frac{1}{2}||\beta||_2^2 = \frac{1}{2}\beta^t\beta$, where β^t is the transposed vector of the estimators.

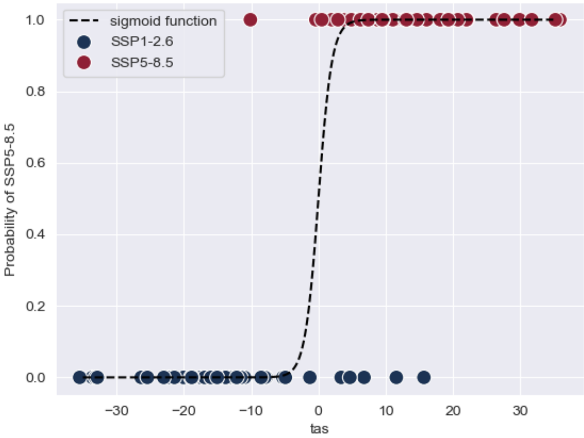


Figure 3.9: Example univariate cross-section in the probability-feature space. The fitted sigmoid function is used to map new realizations to a class probability.

The LR is implemented using the implementation provided by the scikit-learn library, for which we have tuned the regularization strength *'C'* and the *'solver'* of the minimization problem [12]. To alter the regularization strength, we tune *'C'*$\in [100, 10, 1.0, 0.1, 0.01]$, which is the inverse of the regularization strength. Thus, a smaller value specifies a stronger regularization. The *'solver'* is used to minimize the cost function of equation (4), thus estimating its derivatives. Here we have tested *'newton-cg'* and *'liblinear'* solvers; for more details, we refer to the documentation of the scikit-learn library [48]. In summary, the logistic regression model for each cross-section and random state is found using a cross-validated grid search of,

```
param_grid = {
    'C': [100, 10, 1.0, 0.1, 0.01],
    'solver': ['newton-cg',
               'liblinear']
}
```

The *best* model, i.e. the highest performing model on the training-set, is re-fitted with the full training-set and evaluated using the test-set.

Random Forest classifier

Decision trees form the basis for the ensamble method random forest (RF); however, they are also valuable classifiers in themselves. In contrast to regression methods like logistic regression, but similar to the support vector classifier, decision trees are non-parametric, meaning they do not make assumptions concerning the relation between features and targets, which makes them very flexible. The structure of the decision tree is shown (as part of a random forest) in Figure 3.10. The tree consists of nodes in three categories: a root node, internal nodes, and leaf nodes. Connecting these are branches. When one “growes” a decision tree, the gini impuirity g, defined in equation (3.24), is used for the decisions of the optimal split.

$$g = \sum_{i=0}^{K} p_i(1 - p_i) \tag{3.24}$$

Here p_i is the frequency of observations of class i at the node, and K is the number of unique classes. The gini impurity is used at the root of the tree and at every internal node; thus, each of the nodes is to be considered a decision point.

We use the CART algorithm to grow the three. In short, CART looks at one feature, f_i, and a threshold value that f_i must meet in order for the split to occur at t_k. We find the minimized error function by using the ginifactor g,

$$C(f_i, t_k) = \frac{n_{left}}{n} g_{left} + \frac{n_{right}}{n} g_{right} \quad , \tag{3.25}$$

[12] In order to create the LR model, we use the *sklearn.linear_model.LogisticRegression()* estimator from scikit-learn.

which will be used recursively for each internal node until a stopping criterion is met or there are only lief nodes left. In this way, the CART algorithm will perform feature selection since it is optimizing with the gini impurity. However, using this approach, CART is very sensitive to the data sample, and a small change can create a larger change in tree construction [49]. This is referred to as an unstable tree.

If stopping criteria are not implemented, a decision tree will perform splits until there are only lief nodes left. This makes decision trees prone to over-fitting. To combat this, there are common pruning techniques that can be deployed, all of which are categorized into one of two categories: pre-pruning and post-pruning. Pre-pruning involves stopping the growth of the tree early, often using heuristics such as setting a minimum number of training instances for each leaf or limiting the depth of the tree. Post-pruning, on the other hand, involves growing a full tree and then removing nodes and trees to reduce complexity and improve generalization. We will only be conducting pre-pruning by tuning the maximum depth of the decision tree using *'max_depth'*$\in [2,3,5]$ and the minimum number of samples required for a node to split using *'min_samples_leaf'*$\in [4,6,8,10,12,20]$.

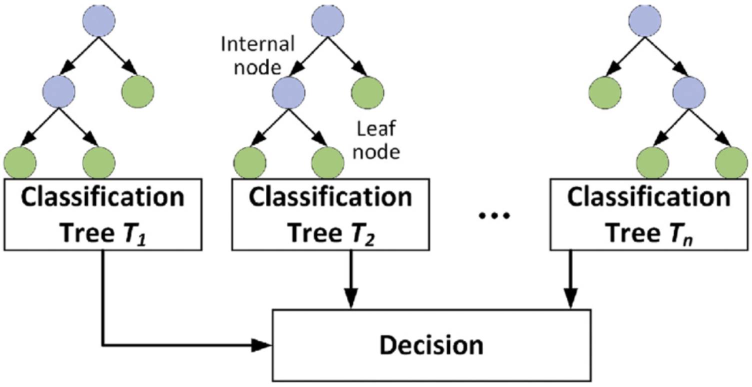

Internal node
Leaf node
Classification Tree T_1
Classification Tree T_2
...
Classification Tree T_n
Decision

Figure 3.10: Schematic overview of a random forest as an ensamble of decision trees. In the random forest classifier an ensamble of decision trees are utilized. The majority vote of the trees is used as the final prediction. Illustration from Zhang et. al (2020 [50]

As mentioned above, the decision tree algorithm is prone to over-fitting. To minimize the risk of over-fitting random forest uses bootstrap aggregation (bagging). Figure 3.10 illustrates how multiple classification trees "vote" use majority voting for the class decision. This is a key property of the bagging technique. The bagging algorithm performs splits of the training set with replacement. One decision tree is then trained on the subset; further, introducing the ensamble to the test-set the majority vote is used as the classification.

Until now, decision trees with bagging have been introduced, leading to an ensemble approach. Strict hyperparameterspaces for pre-tuning will result in trees that are similar, which negates the original intent of ensembling [51]. The number of estimators, or the total number of trees in your forest, is the only extra hyperparameter for decision trees with bagging. We will use *'n_estimators'*$\in [30,50,100,200,500]$ to find the best configuration. To further improve generalization performance, we alter the number of samples that each tree is allowed to check when searching for the optimal split, as given in equation 3.25. Now we have the random forest, with a random subset for training through bagging and a reduced number of searches allowed. We tune the maximum number of features by *'max_features'*$\in ['sqrt','log2']$.

The random forest is implemented using the estimator provided by the scikit-learn library [13]. This allows us to investigate the optimal subspace for the tree-specific hyperparameters at the same time as the forest-specific hyperparameters. In summary, the random forest model for each cross-section and random state is found using a cross-validated grid search of

```
param_grid = {
     'n_estimators': [30, 50, 100, 200, 500],
     'max_depth': [2, 3, 5],
     'min_samples_leaf': [4, 6, 8, 10, 12, 20],
     'max_features': ['sqrt', 'log2']
     }
```

The *best* model, i.e. the highest performing model on the training-set, is re-fitted with the full training-set and evaluated using the test-set.

[13] In order to create the RF model we use the *sklearn.ensemble.RandomForestClassifier()* estimator from scikit-learn.

Support vector classifier

The support vector classifier (SVC) is another supervised learning algorithm for binary classification. The concept of a SVC is the fitting of an optimal hyperplane to separate data points with the largest possible margin. Let one realization be denoted as x in the p dimensional space, we assume, with a linear kernel, that the binary classes are separable into two classes $y_j = \pm 1$. For a feature space of dimension p the separating hyperplane is on of $p-1$ dimensions. Then a linear hyperplane is defined using the general linear model as

$$\mathbf{x} \cdot \tilde{\mathbf{w}} + \tilde{b} = 0 \quad , \tag{3.26}$$

where $b \in \mathbb{R}$ is the intercept and $\tilde{\mathbf{w}}$ is the weights of the hyperplane $\in \mathbb{R}^p$. If the condition of equation (3.26) is not met, we have a classification $y_j = sgn(\mathbf{x} \cdot \tilde{\mathbf{w}} + \tilde{b})$ as illustrated in Figure 3.11. The placement of the hyperplane is, indeed, the learning part of the support vector classifier.

For a robust classifier, the intuitive approach is placing a hyperplane that maximizes the margin M. Equation (3.27) presents the constraint for which we maximize M.

$$y_i \left(\boldsymbol{x}_i \cdot \tilde{\boldsymbol{w}} + \tilde{b} \right) \geq M \quad \text{or,} \quad y_i \left(\boldsymbol{x}_{\boldsymbol{i}} \cdot \boldsymbol{w} + b \right) \geq 1 \tag{3.27}$$

Note that now $||\boldsymbol{w}|| = 1/M$, for the right-hand version, we have defined $\boldsymbol{w} := \frac{\tilde{\boldsymbol{w}}}{M}$ and $b = \frac{\tilde{b}}{M}$. The closest observations relative to the hyperplane are referred to as *support vectors* are illustrated by the points on the dashed lines in Figure 3.11.

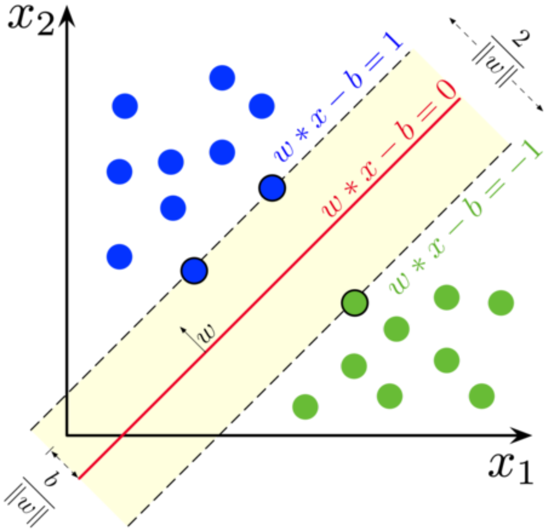


Figure 3.11: Schematic overview of a linear support vector classifier. For a two dimensional featurespace the linear kernel is just a linear line function. The line separates the classes with a margin w and at the edge of the margin we find the support vectors. Illustration from wikipedia.org.

For an easy separable case as Figure 3.11 one could usually perform clustering with a hard-margin SVC, however, hard margins do not allow for wrongful classifications. For more complex relations, this will lead to over-fitting as illustrated in the right-hand subplot of Figure 3.7 (though one does not longer use a linear kernel).

For data where noise is more present, a soft-margin SVC, which allows a proportion of samples (slack) to be on the wrong side of the hyperplane, will be more appropriate [52]. Letting ξ_i be the distance between the slacked samples and the correct class margin, we introduce the soft-margin SVC as

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\frac{1}{n}\sum_i \xi_i$$
$$\text{subject to} \begin{cases} y_i(\boldsymbol{x} \cdot \boldsymbol{w} + b) \geq (1 - \xi_i) & \text{for } i = 1, \ldots, n \\ \xi_i \geq 0 & \text{for } i = 1, \ldots, n \end{cases} \tag{3.28}$$

for some scalar C. C signifies the “cost” of the error allowed by the soft margins. Analyzing equation (3.28) it is apparent that a smaller C will allow a greater proportion of points to be inside the margins, thus broadening the margin. In order to implement the SVC we use the implementation provided by scikit-learn library [14]. For this implementation C is the most important hyperparameter and will be tuned within the range *'C'* $\in [1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02]$ in an effort to improve generalization performance. To fit non-linear hyperplane, we use the kernel trick, which in essence transforms input data via the mapping $\phi : \boldsymbol{x} \mapsto \mathbb{R}^n$ to increase class-cluster separation. Mathematically, we substitute the kernel function $K(\boldsymbol{x}, \boldsymbol{w}) = \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{w})$ for the inner product $\boldsymbol{x} \cdot \boldsymbol{w}$ of equation (3.28). In this thesis, the kernel settings are adjusted using *'kernel'* $\in \{'linear', 'rbf', 'poly', 'sigmoid'\}$. The introduction of non-linear kernels, *'poly'*, *'rbf'* and *'sigmoid'*, necessitates the tuning of the hyperparameter γ, specified by *'gamma'* $\in \{['scale', 'auto']\}$. The *'gamma'* hyperparameter influences the decision boundary's flexibility, thereby managing the bias-variance trade-off. A low value of *'gamma'* makes the decision boundary smoother, thus regularizing the optimization problem, while a high

[14] Scikit-learn provides the *sklearn.svm.SVC()* estimator to create a SVC model.

value of *'gamma'* tends to make the decision boundary more complex, potentially leading to over-fitting. The strength is governed by either a) *'scale'*$=1/(n_features * X.var())$ where *X.var()* is the total variance of the design matrix X (introduced for boruta), or b) *'auto'*$=1/n_features$. Thus, for a data-set with variance $= 1$ *'scale'*$=$*'auto'*. For *'scale'* a higher data-set variance will result in a smaller *'gamma'* and consequently a stronger regularization. Additionally, the *'degree'*$\in \{2, 6, 10\}$ is adjusted for the *'poly'* kernel, similar to tuning the polynomial degree in linear regression. For the *'linear'* kernel, the degree is automatically 2. In summary, the suppor vector model for each cross-section and random state is found using a cross-validated grid search of

```
param_grid = {
    'C': [1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02],
    'kernel': ['linear', 'rbf', 'poly', 'sigmoid'],
    'gamma': ['scale', 'auto'],
    'degree': [2, 6, 10]
    }
```

The *best* model, i.e. the highest performing model on the training-set, is re-fitted with the full training-set and evaluated using the test-set.

3.2.4 Model evaluation [1]

The performance of binary classifications is commonly summarized in a confusion matrix. A generic confusion matrix is shown in Figure 3.12 where the correct classifications, true negative (TN) and true positive (TP) are found on the diagonal. In addition, the confusion matrix holds the false negative (FN) and false positive (FP), both of which are considered errors.

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Figure 3.12: Example Confusion Matrix. The rows hold the true/actual class of the realization, while the columns hold the predicted class by the model.

Based on the cells in the confusion matrix, we can define multiple metrics regarding the performance of the model. We will start by briefly introducing four common metrics and their usage.

The *accuracy* is a measure of the number of correct classifications

$$\text{Accuracy} \;= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad . \tag{3.29}$$

Since it uses all cells of the confusion matrix it is a good choice for data-sets where the classes are well balanced.

The *precision* is the ratio of predictions set to positive that are actually positiv as calculated in equation (3.30). Therefore, precision is a powerful metric for data where the FP is of greater importance then the FN. Further the *recall* is used to measure how many of actual positive samples where classified wrongly as negatives FN as calculated in (3.31).

$$\text{Precision} \;= \frac{\text{TP}}{\text{TP} + \text{FP}} \quad , \tag{3.30}$$

$$\text{Recall} \;= \frac{\text{TP}}{\text{TP} + \text{FN}} \quad . \tag{3.31}$$

The *F1-Score* combines the properties of equation (3.30) and (3.31) into equation (3.32).

$$\text{F1-Score} \;= 2 * \frac{(\text{ Recall } * \text{ Precision })}{(\text{ Recall } + \text{ Precision })} \tag{3.32}$$

Combining the properties, the F1 score balances the weight between classifying wrongfully (FN, FP). It is a common metric for unbalanced classes and is thus powerful when a test-train split is performed between classes without a balance verification [53].

In the thesis, we use the F1-score to inform on the tuned model for each random state. This is done since the train-test split might be unbalanced with respect to number of realizations per class, commonly referred to as class prevalence. Further, to assess the result, we will use the accuracy. This choice is taken on the basis that; a) the consequence of an error is equally as serious for both types of error; b) due to the usage of 50 random states the classes of the will be close to balanced. As a result, the average F1 score and the accuracy across the 50 random states have shown similar relations between the train and test-set as well as close to the same magnitude in initial investigations. The accuracy is then a more natural choice as it is easier to interpret.

Further evaluation and comparison of the model is performed using the Area Under Curve (AUC) metric calculated from the receiver operating characteristic curve (ROC curve). While the *precision* and *recall* utilize the columns of the confusion matrix we define the true positive rate (TPR) and false positive rate (FPR) by the rows, that is:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3.33}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \quad . \tag{3.34}$$

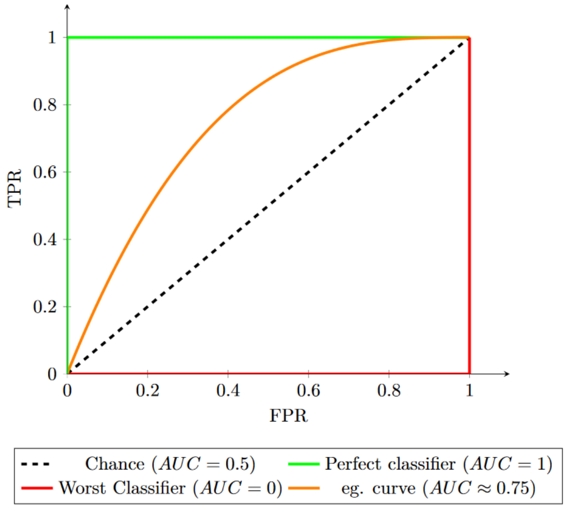


Figure 3.13: Example ROC plot for a perfect, stochastic and worst possible classifier.

Therefore, the TPR is concerned with the treatment of actual positives by the model, and the FPR is concerned with actual negatives. Figure 3.13 displays a schematic ROC curve. Here we note that the TPR is used on the y-axis and the FPR is used on the x-axis. In contrast to the column-derived metrics, the ROC curve is robust to the prevalence changes caused by the random state, as the axes do not interact [54].

The curves of the ROC curve are generated by iteratively testing the model performance at different thresholds of class probability $\in [0, 1]$ for classification. As indicated by the "perfect classifier" of Figure 3.13, a good classifier operates further to the top-left corner as it is capable of achieving a high TPR at low FPR. The point of the curve with the most preferable TPR to FPR ratio can also be used to decide the perfect threshold. However, for this thesis, it is preferable to keep the threshold at a minimum of 0.5 in order to reduce the model bias. In our case, we used 50 random states for our classification; this is also true for the ROC analysis. Since the analysis is performed using the scikit-learn library, the ROC curve is generated using the *predict_proba()* function. This generates the ROC curve at different thresholds for different random states. To average these 50 curves, we have therefore linearly interpolate the TPR of all curves to preset values of FPR $= [0.0, 0.1, \ldots, 0.9, 1.0]$ [48].

The last but also most important reason to use the ROC curve is the AUC metric which allows for fast model comparison. The AUC has the main benefit of being threshold invariant, therefore a model with high AUC is robust. The AUC of a model is given $\in [0, 1]$, as indicated by Figure 3.13 the greater the AUC the better the model.

Chapter 4

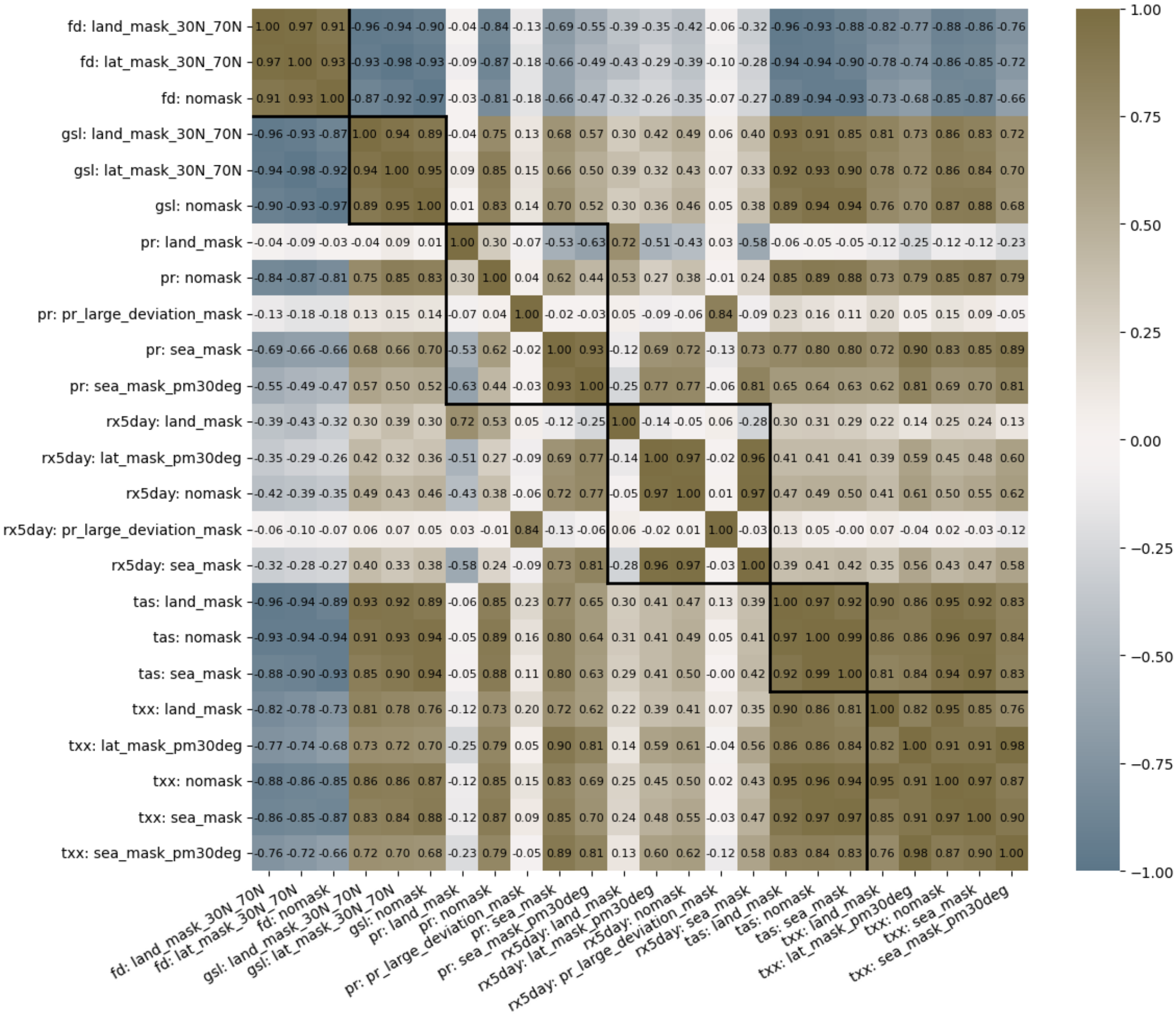
Results and Discussion

4.1 Feature selection

This section presents the results of the feature selection process using the methods described in Section 3.2.2. The set of features available for the selection process is given in Table 3.3. Recall that the goal of feature selection was to identify the features that contribute most effectively to predictive accuracy while maintaining model simplicity. The results highlight the optimal subset of features, given the assumptions of the method. We also want to highlight that we refer to a global-level indicator of climate change as a variable (in this case, tas, pr, txx, rx5day, fd, and gsl), while a masked version (including nomask) is referred to as a feature.

4.1.1 Human supervised selection

The inter-feature correlations for SSP1-2.6 are shown in Figure 4.1. This is the average correlation over the near-term period computed from the scenario ensemble. Because there are only slight differences in correlation between the scenarios, only one scenario is assessed. However, note that the largest differences in the correlation between scenarios are found in features related to precipitation. This underpins the higher noise in precipitation as a variable, consistent with the findings of the IPCC and Tebaldi et al. (2021) [1, 9].



	fd: land_mask_30N_70N	fd: lat_mask_30N_70N	fd: nomask	gsl: land_mask_30N_70N	gsl: lat_mask_30N_70N	gsl: nomask	pr: land_mask	pr: nomask	pr: pr_large_deviation_mask	pr: sea_mask	pr: sea_mask_pm30deg	rx5day: land_mask	rx5day: lat_mask_pm30deg	rx5day: nomask	rx5day: pr_large_deviation_mask	rx5day: sea_mask	tas: land_mask	tas: nomask	tas: sea_mask	txx: land_mask	txx: lat_mask_pm30deg	txx: nomask	txx: sea_mask	txx: sea_mask_pm30deg
fd: land_mask_30N_70N	1.00	0.97	0.91	-0.96	-0.94	-0.90	-0.04	-0.84	-0.13	-0.69	-0.55	-0.39	-0.35	-0.42	-0.06	-0.32	-0.96	-0.93	-0.88	-0.82	-0.77	-0.88	-0.86	-0.76
fd: lat_mask_30N_70N	0.97	1.00	0.93	-0.93	-0.98	-0.93	-0.09	-0.87	-0.18	-0.66	-0.49	-0.43	-0.29	-0.39	-0.10	-0.28	-0.94	-0.94	-0.90	-0.78	-0.74	-0.86	-0.85	-0.72
fd: nomask	0.91	0.93	1.00	-0.87	-0.92	-0.97	-0.03	-0.81	-0.18	-0.66	-0.47	-0.32	-0.26	-0.35	-0.07	-0.27	-0.89	-0.94	-0.93	-0.73	-0.68	-0.85	-0.87	-0.66
gsl: land_mask_30N_70N	-0.96	-0.93	-0.87	1.00	0.94	0.89	-0.04	0.75	0.13	0.68	0.57	0.30	0.42	0.49	0.06	0.40	0.93	0.91	0.85	0.81	0.73	0.86	0.83	0.72
gsl: lat_mask_30N_70N	-0.94	-0.98	-0.92	0.94	1.00	0.95	0.09	0.85	0.15	0.66	0.50	0.39	0.32	0.43	0.07	0.33	0.92	0.93	0.90	0.78	0.72	0.86	0.84	0.70
gsl: nomask	-0.90	-0.93	-0.97	0.89	0.95	1.00	0.01	0.83	0.14	0.70	0.52	0.30	0.36	0.46	0.05	0.38	0.89	0.94	0.94	0.76	0.70	0.87	0.88	0.68
pr: land_mask	-0.04	-0.09	-0.03	-0.04	0.09	0.01	1.00	0.30	-0.07	-0.53	-0.63	0.72	-0.51	-0.43	0.03	-0.58	-0.06	-0.05	-0.05	-0.12	-0.25	-0.12	-0.12	-0.23
pr: nomask	-0.84	-0.87	-0.81	0.75	0.85	0.83	0.30	1.00	0.04	0.62	0.44	0.53	0.27	0.38	-0.01	0.24	0.85	0.89	0.88	0.73	0.79	0.85	0.87	0.79
pr: pr_large_deviation_mask	-0.13	-0.18	-0.18	0.13	0.15	0.14	-0.07	0.04	1.00	-0.02	-0.03	0.05	-0.09	-0.06	0.84	-0.09	0.23	0.16	0.11	0.20	0.05	0.15	0.09	-0.05
pr: sea_mask	-0.69	-0.66	-0.66	0.68	0.66	0.70	-0.53	0.62	-0.02	1.00	0.93	-0.12	0.69	0.72	-0.13	0.73	0.77	0.80	0.80	0.72	0.90	0.83	0.85	0.89
pr: sea_mask_pm30deg	-0.55	-0.49	-0.47	0.57	0.50	0.52	-0.63	0.44	-0.03	0.93	1.00	-0.25	0.77	0.77	-0.06	0.81	0.65	0.64	0.63	0.62	0.81	0.69	0.70	0.81
rx5day: land_mask	-0.39	-0.43	-0.32	0.30	0.39	0.30	0.72	0.53	0.05	-0.12	-0.25	1.00	-0.14	-0.05	0.06	-0.28	0.30	0.31	0.29	0.22	0.14	0.25	0.24	0.13
rx5day: lat_mask_pm30deg	-0.35	-0.29	-0.26	0.42	0.32	0.36	-0.51	0.27	-0.09	0.69	0.77	-0.14	1.00	0.97	-0.02	0.96	0.41	0.41	0.41	0.39	0.59	0.45	0.48	0.60
rx5day: nomask	-0.42	-0.39	-0.35	0.49	0.43	0.46	-0.43	0.38	-0.06	0.72	0.77	-0.05	0.97	1.00	0.01	0.97	0.47	0.49	0.50	0.41	0.61	0.50	0.55	0.62
rx5day: pr_large_deviation_mask	-0.06	-0.10	-0.07	0.06	0.07	0.05	0.03	-0.01	0.84	-0.13	-0.06	0.06	-0.02	0.01	1.00	-0.03	0.13	0.05	-0.00	0.07	-0.04	0.02	-0.03	-0.12
rx5day: sea_mask	-0.32	-0.28	-0.27	0.40	0.33	0.38	-0.58	0.24	-0.09	0.73	0.81	-0.28	0.96	0.97	-0.03	1.00	0.39	0.41	0.42	0.35	0.56	0.43	0.47	0.58
tas: land_mask	-0.96	-0.94	-0.89	0.93	0.92	0.89	-0.06	0.85	0.23	0.77	0.65	0.30	0.41	0.47	0.13	0.39	1.00	0.97	0.92	0.90	0.86	0.95	0.92	0.83
tas: nomask	-0.93	-0.94	-0.94	0.91	0.93	0.94	-0.05	0.89	0.16	0.80	0.64	0.31	0.41	0.49	0.05	0.41	0.97	1.00	0.99	0.86	0.86	0.96	0.97	0.84
tas: sea_mask	-0.88	-0.90	-0.93	0.85	0.90	0.94	-0.05	0.88	0.11	0.80	0.63	0.29	0.41	0.50	-0.00	0.42	0.92	0.99	1.00	0.81	0.84	0.94	0.97	0.83
txx: land_mask	-0.82	-0.78	-0.73	0.81	0.78	0.76	-0.12	0.73	0.20	0.72	0.62	0.22	0.39	0.41	0.07	0.35	0.90	0.86	0.81	1.00	0.82	0.95	0.85	0.76
txx: lat_mask_pm30deg	-0.77	-0.74	-0.68	0.73	0.72	0.70	-0.25	0.79	0.05	0.90	0.81	0.14	0.59	0.61	-0.04	0.56	0.86	0.86	0.84	0.82	1.00	0.91	0.91	0.98
txx: nomask	-0.88	-0.86	-0.85	0.86	0.86	0.87	-0.12	0.85	0.15	0.83	0.69	0.25	0.45	0.50	0.02	0.43	0.95	0.96	0.94	0.95	0.91	1.00	0.97	0.87
txx: sea_mask	-0.86	-0.85	-0.87	0.83	0.84	0.88	-0.12	0.87	0.09	0.85	0.70	0.24	0.48	0.55	-0.03	0.47	0.92	0.97	0.97	0.85	0.91	0.97	1.00	0.90
txx: sea_mask_pm30deg	-0.76	-0.72	-0.66	0.72	0.70	0.68	-0.23	0.79	-0.05	0.89	0.81	0.13	0.60	0.62	-0.12	0.58	0.83	0.84	0.83	0.76	0.98	0.87	0.90	1.00

Figure 4.1: The inter-feature correlations for masked features. Correlations are calculated from the scenario ensemble and represent the mean correlation throughout the near-term period. Thicker black squares outline features generated from the same variable.

The thicker black squares in Figure 4.1 outline the features generated by the same variable. We observe a high absolute value for the correlation between all features generated by temperature. In addition, we generally observe an even higher correlation between features of the same variable. Since many classification algorithms either a) assume feature independence or b) perform worse when features are highly correlated, the correlation matrix suggests that a feature-set should not contain multiple features from the same variable if it is temperature-based. The precipitation features show more variability between features, suggesting that keeping more features derived from the same variable will not affect the performance due to the high correlation.

To assess its usefulness for classification, we investigate distribution plots. Here we will only present some key findings; for a complete insight into visualizations, we refer to *06b - feature selection.ipynb* available in the code (see Appendix A.6). A major component is the analysis of time series, such as those in Figure 3.2 or 3.3. Many of the temperature-based features show similar temporal development within the variable; thus, from a visual analysis of the time-series, they are not easy to rank. For precipitation-based features, some masks are easier to rule out because they do not show any SSP separation; an example is seen in Figure 3.3 for *pr: land_mask*. The remaining ones are again hard to rank based on the time-series.

The most effective visualization has been pairplots, consisting of kernel density plots and scatter plots. In Figure 4.2, we see an example that allows comparing the usefulness of the features. The data show the temporal climatology across 2030–2040, the key period for classification. On the diagonal, we see the density distribution, visualizing the inter-scenario separation of the feature. In general, a feature with high separation is favorable, but as illustrated by *pr: sea_mask* a feature can also interact with other features, allowing for a better separation. The contours shown in the scatter plots of the lower triangle are kernel density estimates seen from above. As such, it visualizes the base of the kernel curve. Counting the realizations that are in the union of the scenario-contours thus allows for an easy estimate of feature usefulness.

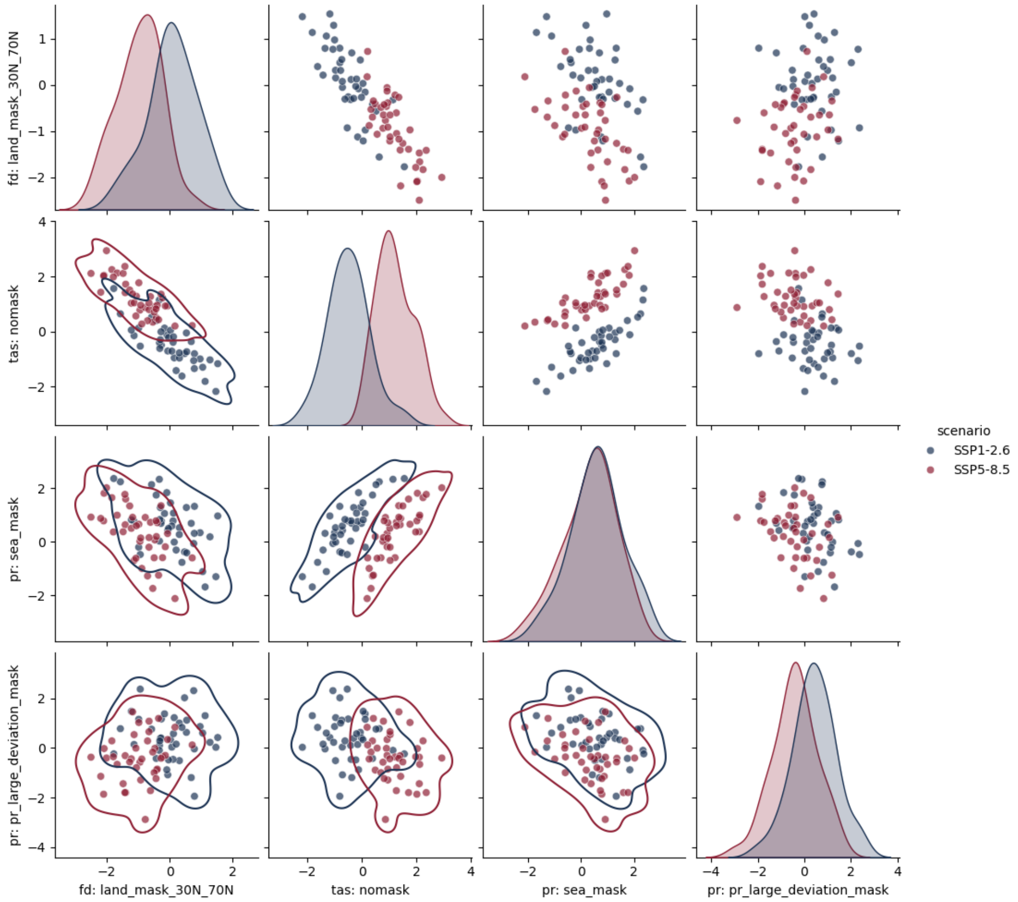


Figure 4.2: Example pairplot allowing for visual analysis of feature usefulness. The data shows the temporal climatology for each realization across 2030-2040. On the diagonal we see the kernel density plot which illustrates the signal to noise ratio as well as how far the scenario separation has developed. In the remaining cells we see the realizations placement in the two dimensional space spanned by the respective features. The contours shown in the scatter plots of the lower triangle are kernel density estimates seen from above.

Choosing the optimal feature set is difficult when based solely on visual analysis. In an attempt to perform human supervised feature selection, we only used one feature from each temperature-based variable in order to avoid highly correlated features. The selection and reasoning are as follows:

- fd and gsl features are highly correlated with all temperature-based features, and in addition gsl shows an unidentified dynamic in its' time-series. Therefore, we include *fd: nomask* as the only feature of these two variables. This is because the ensamble time-series seems to have the highest SNR and the pairplot yields most realizations outside of the contours.
- From precipitation-based features we chose *pr: nomask*, *pr: sea_mask* and *rx5day: land_mask*. This is because their pairplots span useful spaces in interaction with temperature-based variables and they are not highly correlated with each other.
- From the tas features, we chose *tas: nomask* as it interacts similarly to the other tas-features, but does, however, cover a larger area in the real world than the other features.
- From txx we chose *txx: sea_mask* as it shows the most promising separation based on contours.

To summarize, we get the supervised feature-set,

```
supervised_features = [
    'fd: nomask', 'pr: nomask', 'pr: sea_mask',
    'rx5day: land_mask', 'tas: nomask', 'txx: sea_mask'
    ]
```

thus $k_{\text{supervised}} = 6$.

4.1.2 Filter methods

Taking the ten best features into account, the results of the filter methods are consistent between all three methods. Table A.2 in Appendix A.2.3 displays the first ten features of each method. The findings indicate that the temperature-based features have the greatest predictive capacity, as none of the precipitation-based features are included among the ten. Figure 4.3 illustrates the cumulative feature importance of *mrmr_classif()* over the period 2030-2040, revealing that 14 of the top 24 features are related to temperature, while *pr: nomask* is the first precipitation-based feature in the 15th rank.

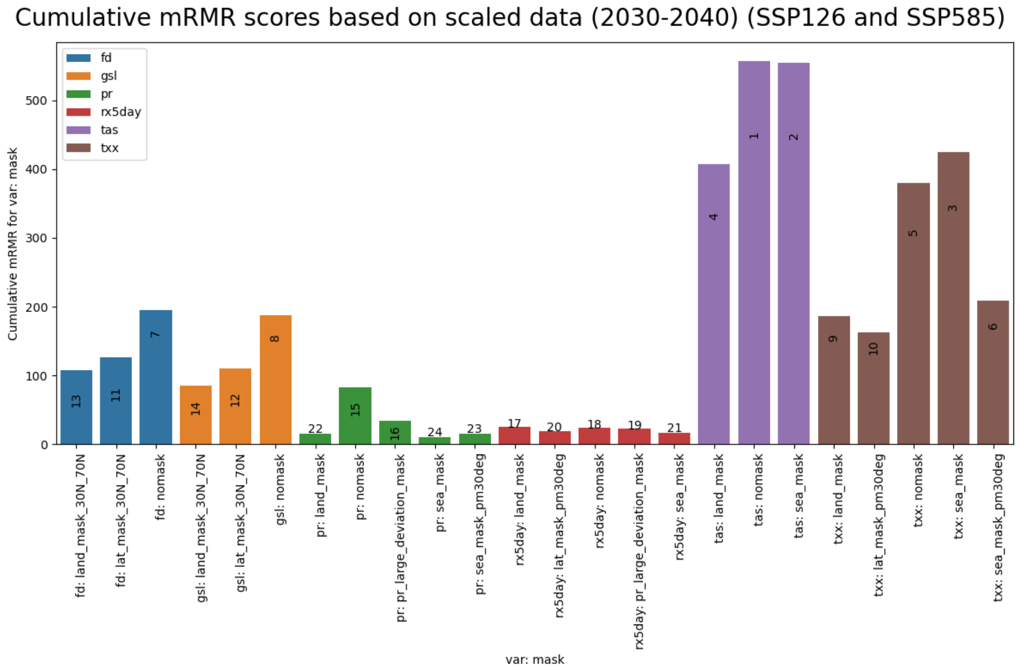


Figure 4.3: The cumulative importance score from mRMR feature selection across 2030-2040. The ranking is displayed in each bar while the color references the original variable.

The feature-set from the filter methods are,

```
mRMR_f_mut_features = [
        'fd: nomask', 'gsl: nomask', 'tas: land_mask', 'tas: nomask',
        'tas: sea_mask', 'txx: land_mask', 'txx: lat_mask_pm30deg',
        'txx: nomask', 'txx: sea_mask', 'txx: sea_mask_pm30deg'
        ]
```

thus $k_{\text{mRMR\_f\_mut}} = 10$. This selection reflects the features found to have the highest correlation towards the scenario.

4.1.3 Boruta(SHAP)

In Boruta, the algorithm self-calculates a threshold for which a feature is considered useful. Figure A.2 in appendix A.2.4 displays the feature importance in 2030 and 2040. In Figure A.2, the box-and-whiskers plots are generated from 100 iterations and show the distribution of the importance measured by the SHAP values. The blue boxes indicate the shadow features; here *shadowMax* is the benchmark feature for the real features to beat. Red boxes denote features that have been dismissed, green boxes signify features that have been approved, and yellow boxes indicate features that are tentative by the algorithm. All tentative features for a cross-section are automatically dismissed. From Figure A.2, we can clearly see that more features are approved as useful by 2040 than by 2030; this is expected from the visual analysis. In contrast to filter methods, Boruta accepts precipitation-based features as useful, with 3 of the 13 accepted features from pr-variable in 2040.

Figure 4.4 shows the acceptance frequency in 2030–2040. With pr-based features having a higher acceptance count than the features of fd and gsl, the difference between Boruta and the filter methods is evident. However, the highest-ranking features are generated from tas and txx.

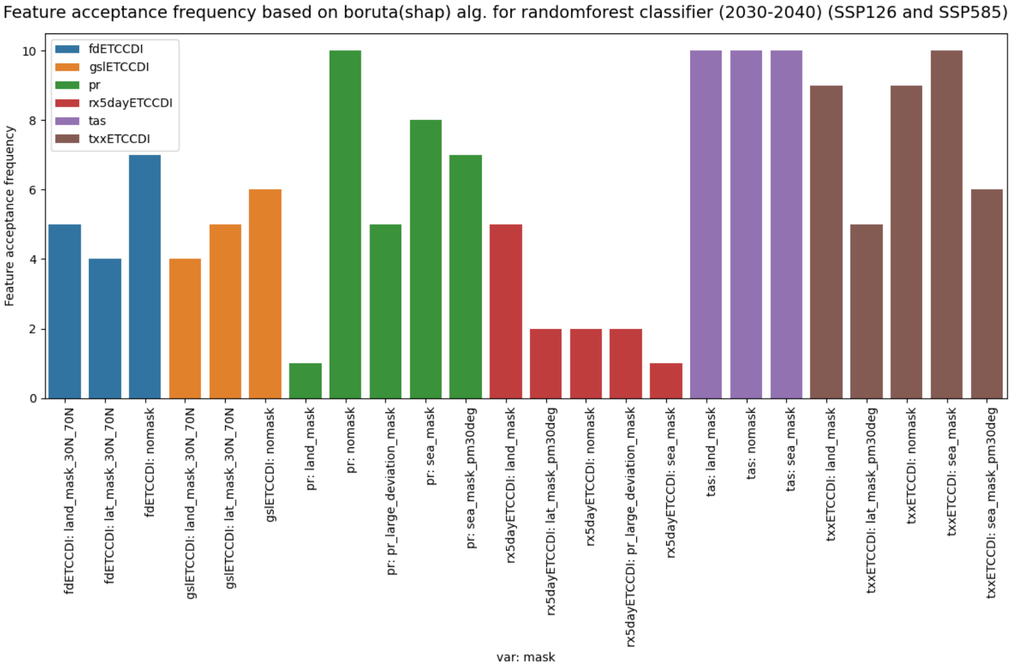


Figure 4.4: The acceptance count from Boruta feature selection across 2030-2040. The color references the original variable.

Here, we include all features with an acceptance count larger than or equal to 8. The resulting feature-set is,

```
boruta_RF_features = [
        'pr: nomask', 'pr: sea_mask', 'tas: land_mask', 'tas: nomask',
        'tas: sea_mask', 'txx: land_mask', 'txx: nomask', 'txx: sea_mask'
        ]
```

thus $k_{\text{boruta\_RF}} = 8$. This selection reflects the features that most frequently beet the shadowfeatures across 2030-2040.

4.1.4 MANOVA assessment

We now have four feature-sets: 1) *nomask_baseline*, which represents the benchmark features with the global version of all variables. 2) *supervised_features*, which are the human selected features based on visual analysis of the scenario distributions within each feature, inter-feature correlation, and subject-specific knowledge. 3) *mRMR_f_mut_features*, which were the top ten features from three different filter methods. And 4) *boruta_RF_features*, which were the features with an acceptance rate higher than 80 % between 2030 and 2040.

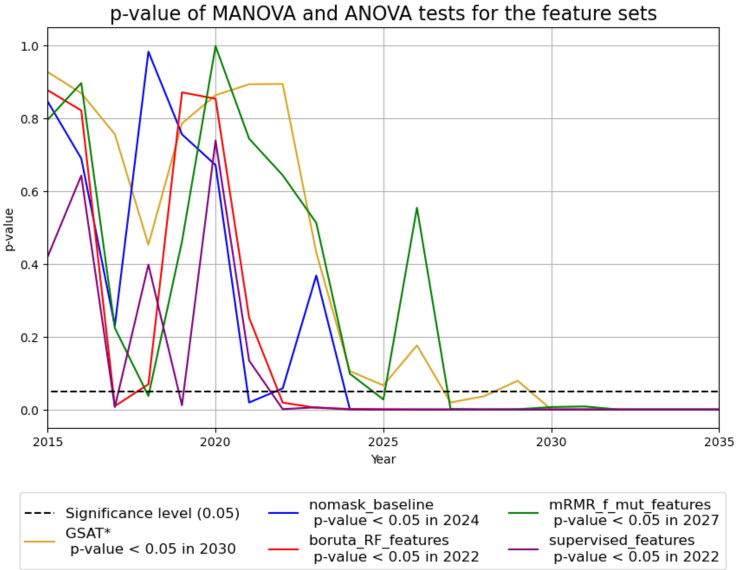


Figure 4.5: The temporal development of MANOVA Wilks lambda p-value for feature-sets and ANOVA p-value for GSAT. GSAT is labeled with as *GSAT** as it is the only set which has a p-value calculated from the test statistic of the univariate ANOVA. Solid lines indicate respective feature selections, while the dashed line indicates a significance level of 0.05. The legend marks the first year where the p-value falls and stays below 0.05.

To assess the expected performance of the feature-set, we performed MANOVA and collected the p-value from the Wilks Lambda test statistics. Figure 4.5 shows the temporal development of the p-value for each feature-set. The p-values indicate the probability of observing the distribution of realizations if the scenarios have the same mean. Thus, a significant p-value, here indicated at 0.05, tells us that the feature-set has differences in means and is suitable for classification. Similarly to the approach presented by Tebaldi et al. (2021), the ANOVA and MANOVA p-values do not have any predictive capabilities in themselves. From the figure, we observe a high fluctuation in the first years; however, all feature-sets relatively early converge below 0.05. From the MANOVA test statistics, it is expected that the boruta- and human-defined feature-sets will have the best predictive abilities, while the baseline- and filter-defined feature-sets will have a delayed development in comparison.

Furthermore, Figure 4.5 also shows the p-value of the univariate ANOVA for GSAT. This is calculated to benchmark the convergence of the feature-sets against the findings of earlier work on SSP separation. As indicated in the legend, the GSAT p-value stays below the 0.05 threshold from 2030 onward. Cross-referencing Table 2.3, we observe the same year of separation reported for the multi-model mean, although the evaluation metrics are not the same.

4.2 Classification

4.2.1 Overview

The estimated classification accuracy for all classification models and feature-sets are presented in Table 4.1. The estimate is based on the classifications of the test-set and is calculated by the mean ± standard deviation across 50 random states and years$\in [2035, 2040]$. These years represent the end of the near-term period for which it is crucial to identify the most effective policies to achieve the goals of the Paris agreement. The highest estimated accuracy of each model is highlighted in red.

The highest estimated classification accuracy for all classifiers is achieved through the utilization of *supervised_features*. With an accuracy of 0.83 ± 0.08, 0.97 ± 0.04, 0.85 ± 0.08, and 0.96 ± 0.04 for GNB classifier, LR, RF, and SVC, respectively. Further, we observe the expected ranking of the feature-sets based on the MANOVA p-value, with

Table 4.1: The estimated classification accuracy from each classifier and feature-set. The accuracy is reported as mean ± standard deviation calculated from test-set classification across 50 random states and years$\in [2035, 2040]$. The highest estimated accuracy of each model is highlighted in red.

feature-set key	**estimated classification accuracy**			
	GNB	**LR**	**RF**	**SVC**
nomask_baseline	0.80 ± 0.09	0.89 ± 0.07	0.82 ± 0.08	0.88 ± 0.08
boruta_RF_features	0.82 ± 0.07	0.96 ± 0.04	0.84 ± 0.08	0.95 ± 0.04
mRMR_f_mut_features	0.80 ± 0.07	0.81 ± 0.09	0.80 ± 0.08	0.80 ± 0.09
supervised_features	0.83 ± 0.08	0.97 ± 0.04	0.85 ± 0.08	0.96 ± 0.04

boruta-, nomask/baseline-, and filter-determined features yielding diminishing accuracy. However, it should be noted that Boruta-determined features achieve a proximate performance of the *supervised_features*. Significant variations are observed in how the classifiers respond to different feature selections, with the GNB classifier showing a spread in estimated accuracies in all feature-sets ranging from $[0.80 - 0.83]$, while the SVC displays a larger range from $[0.80 - 0.96]$.

We will further examine the detailed results for the ROC curve analysis, the temporal development of the accuracy, and a detailed discussion of the algorithms in sections 4.2.2, 4.2.3, and 4.3, respectively.

4.2.2 ROC analysis

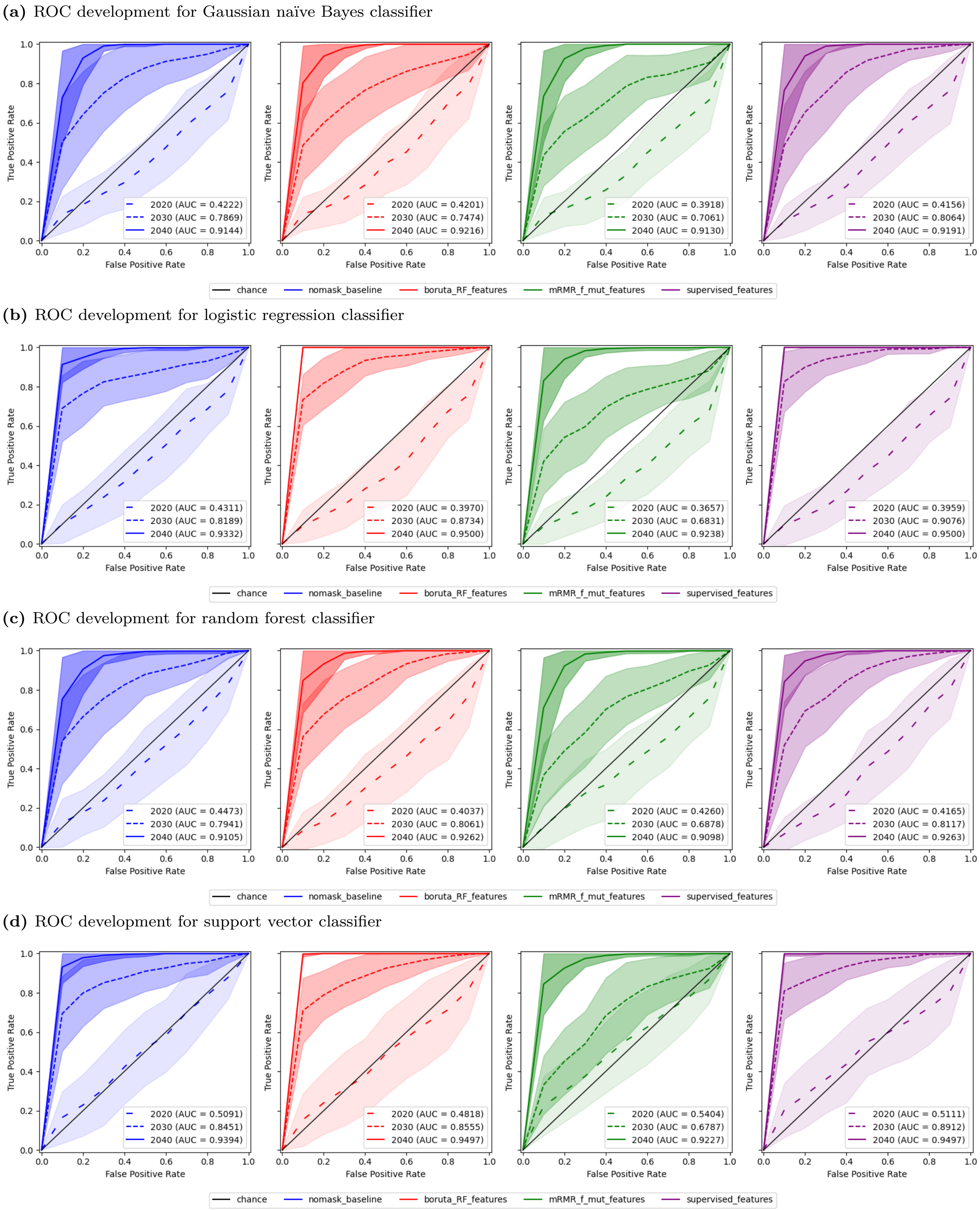
The main significance of the ROC curves is to investigate the robustness and expected performance of the classifiers. This is indicated mainly by the curve shape, which implies the effect of different classification thresholds, and by the AUC metric which measures the performance in a threshold-invariant matter. Figure 4.6 shows the ROC curves for three cross-sections year $\in \{2020, 2030, 2040\}$ for: (a) Gaussian naïve Bayes classifier (b) logistic regression classifier, (c) random forest classifier and (d) support vector classifier. The lines show the mean TPR/FPR ratio, with the shaded uncertainty range based on the standard deviation. Both parameters are calculated across 50 random states in the specific cross-section. All classifiers exhibit consistent temporal development with improvement in AUC.

In general, the 2020 curves lie below the stochastic classifier, with the AUC of the mean ROC curve predominantly falling below 0.5 or marginally exceeding this threshold for the SVC. This suggests that the classification capabilities in 2020 are at best equivalent to a random chance. In addition, no consistent pattern emerges in terms of the feature-sets that yield the most or least effective performance. This observation is substantiated by the *mRMR_f_mut_features* that generate the highest and lowest AUC values, specifically, $\text{AUC}_{\text{max},2020} = 0.5404$ attained by the SVC, and $\text{AUC}_{\text{min},2020} = 0.3657$ attained by the LR.

By 2030, there is a notable improvement in AUC with mean-curve AUC values in the range of $[0.6787, 0.9076]$. The best mean-curve AUC value, $\text{AUC}_{\text{max},2030} = 0.9076$, is attained by LR and *supervised_features*, whereas the minimum, $\text{AUC}_{\text{min},2030} = 0.6787$, is attained with SVC and *mRMR_f_mut_features*. Despite the significant rise in average AUC values across all classifiers, there is still a considerable amount of variability in the ROC curves shape and resulting AUCs, as highlighted by the shaded regions. This suggests that performance has not yet stabilized in various random states.

In 2040, all classifiers' AUC have been stabilized above 0.9. The highest values are attained by LR and SVC with $\text{AUC}_{\text{max},2040} \approx 0.95$ for both *bortua_RF_features* and *supervised_features*. In addition, the variance from the random states has been significantly reduced among all classifiers.

Although the displayed ROC curves is generated from only one cross-section and not a temporal climatology, it illustrates some inherent properties of the classifiers and the data. 1) The mean curves are smooth, indicating that there are no classification thresholds that are more favorable than others in all random states. 2) The general ranking of the feature-sets of the MANOVA p-values is also exhibited by the ROC curves. 3) Average classifiers are not capable of beating a stochastic classifier in 2020, indicating that stronger feature-sets are needed to perform a successful classification for the first years of the near-term period. 4) The uncertainty ranges are significantly reduced by 2040. This indicates that the random state has less impact than in 2030. This has a significant relevance for applications on real-world data, as we at this point have a low impact on the realizations position within the scenario cluster.



(a) ROC development for Gaussian naïve Bayes classifier

(b) ROC development for logistic regression classifier

(c) ROC development for random forest classifier

(d) ROC development for support vector classifier

Figure 4.6: Detailed presentation of classifier ROC curves. The lines show the mean TPR/FPR ratio, with the shaded area showing the uncertainty range based on the standard deviation. Both parameters are calculated across 50 random states. For all panels (a)-(d) the ROC curves for the feature-sets are presented for the respective classifiers. The color corresponds to the respective feature-set, while the diagonal chance line represents a stochastic classifier. The TPR values of the ROC curves are calculated from a linear interpolation of the probabilities to preset values of the FPR. This is done to enable the calculation of mean and uncertainty ranges across the random states, since *predict_proba()* does not generate the same amount of probabilities for each case. The AUC values is calculated based on the mean ROC curve after linear interpolation.

4.2.3 Accuracy development

Although the estimated classification accuracy for the later part of the near-term period is reported in Table 4.1, it is useful to understand the temporal development of the accuracy in more detail. Figure 4.7 shows the temporal development of the accuracy for: (a) Gaussian naïve Bayes classifier, (b) logistic regression classifier, (c) random forest classifier, and (d) support vector classifier. Lines show the mean accuracy, and the shaded area shows the uncertainty range based on the standard deviation. Both are calculated across 50 random states. In order to assess the role of over-fitting, we plot the training-set accuracy in gray. Lastly, as an informal indication of the development we reported the last time, the mean test-set accuracy was observed below 0.8 in the figure legends.

The 0.8 accuracy threshold is passed at different years based on the algorithm and feature-set. The general ranking of the feature-sets based on this crossing is *supervised_features*, *boruta_RF_features*, *nomask_baseline*, and *mRMR_f_mut_features*, with an average crossing year across algorithms of 2031, 2032, 2034, and 2038, respectively. Consequently, it is evident that the selection of feature-sets is critical in determining the effectiveness of the classification performance. Figure 3.5 illustrates how the effectiveness of various feature-sets for classification varies by year. In addition, Figure 4.7 shows the width of the uncertainty range, also called "random-state uncertainty". For all classifiers, the random-state uncertainty begins to taper early. This indicates that the random state has a lower impact on performance later in the near-term period, consistent with the findings in the ROC analysis.Year-to-year fluctuations in accuracy after detrending, or "feature-set noise", manifest at different time steps and magnitudes. This influences the smoothness of the mean accuracy curve. As a result, once past a certain threshold, some feature-sets will more consistently remain above it than others. It is important to note that a static feature set will exhibit some annual variations, as a feature that is highly useful in one cross-section will generate noise in another. To mitigate this effect, the algorithm should include regularization that penalizes features based on cross-section.

Similarly, we can rank the classifiers LR, SVC, RF, and GNB classifier with an average year of crossing of 2031, 2031, 2035, and 2038, respectively. The best model based on this metric is constructed using the *supervise_features* with LR and SVC, attaining a mean-line crossing of accuracy$\geq$ 0.8 in 2026. In general, the classifier determines the shape of the accuracy curves, where the SVC exhibits a rapid increase in growth around 2025, followed by a slowdown between 2030 and 2035, whereas the GNB classifier displays a consistent linear progression throughout the period. This is governed by the learning capacity of the model as well as the practitioner's ability to mitigate over-fitting.

Figure A.3 within Appendix A.3 displays the confusion matrix-equivalent to the results in Table 4.1. That is, the mean confusion matrices across 2035-2040 for the classifiers. The rows hold the true scenario, while the columns hold the predicted scenario. Within each matrix cell, the initial numerical value represents the mean count of classifications for that particular cell within the test-set. The sequential percentage is the percentage of the row that is held in the cell, thus a measure of correct treatments of the specific rows' scenario. The percentage in parentheses is the corresponding percentage for the training-set. The confusion matrices show an average of 12.38 and 11.62 realizations in SSP1-2.6 and SSP5-8.5, respectively, thus having a close to equal sample size per class and confirming the accuracy as a valid evaluation metric. To check for differences in the handling of the scenario, we refer to the test-set percentages. In general, the error rate is higher in SSP1-2.6 than in SSP5-8.5. This is consistent for all algorithms and feature-sets, suggesting that the models are biased toward SSP5-8.5. A bias towards one class is usually related to class imbalances; however, here we have an average sample size that is higher in SSP1-2.6 than in SSP5-8.5, which suggests that SSP1-2.6 is expected to have a higher accuracy. This inconsistency might be due to the intrinsic properties of the scenarios, with SSP1-2.6 being more complex relative to SSP5-8.5. However, additional investigation is required to determine the cause of the contradictory behavior.

4.3 Further discussion

4.3.1 The feature selection process

The feature selection process yielded four different feature-sets. From the MANOVA test statistics as well as the ROC-analysis for 2020, we see that early classification is most likely not attainable for the given feature-sets. In order to improve this, an improved masking process is the first subject to investigate. Although a relatively large number of masks were individually tested for each response variable, all choices were informed based on subjective metrics. This illustrates the full point of the feature selection process; a model can only *learn* what the data contains. So, in order to improve the classification accuracy, including in the early years of 2020, we will need to improve the SNR of the input features. A process suggestion is described in Section 5.2.

One of the purposes of the *nomask_baseline* set was to identify the importance of the masking process. Table 4.1 illustrates that, while masking is necessary, choosing the most useful features also requires a feature selection procedure.

(a) Accuracy development for Gaussian naïve Bayes classifier

(b) Accuracy development for logistic regression classifier

(c) Accuracy development for random forest classifier

(d) Accuracy development for support vector classifier

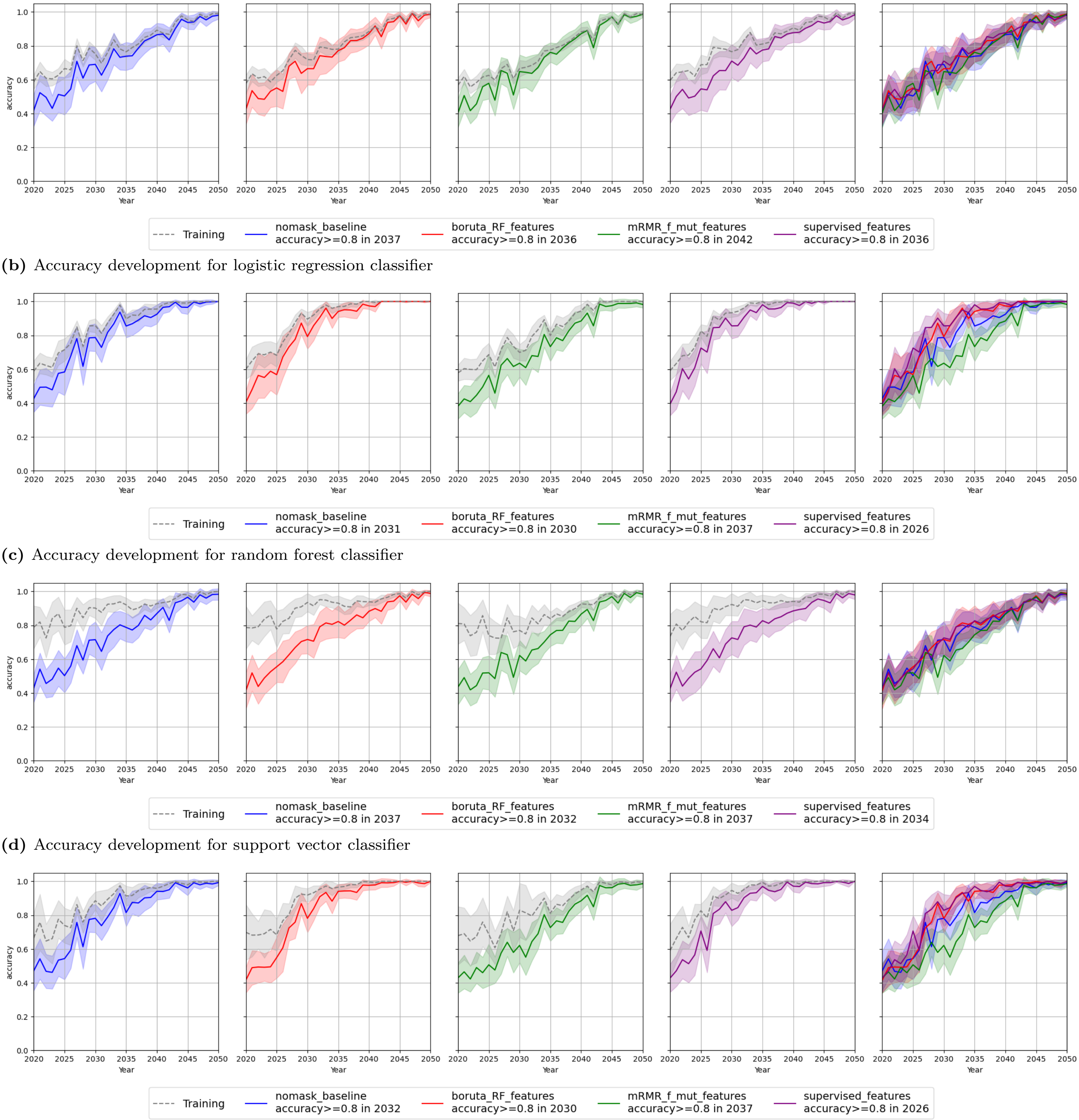


Figure 4.7: Detailed accuracy development of the classifiers. Lines show the mean accuracy, with the shaded area showing the uncertainty range based on the standard deviation. Both are calculated across 50 random states. For all panels (a)-(d) the accuracy development for all years$\in [2020, 2050]$ for the feature-sets are presented for the respective classifiers with a summarizing fifth panel. The color corresponds to the test-set accuracy of the respective feature-set, while the gray lines and shades hold the training-set accuracy. The last time the mean test-set accuracy is observed bellow 0.8 is reported in the legend.

Despite leaving room for improvement, the results from the baseline set are promising with the MANOVA p-values suggesting a separation of scenario means by 2024, which is many years earlier than the separation of the multi-model means of SSP1-2.6/SSP5-8.5 for the GSAT in 2030 reported by Tebaldi et al. (2021) [9]. Thus, the introduction of more indicators of climate change is beneficial for the classification. Furthermore, the results suggest that global observations can be classified with an accuracy of 80 % between 2031 and 2037 which is well within the limits of the near-term period.

The top ten features that consistently scored highest across the filter-based feature selection algorithms were used to create the feature selection set *mRMR_f_mut_features.* Although the evaluation criterion is different between the three algorithms, the fundamental assumption is the same: "a high feature importance makes it a useful feature". However, as illustrated in the results, the feature importance does not necessarily imply increased accuracy, a subtle but imperative difference. In addition, for F-values and MIM, the redundancy of features is not assessed at all. For mRMR, the redundancy is accounted for; however, it is likely that the entire feature-set is too small for the penalty of equation (3.11) to effectively reduce the feature-set. Since the full feature-set in this thesis is small, it is not certain that a high feature correlation will hurt the accuracy due to noise. However, a set of highly correlated features can, as it did here, end up explaining low amounts of the total dataset variance. As a result, the filter-defined feature-set has the overall lowest performance across all evaluated metrics. Considering that *mRMR_f_mut_features* lacked any precipitation-based features, the reported accuracy indicates that precipitation-based features are imperative for early scenario separation. Despite this, the feature-set defined by the filter method demonstrates a quicker convergence in the MANOVA p-value relative to the GSAT ANOVA p-value, as shown in Figure 4.5. This suggests potential advantages in incorporating subsets of both tas and txx into the model training, rather than relying solely on *tas: nomask* (which equates to GSAT after taking the global mean). Nevertheless, the addition of more features might also lead to increased noise, particularly in initial cross-sections. Evidence of this is seen in the significant variation in the p-value in 2026 shown in Figure 4.5, where *mRMR_f_mut_features* exhibits greater fluctuations compared to GSAT.

boruta_RF_features was established using the boruta algorithm relying on the predictive abilities of a tuned RF with SHAP values as indication of feature importance. In comparison to the filter-methods, boruta mitigate the user-specification bias by giving a statistical answer to what features are statistically useless. In all wrapper methods the ML model governs the feature importance. Thus, if the model is over-fitted the feature importance is calculated based on a model with low generalization capacity and the practitioners cut-off threshold will be incorrectly informed. For boruta to break down in a similar fashion the model will have to assign wrong importance to all original values (i.e. low importance to useful features and high importance to useless features) while in parallel assigning the *correct* low importance to *all* shadow features. Since parts of this process is governed by stochastic processes we mitigate this breakdown by iterations making boruta a robust wrapper algorithm. As a result *boruta_RF_features* has an accuracy close to the human supervised features, likely because it is able to identify the precipitation-based features as useful. When comparing the effectiveness of models trained with Boruta features to those trained with filter methods, it is clear that it is critical to select algorithms that not only evaluate feature relevance but also establish a statistical benchmark for feature usefulness [41].

The human *supervised_features* were established through visual analysis and the utilization of subject-specific knowledge. A clear disadvantage of this method is its subjectivity, which greatly decreases the repeatability of the process. Furthermore, the process can be quite lengthy, particularly when dealing with a large set of features, yet it remains an integral component of the exploratory data analysis in any ML-project. The results of the models that are trained and tested on the *supervised_features* have the highest overall performance in all evaluated metrics.

Both the boruta- and human-defined feature-sets show convergence below the 0.05 p-value level in 2022. This is eight years earlier than GSAT in 2030 which underpins the benefit to include precipitation-based features in the model training in order to create separable scenario-clusters in the early cross-sections.

4.3.2 The classification process

Gaussian naïve Bayes classifier

Panel (a) of Figure 4.7 shows the mean line accuracy of the GNB classifier following a nearly linear trend between 2020 and 2050 for all sets of features. This contrasts with most other algorithm-feature-set combinations, which have more logarithmic growth with a deceleration when they converge. From Table 4.1, we see that the average classification accuracy of the GNB classifier is the lowest among all algorithms investigated. To discuss the results, we reiterate the central assumptions of the algorithm: The Gaussian naïve Bayes classifier (GNB classifier) is based on Bayes theorem and the *naïve* assumption that all features are conditionally independent given the scenario. Further, we assume continuous features with a Gaussian distribution within the scenario as well as equal variance. By comparing the accuracy attained using the *supervised_features* set for GNB and LR, it is probable that the GNB classifier is under-fitted. That is, the model has high bias and low variance, which implies that the assumptions of the model are violated. Using the naïve and normality assumptions, the GNB classifier provides a fast classification algorithm. However, the simplicity of Bayes theorem will reduce the learning capacity, making the model resistant to over-fitting. This is also observed in Figure 4.7, with a small difference in the accuracy of the training- and test-set.

When inspecting the correlation plot of Figure 4.1, it is evident that the features are not independent within SSP1-2.6. This is also the case for SSP5-8.5, which has few differences in inter-feature correlation compared to SSP1-2.6. Recalling the derivation of equation (3.15), we invoked the naïve assumption to rewrite the conditional probability of observing a realization x_i from f_i given the scenario and the other features to depend only on the scenario. That is, $P\left(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right) = P\left(x_i \mid y\right)$ for $i \in [1, p]$. In equation (3.15), this allows us to write

$$P(x_1, ..., x_p \mid y_j) = \prod_{i=1}^{p} P(x_i \mid y_j) \quad .$$

When features violate this assumption, the probabilities will be incorrectly calculated, effectively weighting the information from the correlated features more in the posterior probability [55]. Since temperature-based features have a higher count in all feature-sets and are also the ones with the highest correlation, it is probable that the high feature correlation reduces the gain from feature-sets seen in the other classifiers. To effectively mitigate this, we would have to implement correlation-based feature weighting, as outlined by Jiang et al. (2019) [55].

Although, the feature correlation most likely accounts for the low gain from different feature-sets, the *Gaussian* naïve Bayes algorithm also assumes that all realizations come from a Gaussian distribution. This assumption is used to estimate the conditional probability $P(x_i \mid y_j)$ using equation (3.17). A violation of the distribution assumption will reduce the models' learning capacity, as the true distribution is no longer modeled. The normality of the data has not been investigated for all features. However, it has been statistically tested using the Shapiro-Wilks test for the *'nomask'*, *'land_mask'* and *'sea_mask'* version of tas, pr, txx, and rx5day when checking the normality assumption for ANOVA and MANOVA. Here, we observed a violation in 4.9% of the features across all years between 2015 and 2100. Accordingly, further investigations are needed in order to conclude the violation, especially since the sample size of each cross-section is too small for the convergence of the central limit theorem to be invoked. If there are extensive violations of the normality assumption, it is natural to use log-transformations such as the Box-Cox transformation to make the features more "gaussian-like".

The only tunable hyperparameter of the GNB classifier is *'var_smoothing'* which mitigates numerical errors that occur if the variance ratio between features is too small. Figure A.4 in Appendix A.4.1 shows the tune distribution for 2020, 2030, and 2040 for each feature combination. Each snapshot holds the density distribution for the 50 best models, one for each random state. We observe a convergence towards *'var_smoothing'* = 1 as we progress throughout the century, indicating that the correlation between features increases throughout the 21st century. This is also observed in the visual analysis of the correlation. The impact on the model involves accommodating realizations that deviate more from the mean within the scenario, thus smoothing the distributions.

4.3.2.1 Logistic regression

Panel (b) of Figure 4.7 shows the mean line accuracy of the LR, a near logarithmic growth with a deceleration when it converges to 100 % accuracy. From Table 4.1, we see that the average classification accuracy of the LR is the highest among all algorithms investigated. To discuss the results, we reiterate the central assumptions of the algorithm: The logistic regression model is based on the logistic function, which maps the linear combination of features to the probability of the binary outcome. The key assumptions of logistic regression include the binary nature of the class, independence of observations, little to no correlation between features, a linear relationship between features, and the logit of the class probability. Figure 4.7 shows that the training- and test-set accuracy are quite similar, however, with a larger difference in the early years. The average model in the evaluation period of 2035–2040 is a high-accuracy, low-bias model.

Since the observations used for training and testing are realizations of ACCESS-ESM1.5 it is safe to assume observation independence due to the stochastic nature of the ESM. Furthermore, as we are using only two scenarios, the binary class nature is also fulfilled.

As discussed, the features of the four feature-sets are correlated, especially within the temperature-based features. Higher amounts of feature correlations will increase the sensitivity of the regression coefficients (β) of the fitted regression used in the class likelihoods of equation (3.21), thus decreasing the interpability of the fitted model. This is, however, mostly problematic if the model is used to identify feature importance through tests of class separation.

Lieberman & Morris (2014) showed that the predictive capabilities of LR were unaffected even with high levels of inter-feature correlation [56]. So, even if the assumption is violated, the LR is robust in terms of predictive abilities, yielding excellent accuracy and AUC.

If the feature-class relation is non-linear, it will again affect the fitting of regression coefficients and yield a faulty estimation of the class likelihoods of the equation (3.21). This introduces bias and yields a low-accuracy model. If the algorithm is allowed, it will, however, try fitting a relation it is incapable of fitting, hence introducing excessive variance and an over-fitted model. It is likely that this is the case for earlier years, as a successful appearance at this point would signify using a non-linear relation, as one can deduce from Figure 3.5.

In the implementation of LR, we tuned two hyperparameters: the l2 penalty parameter *'C'* and the algorithm for the optimization problem *'solver'*. Figures A.5 and A.6 in Appendix A.4.2 hold snapshots of 2020, 2030, and 2040 for each feature-set. Each snapshot holds the density distribution for the 50 *best models*, one for each random state. In Figure A.5, we show the l2 penalty hyperparameter *'C'*, which is the inverse of the regularization strength. We observe a convergence towards larger values of *'C'* as we progress through the 21st century. Since *'C'* is the inverse of regularization strength, a high value indicates that the training data are becoming more reliable throughout the century, while in the beginning they need stronger damping of features, i.e. they contain more noise in earlier cross-sections [48]. The solver is the backbone of the optimization problem solved in the LR scheme. From Figure A.6, we observe a favoring of the *'newton-cg'* solver as we progress throughout the century.

4.3.2.2 Random forest classifier

Panel (c) of Figure 4.7 shows the mean line accuracy of the RF with the trend being close to linear but deaccelerating as the accuracy converges towards 100 %. From Table 4.1 we see that the average classification accuracy of the RF is the third highest among all algorithms investigated. In contrast to GNB and LR the random forest algorithm does not make any assumptions for the data-set. Consequently, the accuracy is governed solely by hyperparameter tuning. The training- and test-set displays large differences in accuracy across all feature-sets indicating that the models are over-fitted with high variance and low bias. To explain this, we investigate the hyperparameter tune distributions. For the RF, we have tuned four hyperparameters, the pruning parameters penalty parameter *'max_depth'* and *'min_samples_leaf*, the *'max_features'* hyperparameter that governs the subset in the bagging, and the number of decision trees in the forest *'n_estimators'*.

One of the key concepts of the RF is its use of decision tree ensembles. When using ensamble methods, one usually prefers a high variation in the estimators. In essence, we want the individual decision trees to be different, assuming that the majority vote is representative of the expectation value. By using bootstrap aggregation (bagging), one can, in theory, bias individual trees to grow differently. To control the subset size used in the bagging algorithm, we tune *'max_features'*. Figure A.8 from Appendix A.4.3 shows that the tune distribution of *'max_features'* mostly prefers the *'sqrt'* number of features to be used as opposed to 'log$_2$'. This is consistent with empirical evidence found in many studies, such as Probst et al. (2019) (Note that Probst et al. use the R implementation, which denotes *'max_features'* by *'mtry'*) [57].

Figures A.7 and A.8 from Appendix A.4.3 show the tuning distribution of the *'max_depth'* and *'min_samples_leaf'*, respectively. These are pruning techniques that can be tuned if necessary [57]. In initial tuning using random search over a larger hyperparameter space, the absence of these parameters yielded a greatly over-fitted model as all trees were grown to the maximum depth of one sample per node, thus violating the main principle of ensamble learning. Since the sample size is small, we need to keep these heuristic values low and only tune over a max depth of $\{2, 3, 5\}$ and minimum samples per leaf of $\{3, 6, 8, 10, 12, 20\}$. The tune distributions in the Figures show a convergence in the later snapshot toward a maximum tree depth of 2 and minimum samples per leaf of 4. In the earlier years, the tune distributions were more uniformly distributed, suggesting that a low tree depth will be favorable to mitigating over-fitting.

Furthermore, the last hyperparameter needed is the number of trees in the forest tuned by *'n_estimators'*. In general, the number of trees is not a hyperparameter that is subject to "standard" tuning, as empirical and theoretical studies suggest that a larger number of trees are favorable [57]. However, the number of trees drastically increases training time and is therefore kept at a relatively small number in this analysis. Figure A.10 from Appendix A.4.3 shows the tuning distribution of *'n_estimators*. Here we observe a small favoring of fewer trees towards the later snap-shots; however, the distribution does not converge to any specific value.

In conclusion, to mitigate the over-fitting of the random forest, further search for the optimal hyperparameters is needed. We make this conclusion based on the fact that less pruning resulted in a more over-fitted model because the individual trees grew to be similar. In addition to further investigation on the four parameters that are tuned here, we also suggest testing another split criterion than the gini of equation (3.24). For details, we refer to Probst et al. (2019) [57].

4.3.2.3 Support Vector Classifier

Panel (d) of Figure 4.7 shows the mean line accuracy of the SVC with the trend having a nearly logarithmic growth with a deceleration when the accuracy converges towards 100 %. From Table 4.1, we see that the average classification accuracy of the SVC is the second highest among all algorithms investigated. As with the random forest, the support vector classifier does not make any assumptions about the underlying distribution of the data-set. It does, however, assume that the data are linearly separable in the feature space or that they can be transformed into a space where they are linearly separable using the kernel trick. Consequently, accuracy is governed solely by hyperparameter tuning. The training- and test-set display large differences in accuracy across all feature-sets until the test-set accuracy exceeds ~ 0.8. After this, the bias-variance trade-off is better balanced, leading to a model with a low degree of over-fitting. To explain this, we investigate the hyperparameter tuning distributions. For the SVC, we have tuned four hyperparameters: the penalty term *'C'*, which governs the regularization strength, as well as the kernel-specific hyperparameters *'kernel'*, *'gamma'*, and *'degree'*.

To reiterate: *'C'* regularizes equation (3.28), thus a small *'C'* will allow a greater proportion of realizations to be inside the margins, while a large *'C'* will penalize classification errors and thus narrow the margin. Figure A.11 found in Appendix A.4.4 shows the tune distribution of *'C'* for the SVC. In 2020, *'C'* is *heavy-tailed*, that is, the *best* model mostly has a strong or very weak regularization. This is commonly observed when the data-set is noisy and strongly affected by the random statistics, in line with the results in the ROC curves. In 2030 and 2040, the SVC has good classification accuracy; this is attained by weak regularization at *'C'*= 0.1 or no regularization with *'C'*=1. However, convergence towards the closest values to *'C'*= 1 is delayed for *mRMR_f_mut_features*, which also has a weaker accuracy and AUC metric compared to the remaining feature-sets. This indicates that the distribution becomes more linearly separable in the original feature space, which is expected if one compares the distribution of Figure 3.5 with the linear kernel of Figure 3.11.

Figure A.12 found in Appendix A.4.4 presents a view of the tuning distribution for the *'kernel'* hyperparameter across the years 2020, 2030, and 2040. Initially, in 2020, the distribution was nearly uniform, gradually shifting to favor the *'linear'* kernel by 2040. In the distributions for 2030 and 2040, none of the random states identify the *'poly'* kernel as the *best* model. Figure A.13 found in Appendix A.4.4 shows the tune distribution of the *'degree'* which is specific to the *'poly'* kernel. Consequently, there are no realizations in the 2030 and 2040 snapshots. The low rate of the *'poly'* kernel is likely a result of the specified feature space for the *'degree'* parameter. As mentioned earlier, the degree parallels the polynomial degree in a regression line. Given the choices of 2, 6, and 10 for the degree, the poly kernel utilizes 6 and 10, since 2 is associated with the *'linear'* kernel. These are not the most commonly utilized degrees and will likely lead to over-fitting, which is uncovered by the validation-set in the cross-validation. In order to correct this error in the study, it will be beneficial to rerun the experiment with a new hyperparameter space, e.g. *'degree'*$\in \{3, 4, 5, 6, 8, 10\}$.

The *'gamma'* hyperparameter is another regularization hyperparameter that is specific to the *'poly'*, *'rbf'*, and *'sigmoid'* kernels. Since the *'linear'* kernel is one of the most frequent kernels, the plots in Figure A.14 have fewer than fifty realizations in total. For random states where the kernels *'poly'*, *'rbf'*, or *'sigmoid'* are used, the *'scale'* version of *'gamma'* is the most frequent, as opposed to *'auto'*. However, since we train on scaled cross-sections, the variance of the design matrix used in *'scale'* is close to 1 across all years, so there will not be a large difference in which scheme one uses. Consequently, we can assume $\text{gamma}_{\text{scale}} \approx \text{gamma}_{\text{auto}} = 1/\,\text{n\_features}$ and remove the need for tuning this hyperparameter, freeing computational capacity for e.g. more polynomial degrees for the *'poly'* kernel.

Chapter 5

Conclusion and Outlook

5.1 Conclusion

Climate change is a global phenomenon that imposes great risks for both humans and ecosystems. This has prompted more research on the climate of the future, particularly by using ESM experiments. This thesis has focused on using ML methods and algorithms to further the ability of science to separate model realization under different shared socioeconomic pathways. With the ultimate goal of providing a proof of concept for training ML classification models on ScenarioMIP output data creating a feedback tool for the effect of mitigation policies. The analysis is based on the output data from ACCESS-ESM1.5, with 40 model realizations of 2015 to 2100 per scenario.

Although the task of SSP classification through ESM realizations was not previously explored, the evolution of climate change indicators within the ScenarioMIP projections is thoroughly assessed. In the sixth phase of CMIP, Tebaldi et al. (2021) proposed a method for distinguishing SSPs by maintaining a consistent division across the multi-model average of GSAT. According to this criterion, SSP1-2.6 and SSP5-8.5 will be clearly separated by 2030. This is in accordance with the statistical separation of scenario means observed in our study through the convergence of the ANOVA p-value. Similarly, the multivariate MANOVA p-value showed a faster statistical separation for all examined feature-sets. From this, it can be inferred that incorporating a variety of features, rather than solely relying on GSAT, could be advantageous. In addition, we also conclude that precipitation-based features help span a feature space where realizations are easier to classify. Following from this, we can evaluate the feature selection process where four feature-sets were suggested. The *nomask_baseline* feature-set represents the global mean of all response variables. The results suggest that there are significant gains in model accuracy to be made by the masking process, indicated by the comparison of the *nomask_baseline* as opposed to e.g. *boruta_RF_features*. In general, the *supervised_features* set enabled the best models for all algorithms; however, the subjectivity of this process makes it a less consistent pipeline for further work. We recommend the Boruta-determined pipeline as the best method for feature selection in this problem.

For the last years of the near-term period (2035-2040), we attain a maximum classification accuracy of 0.97 ± 0.04 (LR and *supervised_features*) and a minimum accuracy of 0.80 ± 0.09 (GNB classifier and *nomask_baseline* / SVC and *mRMR_f_mut_features*). This suggests that to achieve the benefits of feature engineering, proper model development is essential. This process will look different for all algorithms because they are derived from different principles of separation and thus have different strengths and weaknesses. For the lower-performing algorithms, we propose an additional adaptation by: The GNB classifier requires a feature-set that aligns with the algorithm's distribution assumptions, i.e. reduces feature correlation. In the case of the RF algorithm, extensive hyperparameter tuning is required to align with the data-set size. Based on the analysis of the ROC-AUC and the accuracy of 2030–2040, we conclude that the best algorithms will be either the LR or SVC, which both have strong predictive capabilities. However, for the full near-term period, the SVC might be the better alternative, as it is more stable than LR in the early years with a generally better AUC.

Throughout this thesis, we have shown that the separation of SSP1-2.6 and SSP5-8.5 (where the emission starts to differ in 2015) is attainable with a classification accuracy above 0.8 as early as 2026 based on the mean accuracy across 50 random states. Following the uncertainty in the ROC curve analysis, the dependence towards the random state is, however, greatly reduced towards the later part of the near-term period. For applications on real-world data, this observation is crucial, as it signifies a low impact of the realizations position within the scenario cluster.

We successfully developed a pipeline for training ML models on ESM realizations and estimated the models' generalization abilities using the corresponding test-set. Together with the promising results, this will serve as a stepping stone toward a real-world application that allows effective and confident monitoring of climate system responses to mitigation efforts. This is a critical ability for the science community to provide for policymakers, as the previously established separation time of 20 to 30 years eliminates all room for error in a world aiming to meet the goals of the Paris agreement.

5.2 Outlook

Throughout the work in this thesis, we have identified several directions that can be explored in future work. Some, which will improve the classification models predictive abilities and others, which will further the understanding of earth system modeling as a whole. The suggestions are presented in no particular order with respect to priority and/or difficulty. We do, however, recommend testing one approach at the time in order to identify the response to the individual extensions as opposed to the compound effect.

1) Extension of the analysis

There are several analyses that are natural steps in furthering the analysis and framework given in this thesis. For all extensions there are no guarantees that the optimal feature-set/model combination from the binary case study will be the optimal for the new configuration, as this is specific to the given cross-section/data-set. Further classification algorithms should be considered as all these extensions alter the models' demand for learning capacity. Some suggestions include artificial neural networks (ANN) and XGBoost.

1a) Extending the featurespace: The featurespace in this thesis are presented in Table 3.3. Here all features are either a) spatial subsets of tas or pr, or b) extremes generate from tas or pr. As a result we get highly correlated features which will represent the same dynamics in the ES. From the results we see that even though precipitation-based features have a low degree of SSP-separation they drastically improve the predictive ability of a well developed model. This shows that introduction of more variance in *"new"* directions, i.e. new dynamics, is beneficial. To ensure this does not introduce additional noise, an appropriate feature selection process must be implemented. Recommendations for global climate change indicators are the global sea level and the Arctic sea ice area. Although their changes are influenced by the dynamics of the Earth system and feedback mechanisms we know, from the inclusion of precipitation data, that more indicators could be advantageous, despite the lack of significant anomalies to verify the early detection of ACC signals. Some caution should however be deployed: the end-goal application is to inform policies based on in-situ observational data. Thus, there are a clear requirement for appropriate measurement coverage with quantified uncertainty for all indicators that are included in model training. If not, we are not able to assess the certainty of the scenario estimate.

1b) Re-run using more ESMs: In this thesis we have used realizations of ACCESS-ESM1.5 for the analysis. In order to verify the method it is crucial to re do the analysis with other ESMs. As the classification algorithm responds differently to every data-set consequently using new realizations from other models will either confirm or disprove the readiness toward applications on real-world data. To introduce this extension it is critical to perform a model validation of the individual ESM. Hereunder verifying appropriate magnitude of internal variability as well as a realistic representation of the indicators of climate change compared to the real-world numbers. Due to structural differences in the ESMs, training on one large aggregate data-set of all realizations from all models will a) likely lead to very slow emergence of ACC, and b) reduce interprebility of the data-set. Therefore, if a multi-model approach is tested for real-world applications we recommend a majority vote ensembling where we treat the *best* model from each ESM as a voter. It is also possible to use the model validation to inform a weighting scheme of the vote, thus biasing the vote towards the most representative ESMs.

1c) Re-run using all SSPs: Here we have only covered two, very distinctly different, policy scenarios which has the largest deviations in ERF at the end of the century. In order to create a viably policy feedback-tool additional scenarios should be added, starting with the other *tier 1* scenarios SSP2-4.5 and SSP3-7.0. The best performing algorithms in this analysis are originally binary-classifiers which specialize in problems such as we have investigated here. In order to apply them to multi class purposes they can however be adapted in a one-versus-rest (ovr) scheme supported by libraries such as scikit-learn. Other algorithms are inherently adapted to multi class applications and will therefore not need altering. Additionally a pairwise investigation equivalent to the results of Table 2.3 should be done.

2) Improving Signal to noise ratio of the training data

Naturally, the learning algorithm can not *learn* things that are not present in the data (at least we do not want it to). In many ways, data quality in itself can be viewed as the main component that regulates the predictive capabilities. By improving the signal to noise ratio (SNR) of the individual features, it is therefore likely a better model can be obtained in earlier cross-sections.

2a) Use SNR to identify better masks: The masking process allowed for a large gain in accuracy as opposed to using the global-level indicators as done in *nomask_baseline*. In this thesis, we used subjective metrics such as visual analysis as well as region-based analysis performed by the IPCC to identify the masks for each response variable. In order to improve this, we suggest a data-driven protocol based on data SNR to inform on regions with low noise that will aid in a similar way to precipitation. The appropriate pipeline will include:

a) A calculation of a reference period climatology using realizations of historic simulations for the ESM.

b) A calculation of anomalies to fit a linear regression line. This will allow for an easy estimation of the mean and standard deviation which allows for calculations of the SNR

c) By thresholding the grid of SNR values of each grid-box we effectively create a mask that helps us identify regions with high SNR.

d) By using this SNR mask we can more easily identify regions of high inter-SSP variance using a convolutional neural network (CNN).

2b) Filtering of internal variability: Internal variability introduces large amounts of noise in the quest to identify the effects of mitigation efforts. In 2022 Samset et al. proposed a physics based Green's function to reduce noise in global surface temperature anomaly (GSTA) created from sea surface temperature variations [58]. The main findings include a reduce emergence time from 2035 to 2030 by using the Green's functions filtered GSTA. By using similar filtering techniques on features in the model training it is probable that a faster successfully classification can be made, however at the cost of decreased model-interprebility following the transformation.

3) Other

3a) Temporally dependent methods [1]**:** In this thesis we deploy a temporally independent cross-sectional approach to the classification. In that, we also remove the capability to use the history of a realization to aid the classification. In order to choose if the temporal relations should be taken into account when classifying one key property to investigate is the autocorrelation function (acf). Now, let $\text{acf}(x_t)$ denote the correlation between the time-series and lagged versions (shifting forward or backwards). A significant acf for a lag h, indicates that a current time point x_t is dependent on previous time steps x_{t-h}. The acf is shown to be significant for multiple SSPs for both tas and pr which suggest that there is information to be gained from using the time-dependent nature of our relizations. Initial analysis of the feasibility of common clustering approaches using time-series data showed low levels of applicability to *annual* climatologies as used here. Therefor, we suggest an investigation of seasonal climatologies with the periods DJF, MAM, JJA, SON which will open for a wast array of classification methods. Some example approaches include

a) Distance-based clustering around a baseline time-series per scenario with a clustering that is sensitive to amplitude shifts in order to detect similarity between scenarios and unseen samples.

b) Distance-based clustering using the acf distance metric as the acf was shown in initial analysis to have large inter-scenario differences.

c) Model-based approach should be investigated again, using SeasonalARIMA model with multiplicative residuals which allows for amplitude changes of seasonal patterns.

For more information, we refer to the *Previous work* directory available in the code [30]

3b) Bayesian estimation of uncertainty of the ESM: For a successful feedback tool, it is crucial to understand the certainty of our estimates. From the machine learning models standpoint we estimate the generalization abilities based on the test-set accuracy. For a given cross-section we therefore have a ballpark number of how often we expect a correct classification of the ESM data. For applications to real-world observational data we must however also assess the ESMs' uncertainty. Today, this is in part done by assessing the historical simulations against observational data and for forecasting purposes done through perturbations of the initial values of the simulation as well as using multi-model ensambles to account for structural differences in the numerical model. As previously discussed the multi-model approach adds extra uncertainty to the interpretation of ML model output, and so a one-ESM approach might be preferable. Either way, in order to better quantify the uncertainty from the ESM itself we suggest a Bayesian estimation of uncertainty using a multi-model ensemble of simulations over the historical era to calculate the prior distribution and class probability to define the likelihood function.

Refrences

[1] Intergovernmental Panel On Climate Change. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press, July 6, 2023. ISBN: 978-1-00-915789-6. DOI: 10.1017/9781009157896. URL: https://www.cambridge.org/core/product/identifier/9781009157896/type/book (visited on 04/13/2024).

[2] John M. Wallace and Peter V. Hobbs. *Atmospheric Science*. Elsevier, 2006. ISBN: 978-0-12-732951-2. DOI: 10.1016/C2009-0-00034-8. URL: https://linkinghub.elsevier.com/retrieve/pii/C20090000348 (visited on 04/13/2024).

[3] Intergovernmental Panel On Climate Change (Ipcc). *Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press, June 22, 2023. ISBN: 978-1-00-932584-4. DOI: 10.1017/9781009325844. URL: https://www.cambridge.org/core/product/identifier/9781009325844/type/book (visited on 04/13/2024).

[4] Will Steffen et al. "The emergence and evolution of Earth System Science". In: *Nature Reviews Earth & Environment* 1.1 (Jan. 13, 2020), pp. 54–63. ISSN: 2662-138X. DOI: 10.1038/s43017-019-0005-6. URL: https://www.nature.com/articles/s43017-019-0005-6 (visited on 03/25/2024).

[5] *The Paris Agreement | UNFCCC*. URL: https://unfccc.int/process-and-meetings/the-paris-agreement (visited on 04/14/2024).

[6] Keywan Riahi et al. "The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview". In: *Global Environmental Change* 42 (Jan. 2017), pp. 153–168. ISSN: 09593780. DOI: 10.1016/j.gloenvcha.2016.05.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S0959378016300681 (visited on 03/26/2024).

[7] "Summary for Policymakers". In: *Climate Change 2022 - Mitigation of Climate Change*. Ed. by Intergovernmental Panel On Climate Change (Ipcc). 1st ed. Cambridge University Press, Aug. 17, 2023, pp. 3–48. ISBN: 978-1-00-915792-6. DOI: 10.1017/9781009157926.001. URL: https://www.cambridge.org/core/product/identifier/9781009157926%23pre2/type/book_part (visited on 04/14/2024).

[8] William J Ripple et al. "The 2023 state of the climate report: Entering uncharted territory". In: *BioScience* 73.12 (Dec. 29, 2023), pp. 841–850. ISSN: 0006-3568. DOI: 10.1093/biosci/biad080. URL: https://doi.org/10.1093/biosci/biad080 (visited on 04/14/2024).

[9] Claudia Tebaldi et al. "Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6". In: *Earth System Dynamics* 12.1 (Mar. 1, 2021), pp. 253–293. ISSN: 2190-4987. DOI: 10.5194/esd-12-253-2021. URL: https://esd.copernicus.org/articles/12/253/2021/ (visited on 03/23/2024).

[10] NASA Scientific Visualization Studio. *NASA Scientific Visualization Studio | Earth System Diagram*. NASA Scientific Visualization Studio. Aug. 29, 2018. URL: https://svs.gsfc.nasa.gov/30988 (visited on 03/25/2024).

[11] Lewis Fry Richardson. *Weather prediction by numerical process*. 2nd ed. Cambridge mathematical library. Cambridge university press, 2006. ISBN: 978-0-521-68044-8.

[12] William D. Sellers. "A Global Climatic Model Based on the Energy Balance of the Earth-Atmosphere System". In: *Journal of Applied Meteorology and Climatology* 8.3 (June 1, 1969). Publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, pp. 392–400. ISSN: 1520-0450. DOI: 10.1175/1520-0450(1969)008<0392:AGCMBO>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/apme/8/3/1520-0450_1969_008_0392_agcmbo_2_0_co_2.xml (visited on 03/25/2024).

[13] Gregory M. Flato. "Earth system models: an overview". In: *WIREs Climate Change* 2.6 (Nov. 2011), pp. 783–800. ISSN: 1757-7780, 1757-7799. DOI: 10.1002/wcc.148. URL: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcc.148 (visited on 03/25/2024).

[14] Tapio Schneider et al. "Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations". In: *Geophysical Research Letters* 44.24 (Dec. 28, 2017). ISSN: 0094-8276, 1944-8007. DOI: 10.1002/2017GL076101. URL: https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017GL076101 (visited on 03/25/2024).

[15] Intergovernmental Panel On Climate Change. *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. 1st ed. Cambridge University Press, July 6, 2023. ISBN: 978-1-00-915789-6. DOI: 10.1017/9781009157896. URL: https://www.cambridge.org/core/product/identifier/9781009157896/type/book (visited on 03/25/2024).

[16] Veronika Eyring et al. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5 (May 26, 2016). Publisher: Copernicus GmbH, pp. 1937–1958. ISSN: 1991-959X. DOI: 10.5194/gmd-9-1937-2016. URL: https://gmd.copernicus.org/articles/9/1937/2016/gmd-9-1937-2016.html (visited on 03/26/2024).

[17] Brian C. O'Neill et al. "The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6". In: *Geoscientific Model Development* 9.9 (Sept. 28, 2016), pp. 3461–3482. ISSN: 1991-9603. DOI: 10.5194/gmd-9-3461-2016. URL: https://gmd.copernicus.org/articles/9/3461/2016/ (visited on 03/22/2024).

[18] Detlef P. van Vuuren et al. "A new scenario framework for Climate Change Research: scenario matrix architecture". In: *Climatic Change* 122.3 (Feb. 1, 2014), pp. 373–386. ISSN: 1573-1480. DOI: 10.1007/s10584-013-0906-1. URL: https://doi.org/10.1007/s10584-013-0906-1 (visited on 03/26/2024).

[19] Karl E. Taylor, Ronald J. Stouffer, and Gerald A. Meehl. "An Overview of CMIP5 and the Experiment Design". In: *Bulletin of the American Meteorological Society* 93.4 (Apr. 1, 2012), pp. 485–498. ISSN: 1520-0477. DOI: 10.1175/BAMS-D-11-00094.1. URL: https://journals.ametsoc.org/doi/10.1175/BAMS-D-11-00094.1 (visited on 04/17/2024).

[20] Detlef P. van Vuuren et al. "The representative concentration pathways: an overview". In: *Climatic Change* 109.1 (Aug. 5, 2011), p. 5. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0148-z. URL: https://doi.org/10.1007/s10584-011-0148-z (visited on 03/26/2024).

[21] Keywan Riahi et al. "RCP 8.5—A scenario of comparatively high greenhouse gas emissions". In: *Climatic Change* 109.1 (Aug. 13, 2011), p. 33. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0149-y. URL: https://doi.org/10.1007/s10584-011-0149-y (visited on 04/19/2024).

[22] Toshihiko Masui et al. "An emission pathway for stabilization at 6 Wm2 radiative forcing". In: *Climatic Change* 109.1 (Aug. 13, 2011), p. 59. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0150-5. URL: https://doi.org/10.1007/s10584-011-0150-5 (visited on 04/19/2024).

[23] Allison M. Thomson et al. "RCP4.5: a pathway for stabilization of radiative forcing by 2100". In: *Climatic Change* 109.1 (Nov. 2011), pp. 77–94. ISSN: 0165-0009, 1573-1480. DOI: 10.1007/s10584-011-0151-4. URL: http://link.springer.com/10.1007/s10584-011-0151-4 (visited on 04/19/2024).

[24] Detlef P. Van Vuuren et al. "RCP2.6: exploring the possibility to keep global mean temperature increase below 2°C". In: *Climatic Change* 109.1 (Nov. 2011), pp. 95–116. ISSN: 0165-0009, 1573-1480. DOI: 10.1007/s10584-011-0152-3. URL: http://link.springer.com/10.1007/s10584-011-0152-3 (visited on 04/19/2024).

[25] Brian C. O'Neill et al. "The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century". In: *Global Environmental Change* 42 (Jan. 1, 2017), pp. 169–180. ISSN: 0959-3780. DOI: 10.1016/j.gloenvcha.2015.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0959378015000060 (visited on 04/11/2024).

[26] David Kelly. "Integrated Assessment Models for Climate Change Control". In: *International Yearbook of Environmental and Resource Economics 1999/2000: A Survey of Current Issues* (Jan. 1, 1999).

[27] Monica Ainhorn Morrison and Peter Lawrence. "Understanding Model-Based Uncertainty in Climate Science". In: *Handbook of Philosophy of Climate Change*. Ed. by Gianfranco Pellegrino and Marcello Di Paola. Cham: Springer International Publishing, 2020, pp. 1–21. ISBN: 978-3-030-16960-2. DOI: 10.1007/978-3-030-16960-2_154-1. URL: https://doi.org/10.1007/978-3-030-16960-2_154-1 (visited on 04/21/2024).

[28] *ACCESS-ESM1.5 - Australian Community Climate and Earth System Simulator (ACCESS)*. [Online; accessed 17. Dec. 2023]. Oct. 2021. URL: https://research.csiro.au/access/about/esm1-5.

[29] Linnea L. Huusko et al. "Climate sensitivity indices and their relation with projected temperature change in CMIP6 models". In: *Environmental Research Letters* 16.6 (June 2021). Publisher: IOP Publishing, p. 064095. ISSN: 1748-9326. DOI: 10.1088/1748-9326/ac0748. URL: https://dx.doi.org/10.1088/1748-9326/ac0748 (visited on 04/21/2024).

[30] Johannes Fjeldså. "FYS-STK4155 - Project 3 Is shared socioeconomic pathways separable? A classification problem". In: (Dec. 21, 2023).

[31] Tilo Ziehn et al. "The Australian Earth System Model: ACCESS-ESM1.5". In: *Journal of Southern Hemisphere Earth Systems Science* 70.1 (Aug. 24, 2020). Publisher: CSIRO PUBLISHING, pp. 193–214. ISSN: 2206-5865. DOI: 10.1071/ES19035. URL: https://www.publish.csiro.au/es/ES19035 (visited on 04/11/2024).

[32] Carley Iles et al. *Climate extreme indices and heat stress indicators derived from CMIP6 global climate projections: Product User Guide*. URL: https://confluence.ecmwf.int/display/CKB/Climate+extreme+indices+and+heat+stress+indicators+derived+from+CMIP6+global+climate+projections%3A+Product+User+Guide.

[33] *CMIP6 Data Request MIP Variables search*. CMIP6 Data Request. Mar. 27, 2024. URL: https://clipc-services.ceda.ac.uk/dreq/mipVars.html (visited on 03/27/2024).

[34] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[35] *What is Overfitting in Deep Learning [+10 Ways to Avoid It]*. [Online; accessed 17. Dec. 2023]. Dec. 2023. URL: https://www.v7labs.com/blog/overfitting.

[36] *5. Resampling Methods — Applied Data Analysis and Machine Learning.* [Online; accessed 17. Dec. 2023]. Dec. 2023. URL: https://compphysics.github.io/MachineLearning/doc/LectureNotes/_build/html/chapter3.html?highlight=bias.

[37] Jundong Li et al. "Feature Selection: A Data Perspective". In: *ACM Computing Surveys* 50.6 (Nov. 30, 2018), pp. 1–45. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3136625. arXiv: 1601.07996[cs]. URL: http://arxiv.org/abs/1601.07996 (visited on 03/23/2024).

[38] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to mathematical statistics.* Eighth edition. Boston: Pearson, 2019. 746 pp. ISBN: 978-0-13-468699-8.

[39] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Physical Review E* 69.6 (June 23, 2004), p. 066138. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.69.066138. URL: https://link.aps.org/doi/10.1103/PhysRevE.69.066138 (visited on 04/26/2024).

[40] Chris Ding and Hanchuan Peng. "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". In: ().

[41] Miron B. Kursa and Witold R. Rudnicki. "Feature Selection with the **Boruta** Package". In: *Journal of Statistical Software* 36.11 (2010). ISSN: 1548-7660. DOI: 10.18637/jss.v036.i11. URL: http://www.jstatsoft.org/v36/i11/ (visited on 03/23/2024).

[42] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *arXiv* (May 2017). DOI: 10.48550/arXiv.1705.07874. eprint: 1705.07874.

[43] *Lesson 8: Multivariate Analysis of Variance (MANOVA).* URL: https://online.stat.psu.edu/stat505/book/export/html/762 (visited on 04/27/2024).

[44] Christo El Morr et al. "Naïve Bayes". In: *Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics.* Ed. by Christo El Morr et al. Cham: Springer International Publishing, 2022, pp. 279–299. ISBN: 978-3-031-16990-8. DOI: 10.1007/978-3-031-16990-8_9. URL: https://doi.org/10.1007/978-3-031-16990-8_9 (visited on 04/22/2024).

[45] Harry Zhang. "The Optimality of Naive Bayes". In: ().

[46] Kashishdafe. *Gaussian Naive Bayes: Understanding the Basics and Applications.* Medium. Mar. 23, 2024. URL: https://medium.com/@kashishdafe0410/gaussian-naive-bayes-understanding-the-basics-and-applications-52098087b963 (visited on 04/22/2024).

[47] *scikit-learn/sklearn/naive_bayes.py at main · scikit-learn/scikit-learn.* GitHub. URL: https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/naive_bayes.py (visited on 04/25/2024).

[48] *sklearn.linear_model.LogisticRegression.* scikit-learn. URL: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (visited on 04/25/2024).

[49] *9. Decision trees, overarching aims — Applied Data Analysis and Machine Learning.* [Online; accessed 17. Dec. 2023]. Dec. 2023. URL: https://compphysics.github.io/MachineLearning/doc/LectureNotes/_build/html/chapter6.html.

[50] Mengyun Zhang et al. "Fully convolutional networks for blueberry bruising and calyx segmentation using hyperspectral transmittance imaging". In: *Biosyst. Eng.* 192 (Feb. 2020), p. 159. ISSN: 1537-5110. DOI: 10.1016/j.biosystemseng.2020.01.018.

[51] Jason Brownlee. "Bagging and Random Forest Ensemble Algorithms for Machine Learning - MachineLearningMastery.com". In: *MachineLearningMastery* (Dec. 2020). URL: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning.

[52] PhD Zijing Zhu. "Explain Support Vector Machines in Mathematic Details". In: *Medium* (Dec. 2021). URL: https://towardsdatascience.com/explain-support-vector-machines-in-mathematic-details-c7cc1be9f3b9.

[53] Anuganti Suresh. "What is a confusion matrix? - Analytics Vidhya - Medium". In: *Medium* (Dec. 2021). URL: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5.

[54] Ruchi Toshniwal. *Demystifying ROC Curves.* Medium. Jan. 21, 2020. URL: https://towardsdatascience.com/demystifying-roc-curves-df809474529a (visited on 04/05/2024).

[55] Liangxiao Jiang et al. "A Correlation-Based Feature Weighting Filter for Naive Bayes". In: *IEEE Transactions on Knowledge and Data Engineering* 31 (Aug. 10, 2019), pp. 201–213. DOI: 10.1109/TKDE.2018.2836440.

[56] Mary Lieberman and John Morris. "The Precise Effect of Multicollinearity on Classification Prediction". In: 40 (Jan. 1, 2014), pp. 5–10.

[57] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix. "Hyperparameters and Tuning Strategies for Random Forest". In: *WIREs Data Mining and Knowledge Discovery* 9.3 (May 2019), e1301. ISSN: 1942-4787, 1942-4795. DOI: 10.1002/widm.1301. arXiv: 1804.03515[cs,stat]. URL: http://arxiv.org/abs/1804.03515 (visited on 05/02/2024).

[58] B. H. Samset et al. "Earlier emergence of a temperature response to mitigation by filtering annual variability". In: *Nature Communications* 13.1 (Mar. 24, 2022). Publisher: Nature Publishing Group, p. 1578. ISSN: 2041-1723.

DOI: 10.1038/s41467-022-29247-y. URL: https://www.nature.com/articles/s41467-022-29247-y (visited on 05/09/2024).

Appendix A

Appendices

A.1 Acronyms

ACC	anthropogenic climate change	1
ACCESS-ESM1.5	Australian Community Climate and Earth System Simulator earth system model 1.5	12
ANOVA	analysis of variance	19
Arctic SIA	Arctic sea ice area	5
FAR	First IPCC Assessment Report (published 1990)	1
AR6	Assessment Report 6 (published 2023)	1
AUC	Area Under Curve	29
CMIP	Coupled Model Intercomparison Project	1
CMIP6	Coupled Model Intercomparison Project phase 6	6
DECK	Diagnostic, Evaluation and Characterization of Klima	6
ECS	equilibrium climate sensitivity	10
ERF	effective radiative forcing	4
ES	earth system	1
ESS	Earth system science	2
ESM	Earth system model	2
ETCCDI	Expert Team on Climate Change Detection and Indices	12
fd	frost days	12
FN	false negative	28
FP	false positive	28
FPR	false positive rate	29

GCM	general circulation model	6
GHG	greenhouse gas	1
GMSL	global mean sea level	6
GMST	global mean surface temperature	1
GNB classifier	Gaussian naïve Bayes classifier	22
GSAT	global surface air temperature	5
gsl	growing season length	12
IAM	integrated assessment model	8
IPCC	Intergovernmental Panel on Climate Change	4
long-term period	time period from 2081 to 2100	9
LR	logistic regression	24
MANOVA	multivariate analysis of variance	21
MIM	mutual information maximization	20
MIP	Model Intercomparison Project	6
ML	machine learning	2
mRMR	minimum redundancy maximum relevance	20
near-term period	time period from 2021 to 2040	9
NOAA	National Oceanic and Atmospheric Administration	5
pr	precipitation	1
RCP	representative concentration pathway	6
RF	random forest	21
ROC curve	receiver operating characteristic curve	29
rx5day	maximum 5-day precipitation	12

ScenarioMIP Scenario Model Intercomparison Project . . . 6

SHAP SHapley Additive exPlanations . . . 21

SNR signal to noise ratio . . . 2

SSP shared socioeconomic pathway . . . 1

SVC support vector classifier . . . 27

tas near-surface air temperature . . . 1

TCR transient climate sensitivity . . . 10

TCWV total column of water vapor . . . 5

TN true negative . . . 28

TP true positive . . . 28

TPR true positive rate . . . 29

TSI total solar irradiation . . . 4

txx maximum value of daily maximum temperature . . . 12

WMGHG well-mixed greenhouse gas . . . 4

A.2 Figures and tables

A.2.1 The radiative forcing-SSP matrix

The SSP-RCP scenario matrix. The Tier 1 scenarios represents the main scenarios used to investigate the future climate. The Tier 2 scenarios is implemented to fill plausible gaps in socioeconomic-radiative forcin pairing. Through AR6 SSP1-1.9 is commonly used included in analysis as it presents a scenario where the 1.5 °C target of the Paris agreement is met.

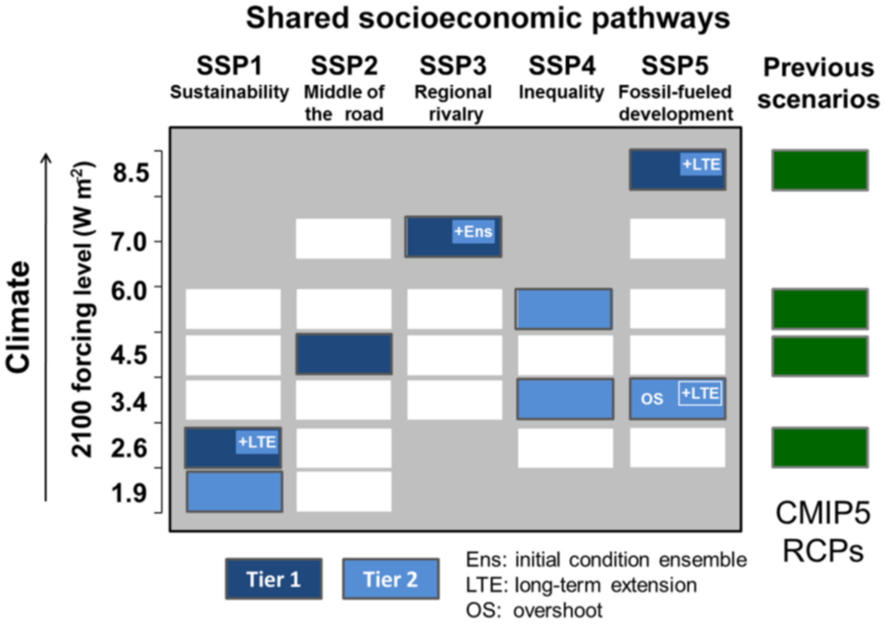


Figure A.1: SSP-RCP scenario matrix illustrating ScenarioMIP simulations. Each cell in the matrix indicates a combination of socioeconomic development pathway (i.e., an SSP) and climate outcome based on a particular forcing pathway that current IAM runs have shown to be feasible [6]. Dark blue cells indicate scenarios that will serve as the basis for climate model projections in Tier 1 of ScenarioMIP; light blue cells indicate scenarios in Tier 2. An overshoot version of the 3.4 Wm^{-2} pathway is also part of Tier 2, as are long-term extensions of SSP5-8.5, SSP1-2.6 and the overshoot scenario, and initial condition ensemble members of SSP3-7.0. White cells indicate scenarios for which climate information is intended to come from the SSP scenario to be simulated for that row. CMIP5 RCPs, which were developed from previous socioeconomic scenarios rather than SSPs, are shown for comparison. Caption and figure from O'Neill et al. (2016) [17].

A.2.2 Table of tested masks

Table A.1 holds the masks tested. Detailed information on mask development and spatial constraints is available in the code (see Appendix A.6).

Table A.1: Overview of tested mask-variable combinations.

Masks	**Variables**					
Short name	**tas**	**pr**	**txx**	**rx5day**	**gsl**	**fd**
nomask	x	x	x	x	x	x
land_mask	x	x	x	x	x	x
sea_mask	x	x	x	x		
lat_mask_pm15deg	x	x				
land_mask_pm15deg		x				
sea_mask_pm15deg	x	x				
lat_mask_pm30deg	x	x	x	x		
land_mask_pm30deg		x	x	x		
sea_mask_pm30deg	x	x	x	x		
lat_mask_0_30N		x				
lat_mask_30S_0		x				
NH_arctic_mask	x					
lat_mask_30N_70N					x	x
land_mask_30N_70N					x	x
pr_decrease_mask		x		x		
pr_increase_mask		x		x		
pr_large_deviation_mask		x		x		

A.2.3 Feature ranking from filter methods

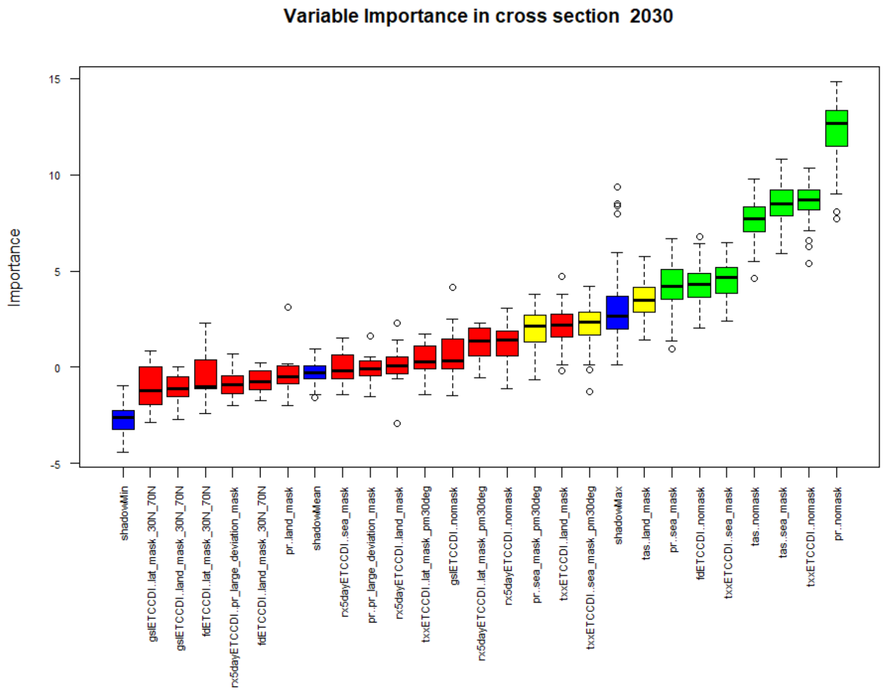
Table A.2 shows the top 10 ranked features of the three filter methods. All methods have the same 10 features as top ranked, and no precipitation-based features are selected.

Table A.2: Overview of top 10 features from filter feature selection methods.

Rank	f_classif	Mut_info	mRMR
1	tas: nomask	tas: nomask	tas: nomask
2	tas: sea_mask	tas: sea_mask	tas: sea_mask
3	txx: sea_mask	tas: land_mask	txx: sea_mask
4	tas: land_mask	txx: nomask	tas: land_mask
5	txx: nomask	txx: sea_mask	txx: nomask
6	txx: sea_mask_pm30deg	txx: sea_mask_pm30deg	txx: sea_mask_pm30deg
7	fd: nomask	gsl: nomask	fd: nomask
8	gsl: nomask	txx: land_mask	gsl: nomask
9	txx: land_mask	fd: nomask	txx: land_mask
10	txx: lat_mask_pm30deg	txx: lat_mask_pm30deg	txx: lat_mask_pm30deg

A.2.4 Feature importance from Boruta(SHAP)

(a) Boruta feature importance in 2030



(b) Boruta feature importance in 2040

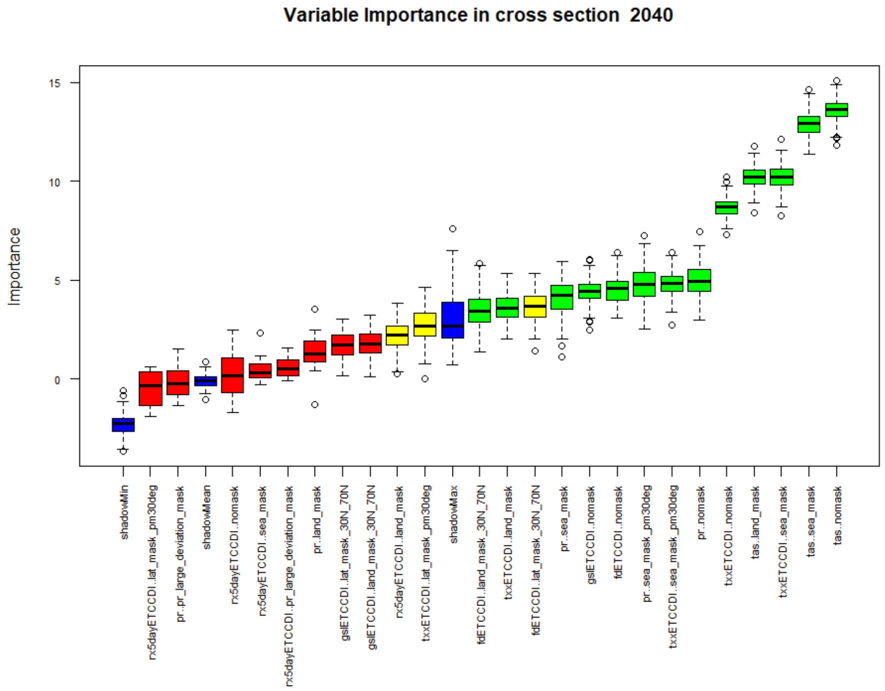
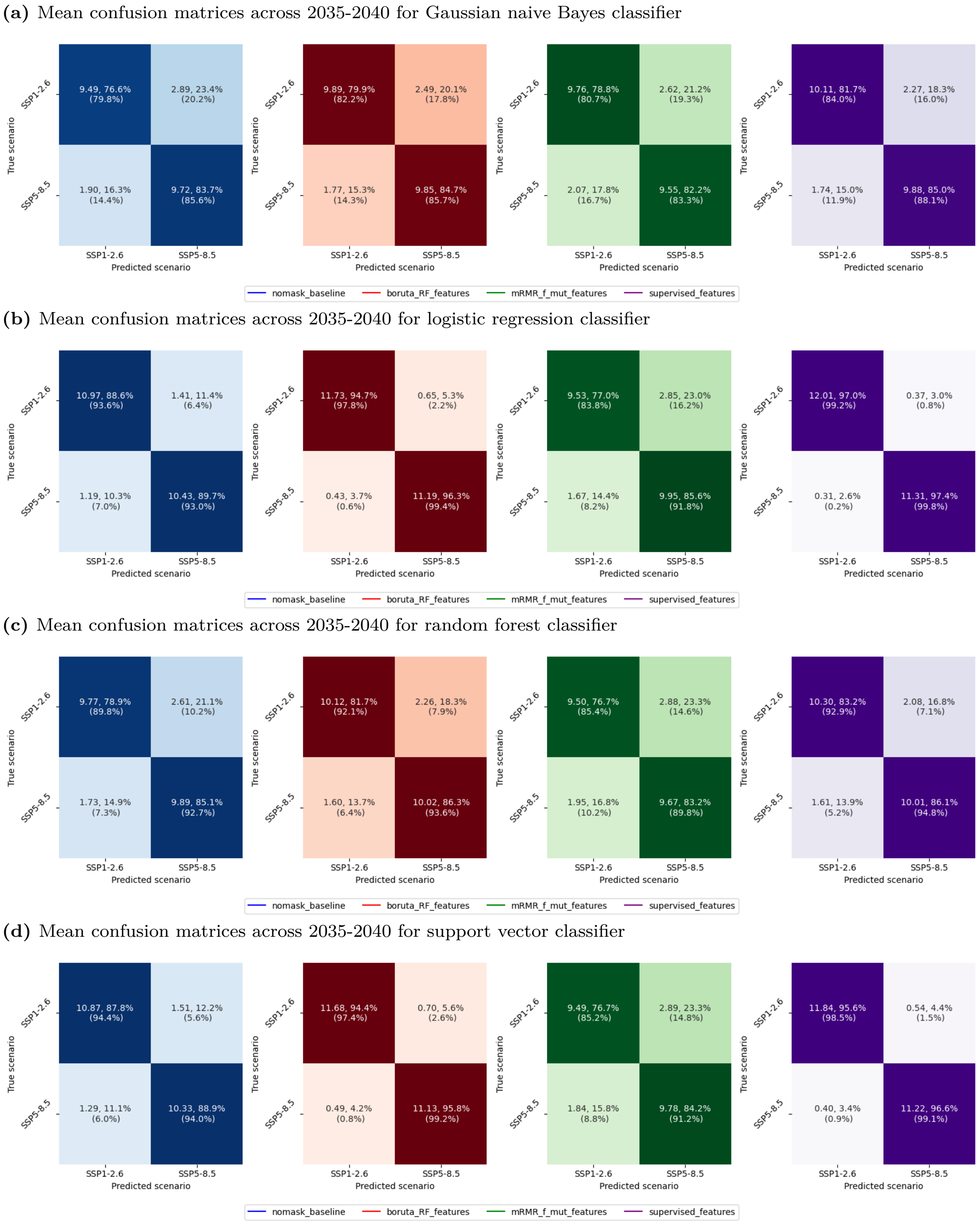


Figure A.2: The boruta feature importance for 2030 and 2040. The box-and-whiskers plot are generate from 100 iterations and show the distribution of the importance measured by SHAP values. Blue boxes indicate shadow features, here *shadoMax* is the benchmark feature for real features to beat. Red boxes denote features that are dismissed, green boxes signify features that are confirmed, and yellow boxes indicate features that are still under consideration by the algorithm.

A.3 Confusion matrices



(a) Mean confusion matrices across 2035-2040 for Gaussian naive Bayes classifier

(b) Mean confusion matrices across 2035-2040 for logistic regression classifier

(c) Mean confusion matrices across 2035-2040 for random forest classifier

(d) Mean confusion matrices across 2035-2040 for support vector classifier

Figure A.3: Mean confusion matrices across 2035-2040 for the classifiers. All numbers are averages calculated across 50 random states. The rows hold the true scenario, while the columns hold the predicted scenario. Within each matrix cell, the initial numerical value represents the mean count of classifications for that particular cell within the test-set, aggregated over 50 random states and spanning the years $\in [2035, 2040]$. The sequential percentage is the percentage of the row that is held in the cell, thus a measure of correct treatments of the specific rows' scenario. The percentage in parentheses is the corresponding percentage for the training-set.

A.4 Tune distributions of hyperparameters

A.4.1 GNB

For the GNB classifier we have tuned only one hyperparameter. This is the *'var_smoothing'* parameter that is used for calculation stability. Figure A.4 holds snapshots of 2020, 2030 and 2040 for each feature combination. Each snapshot holds the density distribution for the 50 *best models*, one for each random state. We observe a convergence towards var_smoothing = 1 as we progress throughout the century.

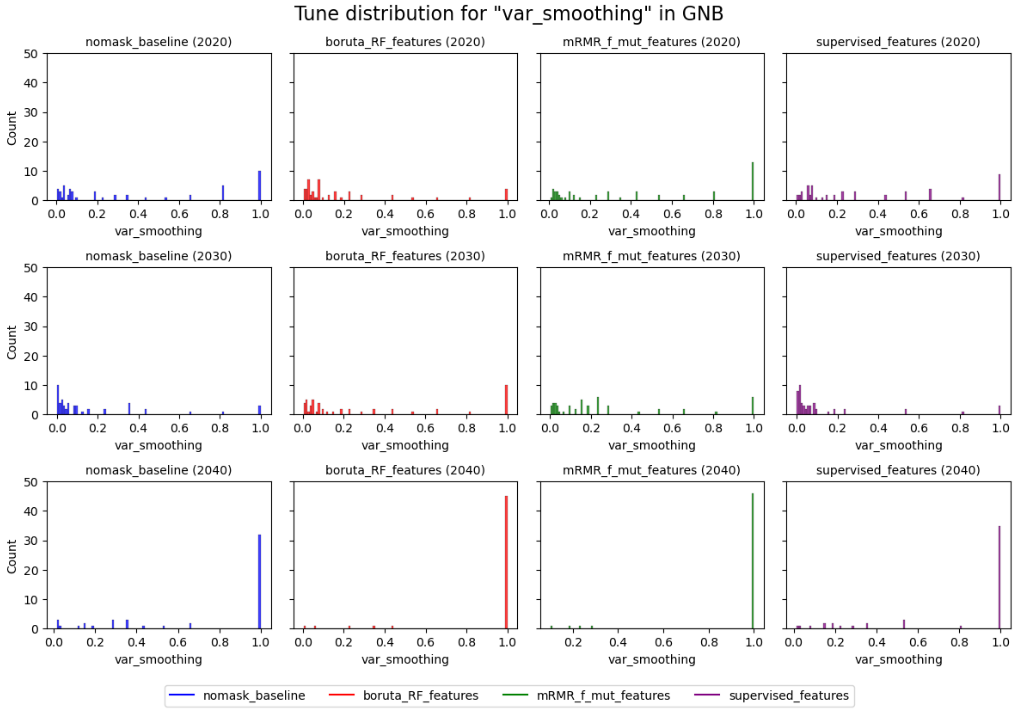


Figure A.4: Snapshots of the tune distribution of *'var_smoothing'* in the GNB classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

A.4.2 LR

For the LR classifier, we have tuned two hyperparameters, the l2 penalty parameter *'C'* and the algorithm for the optimization problem *'solver'*. Figures A.5 and A.6 hold snapshots of 2020, 2030 and 2040 for each feature combination. Each snapshot holds the density distribution for the 50 *best models*, one for each random state. In figure A.5 we show the 'l2' penelty hyperparameter *'C'* which is the inverse of the regularization strength. We observe a convergence towards larger values of *C* as we progress through the century.

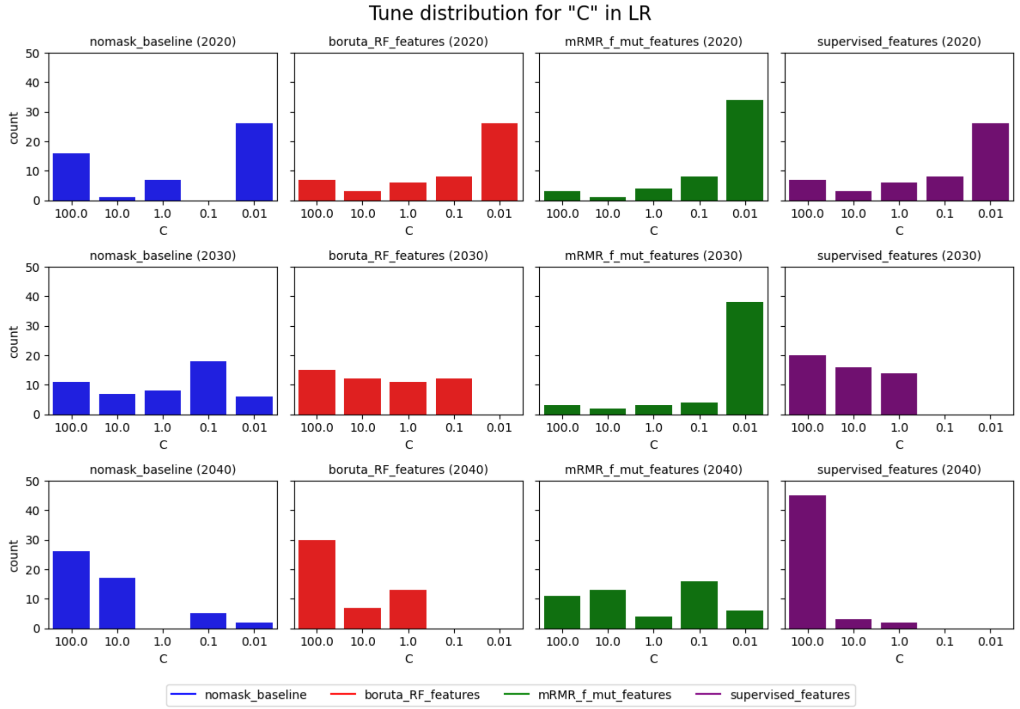


Figure A.5: Snapshots of the tune distribution of *'C'* in the LR classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'solver'* is the backbone of the optimization problem solved in the LR scheme. We observe a favoring of the newton-cg solver as we progress through the century.

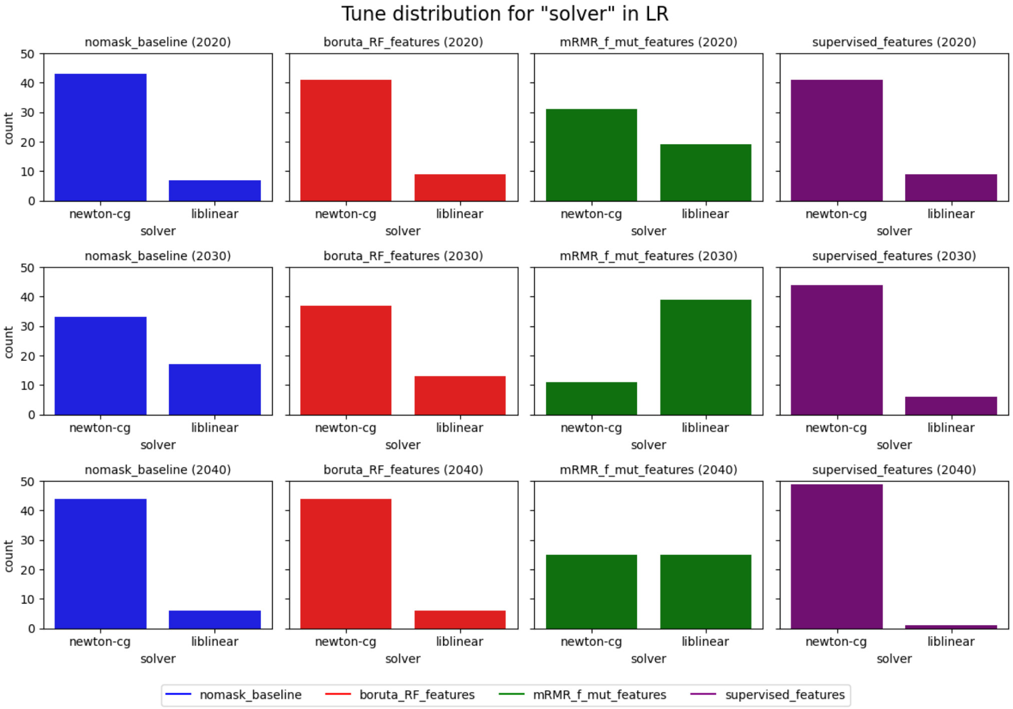


Figure A.6: Snapshots of the tune distribution of *'solver'* in the LR classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

A.4.3 RF

For the RF, we have tuned four hyperparameters, the pruning parameters penalty parameter *'max_depth'* and *'min_samples_leaf*, the *'max_features'* hyperparameter that governs the subset in the bagging, and the number of decision trees in the forest *'n_estimators'*. Figures A.7, A.8, A.9 and A.10 hold snapshots of 2020, 2030 and 2040 for each feature combination.

The *'max_depth'* is a pre-pruning technique that regulates the maximum depth of the decision trees in the random forest. By reducing the depth allowed, we effectively reduce the chance of over-fitting the individual trees. We observe a convergence towards shallow trees towards the end of the near-term period.

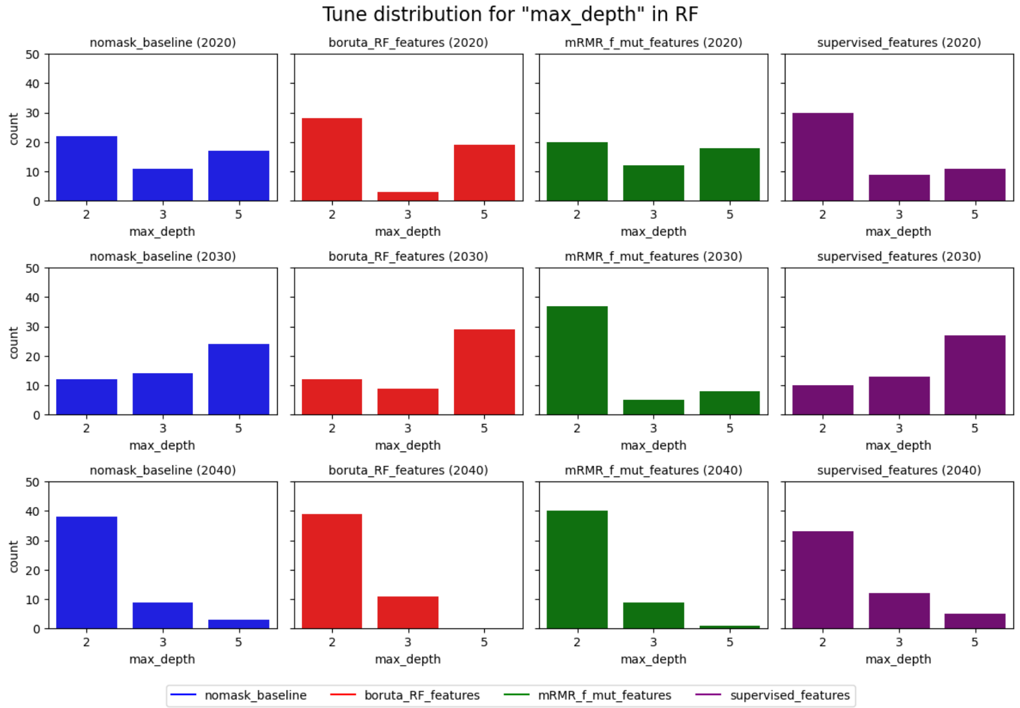


Figure A.7: Snapshots of the tune distribution of *'max_depth'* in the RF classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'max_features'* regulates the number of features considered in a split. We observe a convergence towards *'sqrt'* towards the end of the near-term period.

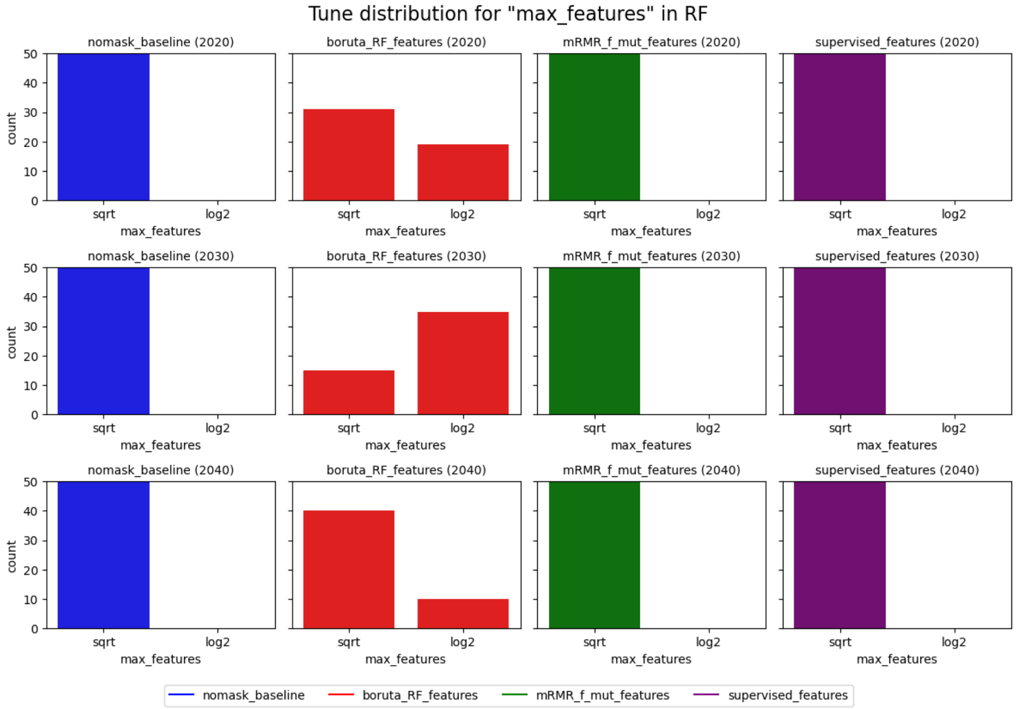


Figure A.8: Snapshots of the tune distribution of *'max_features'* in the RF classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'min_samples_leaf'* is a pre-pruning technique that regulates the minimum number of realizations which has to be in a leaf node for a split to be permisable. We observe a convergence towards the lower values towards the end of the near-term period.

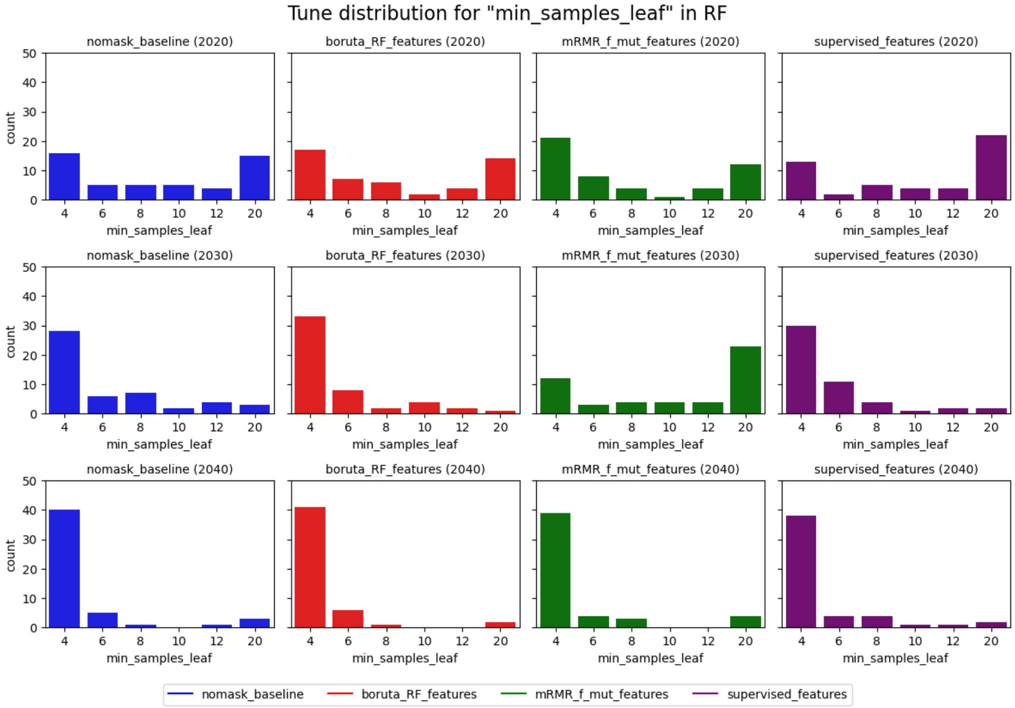


Figure A.9: Snapshots of the tune distribution of *'min_samples_leaf'* in the RF classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'n_estimators'* regulates the number of descision trees in the randomforest. We observe a convergence towards small forests towards the end of the near-term period.

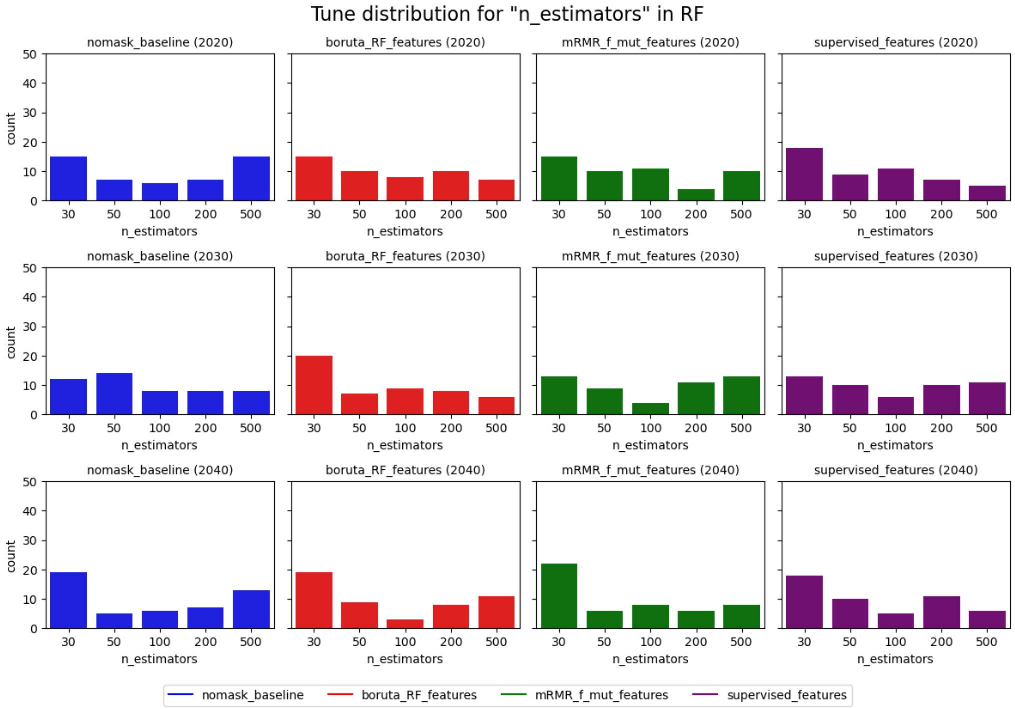


Figure A.10: Snapshots of the tune distribution of *'n_estimators'* in the RF classifier for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

A.4.4 SVC

For the SVC, we have tuned four hyperparameters, the l2 penalty parameter *'C'*, the *'kernel'* of the algorithm, for the *'poly'* kernel we tune the *'degree'* and the *'gamma'* parameter, in which adjusts the radius of a realization influence on the support vectors. Figures A.11, A.12, A.13 and A.14 hold snapshots of 2020, 2030 and 2040 for each feature combination.

The *'C'* is the regularization term of a SVC. A lower value will broaden the margin around the decision boundary. We observe a convergence towards no regularization towards the end of the near-term period.

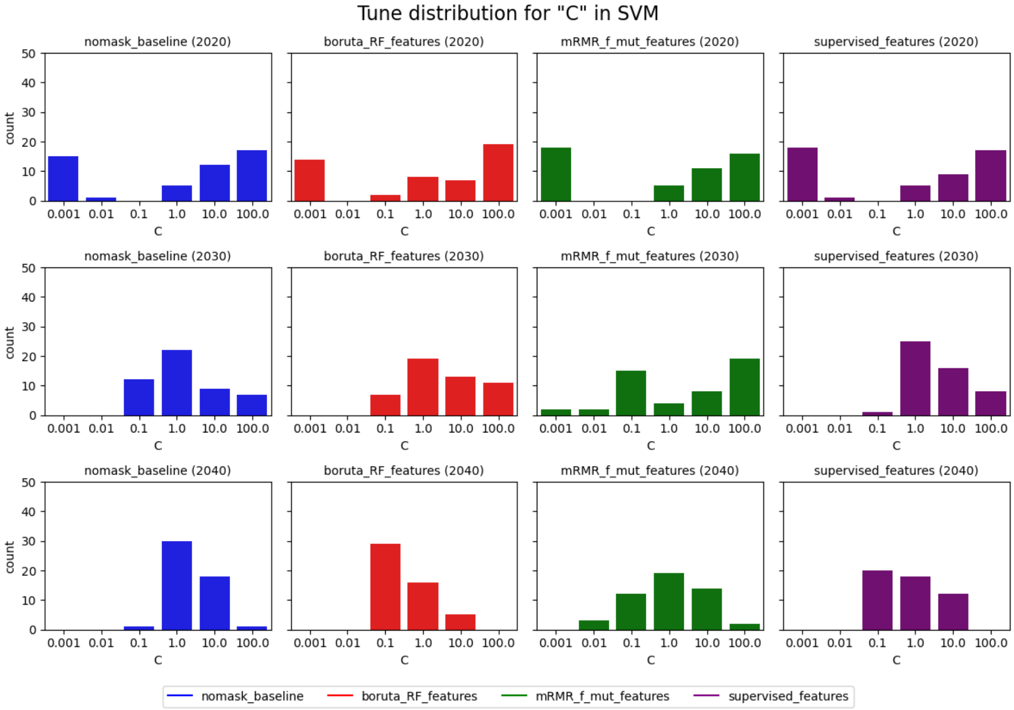


Figure A.11: Snapshots of the tune distribution of *'C'* in the SVC for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'kernel'* of a SVC describes the transformation of the data into a space where it is linearly separable. We observe a convergence towards the *'linear'* kernel towards the end of the near-term period.

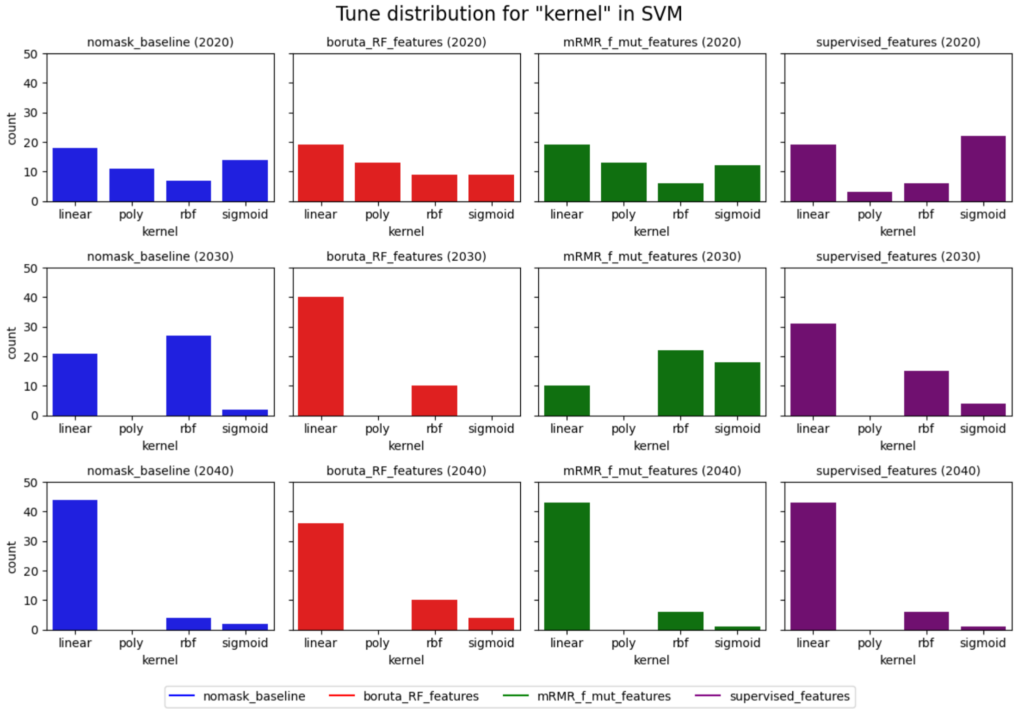
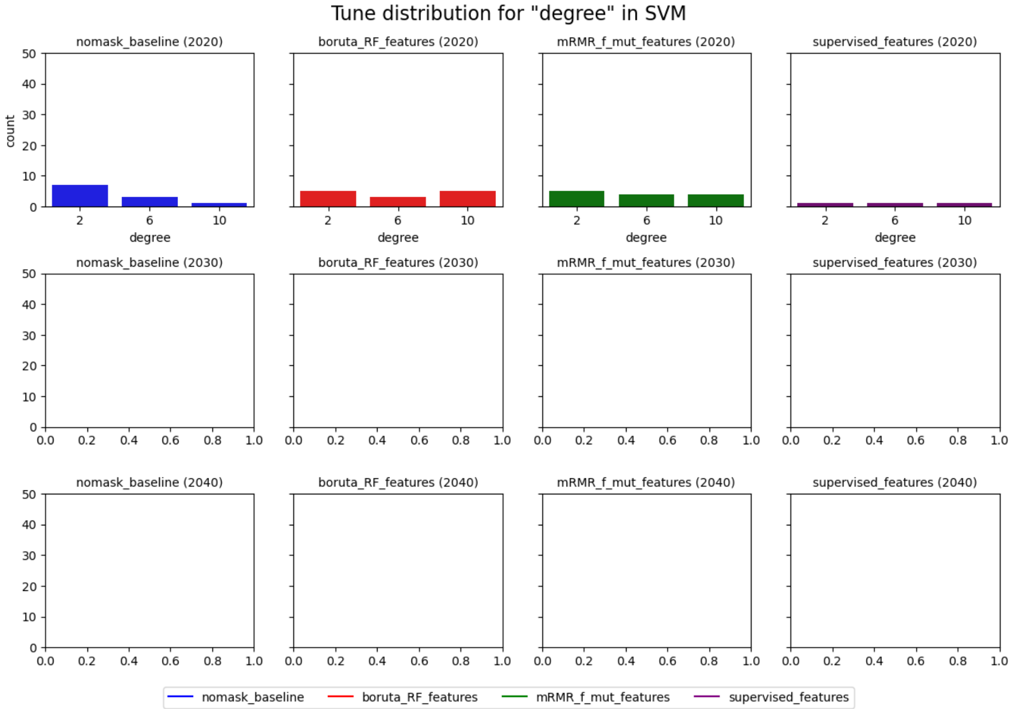


Figure A.12: Snapshots of the tune distribution of *'kernel'* in the SVC for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'degree'* of a SVC describes the degree of the *'poly'* kernel. Since it is specific to only one kernel, we only get a hit when the *'poly'* kernel produces the best model. Since the *'poly'* kernel has no instances of being the *best* model in 2030 and 2040 there are no observations in the plot.



>

Figure A.13: Snapshots of the tune distribution of *'degree'* in the SVC *'poly'* kernel for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

The *'gamma'* is a regularizations parameter specific to *'poly'*, *'rbf'* and *'sigmoid'* kernels. It influences the decision boundary's flexibility and thus regulates the over- and under-fitting of the model. Since the *'linear'* kernel is dominant in the later snapshots, there are fewer instances than 50 in each subplot. We observe a slight favoring of the *'scale'* version of *'gamma'*, however, both expressions will effectively be close to equal since the data is standard scaled before model training.

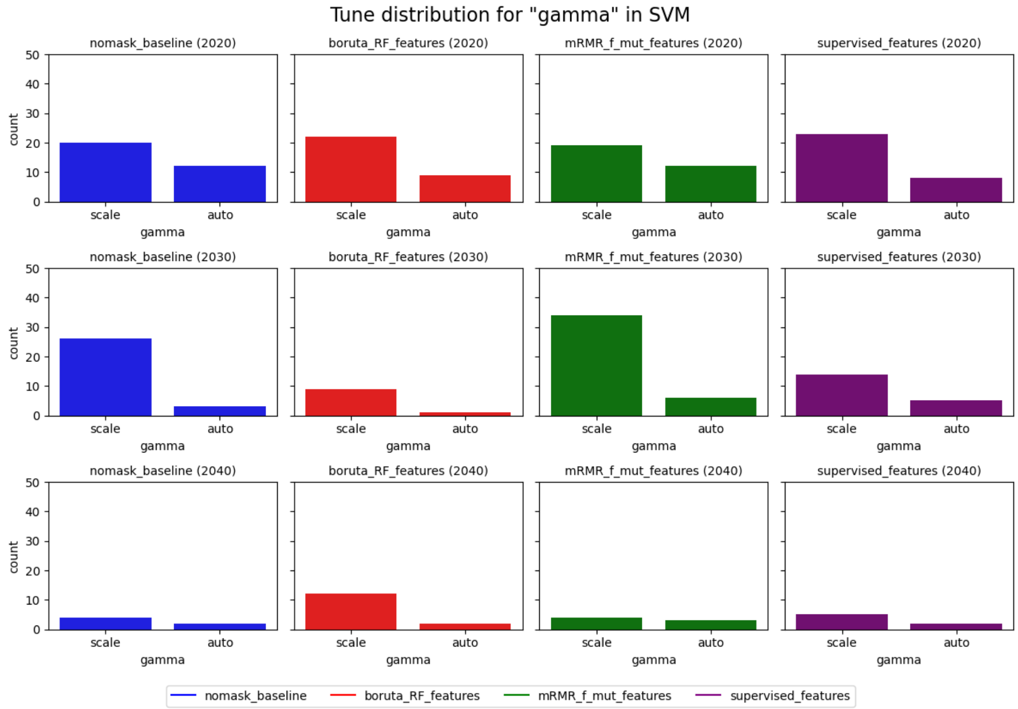


Figure A.14: Snapshots of the tune distribution of *'gamma'* in the SVC *'poly'* kernel for 2020, 2030 and 2040. Each snapshot holds the density distribution for the 50 *best models*, one for each random state.

A.5 Artificial intelligence as a writing aid

Throughout the thesis workflow, generative artificial intelligence (AI) has been used to ensure better linguistic fluency and process effectiveness. The aids include:

- **GitHub Copilot:** I have used GitHub copilot to aid the coding process. Main features of the prompt process include help with plotting and auto-completion. All code structure and functionality are quality guaranteed by me. For more information about this tool see https://github.com/features/copilot
- **Proofreading:** I have used Scribbr's AI proofreader to aid the proofreading process. Here I have not used any user-specified prompt, but the proofreading was conducted using English (US). For more information on this tool, see https://www.scribbr.com/proofreading-editing/.
- **Paraphrasing:** In some sections, I have used Scribbr's AI paraphrasing tool to help ensure a good linguistic fluency. It has only been used to edit individual sentences and has never been used to generate content from scratch. The main configuration I used include: English (US), Fluency, and synonyms set to the lowest setting. For more information on this tool, see https://www.scribbr.com/ai-writing/.

Thus, no section of the thesis is written without human supervision, and the quality of the thesis reflects my work.

A.6 Code availability

The code for all of the analysis used in this thesis is available on GitHub at:

https://github.com/Johannesfjeldsaa/Masterthesis_S23

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway