DIPLOMA THESIS

Creation and evaluation of a database for the Austrian Precipitation Sampling Network

by

Peter Redl, BSc
Student ID 01425015

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

Master programme Technical Chemistry, Study Code: 066 490

conducted at Institute of Chemical Technologies and Analytics, TU Wien

Advisors: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Anne Kasper-Giebl
Ao.Univ.Prof. Mag.rer.nat. Dr.rer.nat. Johann Lohninger

Vienna, February 28, 2022

Peter Redl

Anne Kasper-Giebl

Acknowledgement

For my master thesis I was given the opportunity to take part in and further develop the Austrian Precipitation Sampling Network. I am humbled that I was able to contribute to this multi decade research program. On the way I was able to deepen my data science and project management skills in addition to learning a lot about environmental science.

At this point I would like to express my sincere gratitude to everyone who supported my work in a direct or indirect way. This includes but is not limited to the members of the environmental analysis working group who created an amazing working atmosphere even in times of social distancing. Furthermore, I would like to thank our partners at the federal governments and all collaborators who worked ceaselessly to collect data over decades. For their incredible support I would like to thank my family, including my brother who encouraged my interest for data science. For his mentoring in statistics and important notes to section 3, I thank Hans Lohninger. A very special thanks goes to my supervisor Anne Kasper-Giebl for her unrestricted commitment to this project.

Without you this work would't have been possible.

Abstract

Long term measurement campaigns like the Austrian Precipitation Sampling Network continuously need to improve their workflows to meet current scientific needs. This applies not only to measurement but also to data processing. This work focusses on the implementation of a new database for precipitation data and it presents several use cases.

The created database includes all available samples processed by the Austrian Precipitation Sampling Network since its start in 1984. It meets the requirements of Global Atmosphere Watch for level 1 data and it includes a flag system, which represents the results of the ongoing data review process. The developed workflow ensures data integrity from measurement to reporting by using the unified database as the essential center piece. According Jupyter Notebooks, containing input and output Python scripts, were realized to enable straightforward data handling.

Based on the merged dataset it was possible to create a random forest classifier, that transfers the current standard of data reviewing on older parts of the dataset, that were not reviewed under the same testing regime. The classification showed that possibly overlooked contaminated samples allow only minor changes to the observed trends. However, the created classifier cannot replace the manual data review process due to the fluid border between valid and contaminated samples.

Further examples of database applications focus on seasonality and trends. Clear seasonality is observed for ammonium, nitrate and sulfate concentrations, which peak in spring. Precipitation depth reaches its maximum in summer with the only exception of mount Sonnblick where no precipitation seasonality is identifiable. In addition seasonal time series were investigated separately. Faster decreasing sulfate concentrations in spring and summer compared to fall and winter were discovered in the inner-alpine region. Finally it was tested whether sulfur and nitrogen deposition time series are better reflected with two separate rather than one single Theil-Sein estimator. A comparison with emission data showed that this is only possible for sulfur and only at four out of twelve tested stations. Therefore, the use of one linear approximation is still a good option for most time series. Although, one possibility to better reflect the current situation is to exclude the oldest parts of the datasets and select a uniform starting point, like 1991.

The developed workflows proofed to be effective in facilitating data reviewing, reporting and analysis in general. By reducing the data wrangling effort, they free time for actual analysis. The creation of the database was an important step that will lead to the expansion of data reporting in the future.

Contents

1. Introduction

Deposition describes the insertion of atmospheric trace substances on surfaces like water, soil, buildings or vegetation. It can occur as dry or as wet deposition. The first one describes the direct transport and adsorption of gases or particulate matter without prior dissolution in an aqueous phase. Whereas wet deposition describes transport and deposition with prior dissolution in a liquid phase like rain, clouds or fog. This means that the terms wet and dry refer to the mechanism of transport, not the nature of the surface itself. (Finlayson-Pitts and Pitts, 2000)

1.1. History

A short historical outline is provided by Erisman and Draaijers (1995). This section highlights some of the important milestones in deposition research, which are discussed in more detail in the above mentioned outline. First wet deposition analysis was performed around 1750 by Andreas Sigismund Marggraf who distilled sampled rainwater. He found HCl, HNO_3, NaCl and lime. In the following century more and more components were identified. Sulfates were found by Julia de Fontenelle in 1819. 1827 Von Liebig discovered that plants were fertilised by NH_3 introduced by precipitation. In the second half of the 19th century a similar effect caused by NO_3^- was shown by Jean-Baptiste Boussingault. He was the first who analyzed rime, fog, snow and dew for NH_4^+ and NO_3^-. At the same time first dry deposition measurements were carried out and the environmental impact of H_2SO_4 caused by coal burning and industry emission was investigated. The first big deposition measurement campaign focused on influence on public health was initialized in England by R. A. Smith who measured at multiple locations around 1870. In the following decades additional data was collected in Germany, France and Russia. (Erisman and Draaijers, 1995)

1.2. Measurement networks

The emergence of the first international measurement networks is outlined by Fowler et al. (2020). Culminating with the great smog of London in 1952, pollution became a defining issue. Research on atmospheric pollution expanded rapidly and the first monitoring networks were formed. In 1955 the European Air Chemistry Network (EACN) was founded by Scandinavian scientists. It enabled the creation of deposition maps and investigation on long-term changes in precipitation chemistry. Through the 1950s and 1960s steadily increasing acid and sulfat concentrations in precipitation were observed. This phenomenon was called acid rain and the first major conference on the subject was held in 1975. It was recognized that sulfur emissions from industrialized Europe influenced even remoter areas like Scandinavia

through long-range transport. Freshwater acidification caused a decline in fish populations and extensive die-back of forests was observed in the most polluted regions. The requirement to reduce European emissions became clear and therefore the Convention on Long-Range Transboundary Air Pollution (LRTAP) was established. The EACN programme ended in 1976 but many stations continued within the European Monitoring and assessment Programme (EMEP), which is focused on monitoring, modeling and evaluation of long-range transmission of air pollutants in Europe. At the global scale, monitoring is now coordinated through the Global Atmosphere Watch (GAW) Programme, which is part of the World Meteorological Organization (WMO). During the 1980s the control on emission through LRTAP protocols showed effect and sulfur and acid deposition declined steadily. In 2016 SO_2 emissions in Europe and North America were reduced by approximately 90 % from their peak values in the 1970s and 1980s. (Fowler et al., 2020)

In Austria first systematic studies on rainwater composition began in 1957 with one station in Retz, which was part of EACN and the Background Air Pollution Monitoring Network (BAPMON) (Cehak and Chalupa, 1985; Puxbaum et al., 2002). But the Austrian Precipitation Sampling Network with stations in several federal states startet full operation as early as 1984. Figure 1.1 shows all active and inactive sites of the network. Some stations were only in operation for a short period of time, while others provide multi decade datasets. An overview on the available data is given in figure 1.2. All sites are listed in table 2.1 on page 10. Three of the active sites (Sonnblick, Masenberg, Haunsberg) are also part of the EMEP network, which currently includes 17 active stations in Austria (`https://projects.nilu.no/ccc/sitedescriptions/at/index.html`). In addition Sonnblick is one out of 20 GAW Global stations to provide background data.

1.3. Analytical methods

Sampling and analysis is based on a guideline of the Ministry of Health and Environmental Protection (BMGU, 1984) and on the GAW Manual (Allen, 2004). Therefore, the used method for precipitation analysis provides international comparability. Regular participation on round robin tests is used to further develop this comparability.

All stations of the Austrian Precipitation Sampling Network are equipped with Wet And Dry Only precipitation Samplers (WADOS) by Kroneis GmbH. This instrument enables separated sampling of dry and wet deposition. For wet precipitation sampling it meets the standards of WMO, including the GAW Manual Field Protocols updated in July 2021 and the standards of ISO for atmospheric dustfall sampling. The device has one container for wet and one container for dry sampling. An electronic driving mechanism moves the lid from the wet container to the dry container when the precipitation sensor detects rain or snow. In this case precipitation is transferred through a polypropylene funnel in a high-density polyethylene (HDPE)

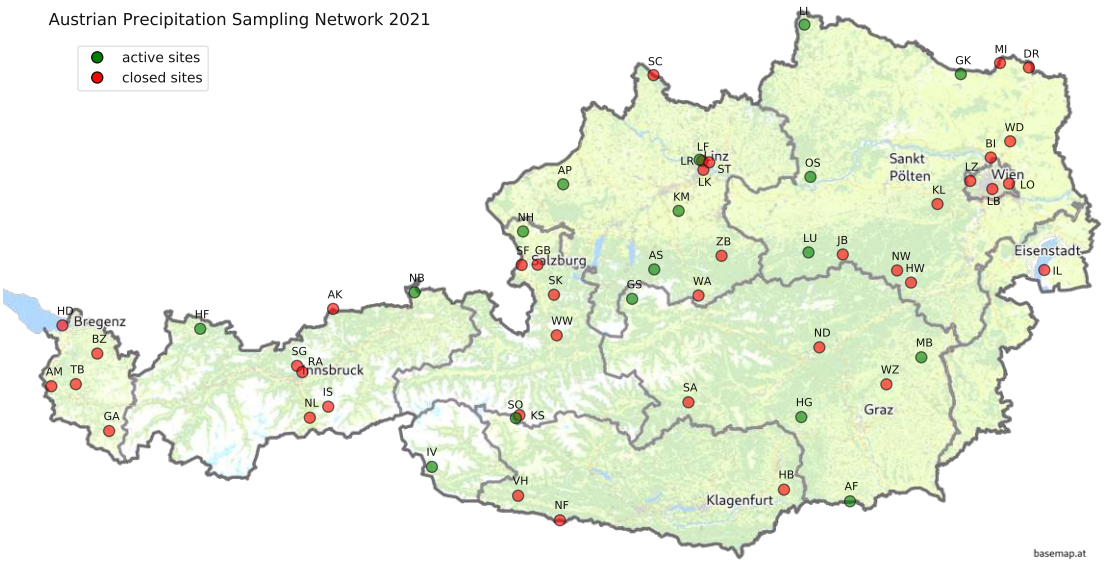


Figure 1.1: Active and inactive sites of the Austrian Precipitation Sampling Network

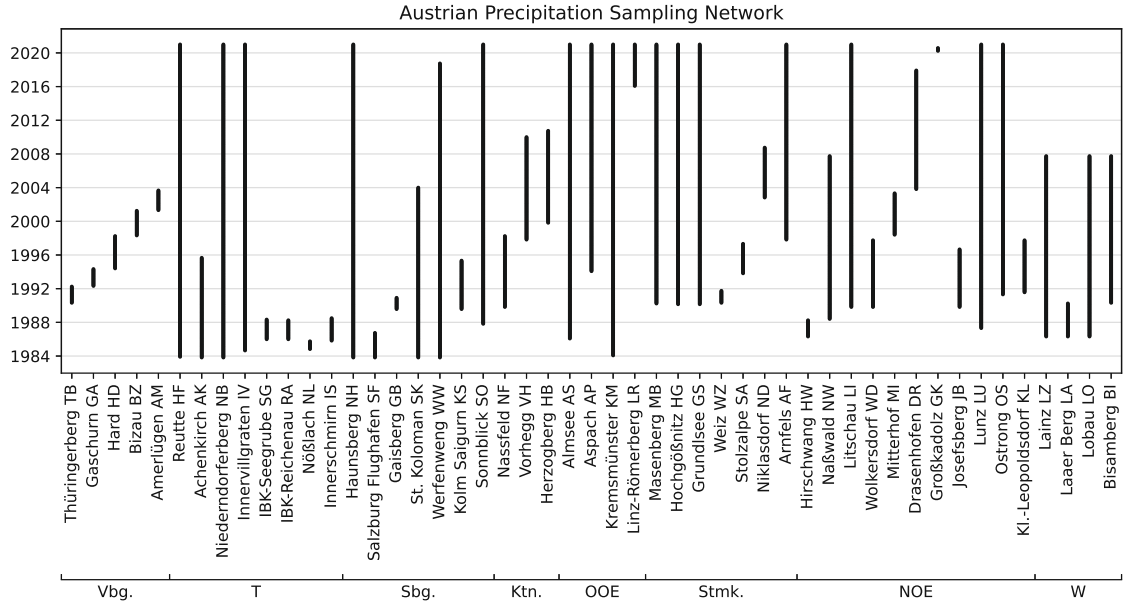


Figure 1.2: Available data of the Austrian Precipitation Sampling Network

sample-bottle. Bottle change is performed manually in the morning or at the Styrian stations automatically at midnight. The precipitation detector is heated to 20 °C to prevent dew formation on the one hand and melt snow on the other hand. During precipitation events temperature is increased to 50 °C to ensure a fast evaporation after the event. Five minutes after the end of the event the lid returns to the dry deposition sampling position. The base of the collecting funnel is heated as well to melt sampled snow and ice. It is kept between 8 and 10 °C to minimize loss through evaporation. (Kroneis GmbH, 2005)

Samples are stored in refrigerators before and after transportation to the laboratory. To prevent redox reactions and microbial activity, samples from Sonnblick Observatory are frozen after sampling and thawed right before measurement. Procedures and schedules vary for different federal states. For example Styrian samples are delivered annually blocked to the Institute of Chemical Technologies and Analytics (CTA), while samples from Lower Austria arrive on a regular basis. Other federal states like Tyrol, Salzburg and Upper Austria operate their own laboratories. An overview on the responsible laboratories is given in appendix A.1. For detailed information on measurement, the annual reports on wet deposition, published by the federal governments, can be consulted. In table 1.1 the lab equipment used at CTA in 2021 is given as an example for ion analysis. For pH analysis an InLab Pure Pro-ISM electrode by Mettler-Toledo was used. Conductivity is measured with a Mettler-Toledo InLab 720 electrode (conductivity range 0 - 500 $\mu S\,cm^{-1}$, temperature range 0 - 100 °C).

Table 1.1: Analysis system at TU Wien 2021

	cation analysis	anion analysis
system	Dionex-Aquion	Dionex ICS 1100
column	Dionex Ion Pac CS16	Dionex Ion Pac AS22
precolumn	Dionex Ion Pac CG16	Dionex Ion Pac AG22
eluent	38 mM MSA	4.5 mM Na_2CO_3/1.4 mM $NaHCO_3$
flow	1 mL min^{-1}	1 mL min^{-1}
suppressor	Dionex CSRS 500 - 4 mm (electrochemical)	Dionex AERS 500 - 4 mm (electrochemical)
regenerant	cycled eluent	cycled eluent
sampling loop	150 µL	100 µL
detection	conductivity cell	conductivity cell
software	Chromeleon 7.2.9	Chromeleon 7.2.9

Naturally the lab equipment and the measurement parameters have been subject to change since operation start. For example bivalent cations were determined by atomic absorption spectroscopy until 1993. After that they were measured via ion chromatography like the other ions. In addition the laboratories in Salzburg and Tyrol use different equipment as well. A summary of all used systems and their

respective detection limits (LODs) is given by Firmkranz (2019). Typically all ion LODs are in the range of 0.01 - 0.03 mg L^{-1}. Less than 1 % of all samples exhibit Cl^-, $NO_3{}^-$, $SO_4{}^{2-}$ and Ca^{2+} concentrations below the LOD. Na^+ and $NH_4{}^+$ was below the LOD in around 2 % of the samples. The lowest overall concentrations are found for K^+ and Mg^{2+}, which leads to 9 and 7 % of the samples below the LOD. If only the background station on mount Sonnblick is considered, the shares for Cl^-, $NO_3{}^-$, $SO_4{}^{2-}$, Ca^{2+} and Na^+ increase to 1 - 3 %. $NH_4{}^+$ concentrations are below the detection limit for 6 % of these samples. Regarding K^+ and Mg^{2+} the same is true for approximately one quarter of the Sonnblick samples. The numbers for this short overview represent the proportions of samples with concentrations below or equal to 0.01 mg L^{-1}. However, a more detailed evaluation of this subject will be part of future work.

1.4. Aim of this work

This thesis follows the work of Schreiner (2017) and Firmkranz (2019). Both were conducted at the Institute of Chemical Technologies and Analytics. Schreiner elaborates on trends and seasonality of concentration and deposition data. Several statistical methods are investigated and map plots are used to show local influences. Her work is based on monthly and annual data from 1983 to 2014. Firmkranz (2019) uses individual sample data from 2014 to 2017 to calculate ion and conductivity balances. These are combined in Miles and Yost diagrams (Miles and Yost, 1982) to uncover relations between sample composition and local influences on sampling. The thesis is focused on measurement and workflow optimization and therefore contains detailed information on measurement.

This work tries to combine the scope of Elisabeth Schreiners work, that covered the complete time series for many stations, with the granular approach of Julia Firmkranz who investigated individual samples. Therefore, a new database was created that includes all available data of the Austrian Precipitation Sampling Network (see fig. 1.2 on page 3). The goal of this database is to provide level 1 data, which according to GAW is defined as instrument data processed to physical parameters without time aggregation or contamination removal (see `https://www.gaw-wdca.org/Submit-Data/Advanced-Data-Reporting/Level-1`). In addition it should include all information necessary to calculate aggregated corrected results, that can be used in reports. This is achieved via integration of a flag system. A previous attempt to create a unified precipitation sample database was not successful because it was designed as a data backup with no data integrity checks. The new database is planned to be tightly integrated in the report creation process. Therefore, a complete redesign of the data processing workflow is necessary. Previously the workflow consisted of multiple Excel files, a python script for trend analysis and an Origin template to create map plots. The new process is centred around an import script and an export script for the database written in Python.

The unified database enables new opportunities in per sample analysis. This work features a random forest classification to find overseen contaminated samples and an assessment of their impact on trends.

But the database also simplifies more traditional analysis types that are based on aggregated data because the aggregation process is streamlined and can be used with different aggregation periods and for every station of choice. This greatly accelerates the data acquisition process, which frees up more resources for the actual analysis. In this work a seasonal trend analysis is performed that reveals which components follow a seasonal trend and whether there are differences between the various stations. In addition the concentration development over time is investigated separately for each season. Lastly the observed depositions are compared with emission data to identify common trends and to verify the linear model, which is currently used to describe deposition trends.

2. Creation of a precipitation database

2.1. Previous work

At the Institute of Chemical Technologies and Analytics samples of the Austrian Precipitation Sampling Network are usually processed by dedicated employees who are also responsible for report creation. These reports are prepared for the participating federal governments on an annual basis. The needed knowledge is passed on between personal. However, all responsible people, which are listed in appendix A.2, developed and shaped the existing workflow to ensure state of the art analysis.

Figure 2.1 depicts a simplified version of the previous data processing workflow that is used for report creation. At first the data is imported from the source, which can be the output of the measurement system or an intermediate Excel sheet that contains necessary conversions. *Ionenprüfen* is an Excel file template that incorporates macro code that performs time aggregation and creates multiple plots and tables that are needed for quality assurance. This includes ion and conductivity balances as well as additional statistics. *Ionenprüfen* files are created for every station and cover one period (Oct. - Sep.). Annual and monthly aggregated data is then transferred to the long-term Excel sheet *langjährig*, which exists for every station and contains data of all previous periods. In addition non aggregated data is added to the Access database that contains tables for every station. However, the content of the Access database is not used for further steps. As the scope of the reports expanded over the years additional analysis steps were added. Some of these were not integrated in the existing *ionenprüfen* and *langjährig* templates. For example a python script was used for statistical trend analysis, which created the necessity for an Excel python interface. This was executed via separate CSV files for concentration and deposition, which further increased maintenance effort.

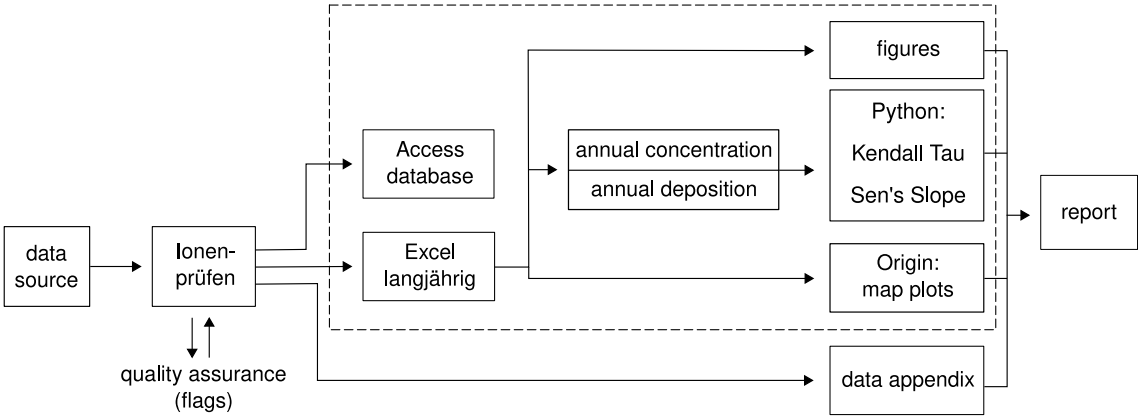


Figure 2.1: Previous data processing workflow for report creation

The previous and grown workflow has two main downsides. Firstly the complex structure requires some manual copy and paste steps. These are prone to error due to column mix-ups. In addition no automated update process prevents the transmission of outdated data and unless increased attention is spent it may appear in final results. Secondly, the terminal position of the Access database does not encourage a continuing data review process. This led to incomplete and in some cases inhomogeneous datasets.

2.2. Unifying approach

To counter above mentioned problems a new workflow was designed. Its main idea is to replace all tasks and files within the dashed box of figure 2.1 with a new database and scripts for data import and export. *Ionenprüfen* was not included as Excel provides several features that proofed to be effective for quality assurance like context colored cells or live-updated graphs.

For the new database several file format options were explored like SQLight or a revised Access database. Finally it was realized that no sophisticated solution is necessary to store the data. In fact a simple solution is preferred because operation has frequently to be taught to new staff. Therefore, a text file with comma-separated values (CSV) was used. Most programmes can open CSV files with one million rows or more. Currently almost 80 000 samples were processed for the Austrian Precipitation Sampling Network. This means there is no bottleneck to be expected. In addition CSV files have shown exceptional longevity in contrast to constantly changing Microsoft file formats.

For the scripts Python was used because it is most popular among data scientists as well as software engineers (Capellupo, 2021) and an emerging programming language in general (see `https://www.tiobe.com/tiobe-index/`). As development environment Jupyter Notebook was used, as it features interactive inline charts for fast data analysis and separated code cells to structure the scripts.

2.3. Database initialization

As stated in section 2.1 previous attempts to create and maintain databases with all precipitation samples were not successful. Nevertheless, parts of incomplete datasets were used for the foundation of the new database that was codenamed rainybase. All parts were imported and merged via script to prevent errors. Overall more than 150 files were combined. Most of them contain only data of one station. Therefore, station IDs were not present and had to be added. The Austrian Precipitation Sampling Network consists of 46 station, although not all of them are still active. Station information is given in table 2.1.

Table 2.1: Stations of the Austrian Precipitation Sampling Network

ID	name	state	latitude	longitude	height	start	end
TB	Thüringerberg	Vbg.	47.2181	9.7847	960 m	Apr 90	Mar 92
GA	Gaschurn	Vbg.	46.9917	10.025	990 m	Apr 92	Apr 94
HD	Hard	Vbg.	47.5022	9.6881	400 m	May 94	Mar 98
BZ	Bizau	Vbg.	47.3661	9.9394	700 m	Apr 98	Mar 01
AM	Amerlügen	Vbg.	47.2081	9.6081	770 m	Apr 01	Aug 03
HF	Reutte	T	47.4857	10.6819	930 m	Nov 83	
AK	Achenkirch	T	47.5819	11.6403	840 m	Oct 83	Aug 95
NB	Niederndorferb.	T	47.6621	12.2269	680 m	Oct 83	
IV	Innervillgraten	T	46.8183	12.3528	1730 m	Aug 84	
SG	IBK-Seegrube	T	47.3067	11.38	1960 m	Dec 85	Apr 88
RA	IBK-Reichenau	T	47.2767	11.4181	570 m	Dec 85	Mar 88
NL	Nößlach	T	47.0561	11.4722	1420 m	Oct 84	Sep 85
IS	Innerschmirn	T	47.1094	11.605	1570 m	Oct 85	Jun 88
NH	Haunsberg	Sbg.	47.9564	13.01	520 m	Oct 83	
SF	Sbg. Flughafen	Sbg.	47.7944	13.0003	433 m	Oct 83	Sep 86
GB	Gaisberg	Sbg.	47.7958	13.1147	1010 m	Jul 89	Nov 90
SK	St. Koloman	Sbg.	47.6503	13.2328	1020 m	Oct 83	Dec 03
WW	Werfenweng	Sbg.	47.4542	13.2528	940 m	Oct 83	Sep 18
KS	Kolm Saigurn	Sbg.	47.0683	12.9842	1600 m	Jul 89	Apr 95
SO	Sonnblick	Sbg.	47.0542	12.9578	3106 m	Oct 87	
NF	Nassfeld	Ktn.	46.5603	13.2758	1530 m	Oct 89	Mar 98
VH	Vorhegg	Ktn.	46.6786	12.9744	1020 m	Oct 97	Dec 09
HB	Herzogberg	Ktn.	46.7083	14.8917	540 m	Oct 99	Sep 10
AS	Almsee	OOE	47.7728	13.9561	591 m	Jan 86	
AP	Aspach	OOE	48.1842	13.2997	430 m	Jan 94	
KM	Kremsmünster	OOE	48.0556	14.1319	384 m	Jan 84	
LR	Linz-Römerb.	OOE	48.303	14.2821	262 m	Jan 16	
MB	Masenberg	Stmk.	47.3481	15.8822	1137 m	Mar 90	
HG	Hochgößnitz	Stmk.	47.0592	15.0167	900 m	Feb 90	
GS	Grundlsee	Stmk.	47.6306	13.7967	954 m	Feb 90	
WZ	Weiz	Stmk.	47.2175	15.6303	456 m	Apr 90	Sep 91
SA	Stolzalpe	Stmk.	47.1306	14.2028	1302 m	Oct 93	Apr 97
ND	Niklasdorf	Stmk.	47.3961	15.1469	510 m	Oct 02	Sep 08
AF	Arnfels	Stmk.	46.6519	15.3678	763 m	Oct 97	
HW	Hirschwang	NOE	47.7092	15.8078	500 m	Apr 86	Mar 88
NW	Naßwald	NOE	47.7678	15.7072	600 m	May 88	Sep 07
LI	Litschau	NOE	48.956	15.039	560 m	Oct 89	
WD	Wolkersdorf	NOE	48.3922	16.5227	180 m	Oct 89	Sep 97

Table 2.1: Stations of the Austrian Precipitation Sampling Network

ID	name	state	latitude	longitude	height	start	end
MI	Mitterhof	NOE	48.7706	16.4497	179 m	May 98	Apr 03
DR	Drasenhofen	NOE	48.7489	16.6578	216 m	Oct 03	Nov 17
GK	Großkadolz	NOE	48.7122	16.1931	191 m	Mar 20	Jul 20
JB	Josefsberg	NOE	47.845	15.3156	1010 m	Oct 89	Aug 96
LU	Lunz	NOE	47.855	15.0686	618 m	Apr 87	
OS	Ostrong	NOE	48.22	15.0825	575 m	Apr 91	
KL	Kl.-Leopoldsd.	NOE	48.0889	15.9989	400 m	Jul 91	Sep 97
LZ	Lainz	W	48.2006	16.2353	230 m	Apr 86	Sep 07
LB	Laaer Berg	W	48.1614	16.3942	250 m	Apr 86	Mar 90
LO	Lobau	W	48.1875	16.5142	155 m	Apr 86	Sep 07
BI	Bisamberg	W	48.3136	16.3831	310 m	Apr 90	Sep 07

To complement the measurement data, a flag system was implemented. It consists of one flag column for every measurement column. That leads to a total of 29 columns in the database, which are listed in table 2.2.

Table 2.2: Information of a database entry

columns	description
Datum	date of the precipitation event (as sampling takes place at midnight or in the morning this is the day before sampling)
Ort	station ID; see table 2.1
NS / NS_flag	precipitation amount in mm and according flag
LF / LF_flag	conductivity in $\mu S\,cm^{-1}$ and according flag
pH / pH_flag	pH value and according flag
NH4 / NH4_flag	NH_4^+ concentration in $mg\,L^{-1}$ and according flag
Na / Na_flag	Na^+ concentration in $mg\,L^{-1}$ and according flag
K / K_flag	K^+ concentration in $mg\,L^{-1}$ and according flag
Ca / Ca_flag	Ca^{2+} concentration in $mg\,L^{-1}$ and according flag
Mg / Mg_flag	Mg^{2+} concentration in $mg\,L^{-1}$ and according flag
Cl / Cl_flag	Cl^- concentration in $mg\,L^{-1}$ and according flag
NO3 / NO3_flag	NO_3^- concentration in $mg\,L^{-1}$ and according flag
SO4 / SO4_flag	SO_4^{2-} concentration in $mg\,L^{-1}$ and according flag
Pb / Pb_flag	Pb^{2+} concentration in $\mu g\,L^{-1}$ and according flag
Cd / Cd_flag	Cd^{2+} concentration in $\mu g\,L^{-1}$ and according flag
Anmerkungen	text-based note

2.4. Flag system

Flags were specified to enable easy conversion to the flag system outlined by the GAW Manual (Allen, 2004) but also to reflect measurement reality of the Austrian Precipitation Sampling Network. In contrast to the GAW system, which uses letters, these flags are coded as single digit numbers. A list is given in table 2.3. The quality criteria mentioned for flag 2 are based on the laboratory operations section of the GAW Manual. However, no fixed limits regarding the ion and conductivity balances are used.

Table 2.3: Flag code description

flag	description	GAW equivalent
1	valid measurement	V
2	valid - does not meet quality criteria	V
3	valid - measurement below LOD - actual value reported	L
4	valid - measurement below LOD - replaced with LOD	D
5	invalid - contaminated	X
6	invalid - malfunction	M
7	no analysis - low volume	M
8	no analysis - missing sample	M
9	used only temporarily for extrapolated values	-

The majority of the merged files did not include a flag system. Invalid data was color coded, annotated or simply not used for further calculations. Therefore, flags had to be assigned manually in most cases. Many tools were created to facilitate and accelerate this process. One useful approach was to compare monthly aggregated data from *langjährig* files and from the database. For this purpose, an aggregation function (see section 2.6.3) was created. It calculates aggregated values based on the database in a similar way as in the old workflow, which led to the *langjährig* files. The difference is that the new function has proper date recognition as opposed to the macro in *ionenprüfen* that only works for one period. The general idea of both calculations is to produce aggregated concentrations weighted by precipitation amount. This must be done without considering contaminations. Therefore, the * in equation 1 denotes that the flagged measurements are set to NaN. For NO_3^-, NH_4^+ and SO_4^{2-} a conversion factor ($\frac{14}{62}$, $\frac{14}{18}$ and $\frac{32}{96}$) is used. It converts the ion masses to corresponding atom masses as concentrations are usually given in nitrogen and sulfur mass per volume.

$$\frac{\sum_{\text{month}} (\text{NS} \cdot \text{concentration}^*)}{\sum_{\text{month}} \text{NS}^*} \cdot f_{\text{conversion}} = \text{monthly concentration mg/L} \qquad (1)$$

To expose deviations the difference between the newly calculated aggregated concentrations and the values found in *langjährig* files were plotted as a time series. Missing flags in the database are one of several reasons for spikes in those graphs. All deviating months were manually checked and if possible differences were resolved by flagging single events or by data adjustments according to the original source files.

2.5. Encountered deviations

There are several other issues beside missing flags that had to be addressed. Some of them impacted the data more gravely than others and therefore had to be solved first like the date offset or inhomogeneously applied conversion factors. Non resolvable issues were listed in deviation reports for all stations, which are available at the environmental analysis working group. This way a clearly defined transition to the new system was ensured. The most frequently encountered problems are outlined in the following sections.

2.5.1. Date offset

As stated in table 2.2 the date given in the database resembles the date of the precipitation event, which more often than not is the day before sampling because the manual bottle change takes place in the morning and the automated change at Styrian stations is performed at midnight. Therefore, the sampling date has to be reduced by one day. Inconsistencies in the source files led to samples ending up in the wrong months. Although this is irrelevant for long term analysis, it hampers the above described approach to find missing flags. Fortunately periods with said date offset are easily identifiable by characteristic $\wedge$-shaped artifacts in the divergence line. Dates in these periods were shifted to match the results in the *langjährig* files.

2.5.2. Rounded intermediates

Missing flags and shifted days were not the only reasons for deviations between the results based on the database and the values given in the *langjährig* files. Especially in the first decade of the Austrian Precipitation Sampling Network intermediate results were rounded, which leads to deviations for almost every month before September 1994. Therefore, stations that started their operation very early like Höfen, Niederndorferberg, Innervillgraten, Haunsberg or Werfenweng are particularly affected. But also stations with low concentrations like Sonnblick show significant deviations that are caused by rounding. Like the date shift, this has no impact on long term data analysis. Some values, close to the detection limit

can deviate up to 50 % (e.g. 0.0149 → 0.01) or more if intermediates were rounded too. This is especially true for species with low concentrations like K^+ or Mg^{2+}. For the compared monthly aggregates this means average deviations in the order of ±10 %. Unlike the date shift, there were no attempts undertaken to compensate for this. According to the philosophy to round only final results, no rounding is applied in the database. Unfortunately these deviations complicate the data review process for the period before 1994, because deviations with other reasons are hidden in the noise. In addition, calculations and intermediate results of that time have not been preserved, which aggravates the process even further. Therefore, the review of data before September 1994 is not as complete as the rest and the causes for some differences could not be resolved.

2.5.3. Sample overflow

One resolvable issue was characterized by differing precipitation amounts. For some samples the rain volume stated by the station is overruled by measurements of the Austrian hydrographic service. This is the case for short out-of-service periods or for strong precipitation events that exceed the sample capacity of some of the used samplers. A WADOS equipped with a 1 L bottle exhibits sample overflow at precipitation events beyond 31 mm. In these cases the precipitation amount in the database was replaced with the available information from the hydrographic service. Therefore, the precipitation amounts given in the database resemble the amounts used in the respective *langjährig* files. Nevertheless, the precipitation amount distribution shows an accumulation around 30 mm, which means that some events with overflow were overlooked in the past. The Styrian stations but also Niederndorferberg and Haunsberg are mainly responsible for this. It is unlikely that this issue can be solved retroactively but the GAW Manual (Allen, 2004) introduced a solution for current sampling. It requires the use of a designated precipitation gauge in parallel with the precipitation chemistry sampler. Manual gauges are preferred, however correct gauge selection can be made by the national hydrographic service.

2.5.4. Duplicated samples

Another cause of deviations is the occurrence of duplicated samples in the original datasets. Duplicates may arise from different reasons. In the original files about 2 % of the samples have no unique date and station combination but about 1.7 % are caused by experimental measurements at the start of the campaigns in Innervillgraten and Litschau. In most other cases they originate from additional samples, actually taken to check the cleanliness of the WADOS, that were mistaken for actual samples. At most stations the sampling device is rinsed regularly for cleaning purposes. To identify device contamination sometimes rinsed water is sampled and measured. This creates the possibility of mix-ups. Other reasons for duplicates are

the shipping of two sample bottles for one event or incorrectly reported dates. Only one of each duplicated sample pair remained in the database. Whenever a sample of rinsed water was identified it was removed, in all other cases one of the duplicates was excluded. If none of the samples showed irregularities that justified exclusion, the selection was made arbitrarily.

2.5.5. Conversion factor

A common problem concerning NO_3^-, NH_4^+ and SO_4^{2-} is caused by the conversion factor. In some datasets factors were already included in base data. As the conversion is part of the calculation, this inclusion had to be reverted for the database. In view of the deviation plots and the actual data it was possible to uniformly exclude conversion to nitrogen or sulfur in the concentration values given in the database.

2.6. Database usage

2.6.1. Data import

Data import is realized with a simple Python script. New *ionenprüfen* files are added to a list of files that are concatenated with the main database repository. At the time of this work this repository is a CSV file including all samples before October 2017. The repository and the *ionenprüfen* files are not changed by the process. The complete database is newly formed out of all source files at every script execution. Therefore, at each run all source file changes are included in the newly compiled database.

In the future it may become beneficial to include a fixed part of the growing list of *ionenprüfen* files in a new main database repository file. Both to shorten the list and to be protected against file standard deprecation, that may be caused by future Microsoft Office updates. It has to be considered, that changes are only transferred to the database if they are carried out in one of the source files listed in the script. For example older *ionenprüfen* files that are already included in the repository will not be considered.

2.6.2. Flagsystem application

After the database creation it can be used for analysis. Regardless of the desired type of analysis some procedures are strongly recommended due to the level 1 character of the database. As already stated, level 1 data means that all measurement results, that are not clearly caused by instrument malfunction, are converted into physical units and are transferred in the database. However, the flag system introduced

in section 2.4 is used to mark contaminated samples within a review process. It is strongly recommended to exclude invalid samples from further calculations. One reasonable way to do this in Python is shown in listing 1. Please note that the pandas and numpy packages are necessary for this code to work and that the database is imported as db.

Listing 1: Python code snippet for flag usage

```
notflagged = ['Datum','Ort','Anmerkungen']
components = db.columns[~db.columns.isin(notflagged)][::2]
flagdict = dict(zip(components + '_flag', components))
for flag,col in flagdict.items():
    db.loc[db[flag].isin([5,6,7,8]),col] = np.nan
```

2.6.3. Aggregated concentrations and depositions

Many parts of the precipitation data analysis are performed on aggregated concentration or deposition values. For this purpose, the function in listing 2 was written to facilitate the aggregation process. It is based on the calculations that were already in place in the *ionenprüfen* files. However, it was custom-made for usage with the new database and it offers additional features. The function uses the `pandas.DataFrame` resample method and therefore all pandas DateOffset objects can be passed to it. This means that the aggregation period can be changed easily by setting the `freq` parameter to 'M' for monthly, 'Y' for annual, '1Q-DEC' for seasonal or 'AS-OCT' for period (Oct.-Sep.) aggregation. The `stations` parameter takes a list of station codes, which is useful if aggregated data from a limited number of stations is needed. The function returns two dataframes, one with concentration and one with deposition values.

Listing 2: Function for concentration and deposition aggregation

```
def aggr_db(freq='Y', stations=db.Ort.unique()):
    concentration, deposition = pd.DataFrame(), pd.DataFrame()
    for station in stations:
        data = db[db['Ort']==station]
        data.index = pd.DatetimeIndex(data.Datum)
        df, summe = pd.DataFrame(), pd.DataFrame()
        con, dep = pd.DataFrame(), pd.DataFrame()
        df['NS'] = data.NS
        df['H'] = 1000*10**(-data.pH)
        df['Hmass'] = df.NS*df.H
        summe['Hmass'] = df.Hmass.resample(freq).sum()
        df['NS_H'] = df.NS
        df.NS_H[pd.isna(df.H)] = 0
        summe['NS_H'] = df.NS_H.resample(freq).sum()
```

```
        summe['NS'] = data.NS.resample(freq).sum(min_count=1)
        summe['H_korr'] = summe.Hmass/summe.NS_H
        con['H'] = (summe.H_korr)
        dep['H'] = (summe.H_korr*summe.NS)/100
        summe.H_korr.replace(0, np.nan, inplace=True)
        con['pH'] = -np.log10(summe.H_korr/1000)
        dep['pH'] = con['pH']
        conv = {'Na':1,'NH4':(14/18),'K':1,'Ca':1,'Mg':1,
                'Cl':1,'NO3':(14/62),'SO4':(32/96),'Pb':1}
        convdict = {'NH4':'NH4N','NO3':'NO3N','SO4':'SO4S'}
        for ion in conv.keys():
            df[ion+'mass'] = df.NS*data[ion]
            summe[ion+'mass'] = df[ion+'mass'].resample(freq).sum()
            df['NS_'+ion] = df.NS
            df['NS_'+ion][(pd.isna(data[ion]))|(data[ion]==0)] = 0
            summe['NS_'+ion] = df['NS_'+ion].resample(freq).sum()
            con[ion] = (summe[ion+'mass']/summe['NS_'+ion])*conv[ion]
            dep[ion] = (con[ion]*summe.NS)/100
        con.rename(convdict, axis=1, inplace=True)
        dep.rename(convdict, axis=1, inplace=True)
        con['Nges'] = con.NO3N + con.NH4N
        dep['Nges'] = dep.NO3N + dep.NH4N
        con['Datum'], dep['Datum'] = con.index, con.index
        con['Ort'], dep['Ort'] = station, station
        con['NS'], dep['NS'] = summe.NS, summe.NS
        concentration = concentration.append(con)
        deposition = deposition.append(dep)
    return concentration,deposition
```

The way deposition is calculated is shown in equation 2. Like the concentration calculation, the deposition calculation has to be unaffected by removed contaminated measurements. Therefore, equation 2 starts similar to equation 1. The concentrations are multiplied by the precipitation amounts. Resulting masses are aggregated over the desired period and divided by the sum of precipitation amount. Once again the * denotes that contaminated samples are excluded. This leads to a corrected concentration that is multiplied by the original precipitation amount to obtain the aggregated deposition. As in equation 1, a conversion factor is used for NO_3^-, NH_4^+ and SO_4^{2-} ($\frac{14}{62}$, $\frac{14}{18}$, $\frac{32}{96}$) to convert ion depositions to nitrogen and sulfur depositions.

$$\frac{\sum\limits_{\text{period}} (\text{NS} \cdot \text{concentration}^*)}{\sum\limits_{\text{period}} \text{NS}^*} \cdot f_{\text{conversion}} \cdot \frac{\sum\limits_{\text{period}} \text{NS}}{100} = \text{deposition kg/ha} \qquad (2)$$

2.6.4. Missing data extrapolation

Occasional downtimes at certain stations can cause gaps in time series when no samples are present in a complete aggregation period. To distinguish these periods from times when actually no precipitation occurred, it is necessary that according precipitation amounts are included in the database. This can be achieved either by including data from manual gauges at the respective station or by using data from the Austrian hydrographic service.

These included values add to the overall sum of precipitation depth. As shown in equation 2 it is assumed that the non-measured precipitation corresponds to the weighted average concentration of the measured samples. This means deposition is automatically extrapolated to a certain degree. In the event of prolonged outages there is a risk of distortion because seasonality is not taken into account. Therefore, in these cases, extrapolation must be performed separately based on the same months of previous and following years.

Listing 3 presents an implementation that calculates precipitation-amount-weighted mean values based on the results of the same month of the two previous and the two following years. These mean values are then set for all samples in the concerning month. The flags of extrapolated values are set to 9 (see table 2.3). This way it is possible to calculate seasonally undistorted depositions after listing 2. If the concentrations are not to be affected by this extrapolation, they can be recalculated after a removal of all values flagged with 9, which is easily possible with a modified version of listing 1. `needcalc` represents a list of all station and month combinations that need extrapolation. Future elongated outages will be added to this list.

Listing 3: Missing value calculation

```
conc,dep=mitteln('M',orte)
def calc_missing(ort,monat):
    year=pd.to_datetime(monat, format='%Y-%m')
    oneyear=pd.offsets.DateOffset(years=1)
    years=[year - 2*oneyear+MonthEnd(1),year - 1*oneyear+MonthEnd(1),
           year + 1*oneyear+MonthEnd(1),year + 2*oneyear+MonthEnd(1)]
    environment=conc.loc[(conc.Ort==ort)&(conc.Datum.isin(years))]
    components=['H','pH','Na','NH4N','K','Ca','Mg','Cl','NO3N',
                'SO4S','Pb','Nges']
    inter=environment[components].mul(environment.NS, axis=0).sum()
    proj=inter/environment.NS.sum()
    proj['Datum']=year+MonthEnd(1)
    proj['Ort']=ort
    proj['NS']=conc.loc[(conc.Ort==ort)&
                (conc.Datum==(year+MonthEnd(1)))].NS.values[0]
    conversion={'NH4N':(14/18),'NO3N':(14/62),'SO4S':(32/96)}
    for ion in ['NH4N','NO3N','SO4S']:
        proj[ion]=proj[ion]/conversion[ion]
```

```
    proj.rename({'NH4N':'NH4','NO3N':'NO3','SO4S':'SO4'},
                axis=1,inplace=True)
    interdic=proj.to_dict()
    interdic.update({'NS_flag':9,'pH_flag':9,'Na_flag':9,'NH4_flag':9,
                     'K_flag':9,'Ca_flag':9,'Mg_flag':9,'Cl_flag':9,
                     'NO3_flag':9,'SO4_flag':9,'Pb_flag':9})
    row=db.loc[(db.Ort==ort)&(db.Datum>year-MonthEnd(1))&
                (db.Datum<=year+MonthEnd(1))]
    for key in row.columns[~row.columns.isin(['NS','Datum'])]:
        row[key] = interdic.get(key)
    row['Anmerkungen']='ber. '+str(list(environment.Datum.dt.year))
    return row

def insert_missing(ort,monat):
    year=pd.to_datetime(monat, format='%Y-%m')
    db.loc[(db.Ort==ort)&(db.Datum>year-MonthEnd(1))&
            (db.Datum<=year+MonthEnd(1))]=calc_missing(ort,monat)

needcalc=[['IV','1986-02'],['IV','1986-03'],['IV','1986-04'],
            ['SO','2014-01'],['WW','1983-10'],['WW','1983-11'],
            ['SO','2014-02'],['SO','2014-03'],['SO','2014-04'],
            ['SO','2014-05'],['SO','2014-06'],['OS','2013-12'],
            ['DR','2014-11'],['DR','2014-12'],['AF','2015-06'],
            ['AF','2015-07'],['GS','2013-11'],['GS','2013-12'],
            ['LI','2014-07'],['LI','2015-07']]
for ort,monat in needcalc:
    insert_missing(ort,monat)
```

2.6.5. Data export

The data processing or export script that was written in the course of this work includes all of the above mentioned code pieces as individual cells in a Jupyter Notebook. Furthermore, cells for data export or data plotting are included. The code blocks were developed with modularity in mind. Most of them provide options on how the analysis shall be performed. In most cases the queried time span, component or station can be set in a simple way. Thus even inexperienced Python users can create all graphs and data exports necessary for report creation.

3. Contamination detection by random forest classification

The newly formed database enables new possibilities like analysis on a per-sample basis. One example for this shall be given in this section. The goal is to create a classifier, which can detect contaminated precipitation samples within the database. Since October 1, 2014 all contaminated samples were manually identified by the same procedure under supervision of Anne Kasper-Giebl. Identification of invalid samples is based on ion as well as conductivity balances and a comparison with regional (e.g. long-range transport of mineral dust as identified via meteorological modelling) as well as local (e.g. information about construction activities, farming or other possible contaminations) conditions. Before 2014 the screening for contaminated samples was conducted by different personal (see appendix A.2). Although the basic considerations (e.g. ion and conductivity balances) most likely remained the same, different rules will have been applied as no defined settings for a mismatch of the respective balances are reported. The classifier should adjust the older part of the database to the newer and stricter testing regime and therefore answer if the observed trends are influenced by overlooked contaminated samples.

3.1. Classifier characterization

Decision trees classify samples by their properties, which in this case are measurement results of precipitation samples. The used classes are valid and contaminated. Figure 3.1 depicts the schematics of a decision tree as it was used in this work. A closer look on the first nodes of the tree is given in figure 3.2. It shows that every node separates the data by a specific criterion. The Classification and Regression Trees (CART) algorithm by Breiman et al. (1984) is used to create complex trees based on training datasets. For every node it selects the decision criterion which reduces Gini impurity the most. Gini impurity is defined as the possibility that a random sample is in the wrong class, when it is put in the majority class of a node. In the case at hand Gini impurity can be calculated according to equation 3.

$$\text{Gini Impurity} = 1 - \left(\frac{\text{valid samples}}{\text{all samples}}\right)^2 - \left(\frac{\text{cont. samples}}{\text{all samples}}\right)^2 \tag{3}$$

Random forest classification uses several decision trees based on different slices of the dataset. Final classification is usually made by a majority vote of the trees, although this parameter can be modified. The use of a forest over a single tree increases the performance of the classifier as overfitting issues are eliminated. (Breiman et al., 1984)

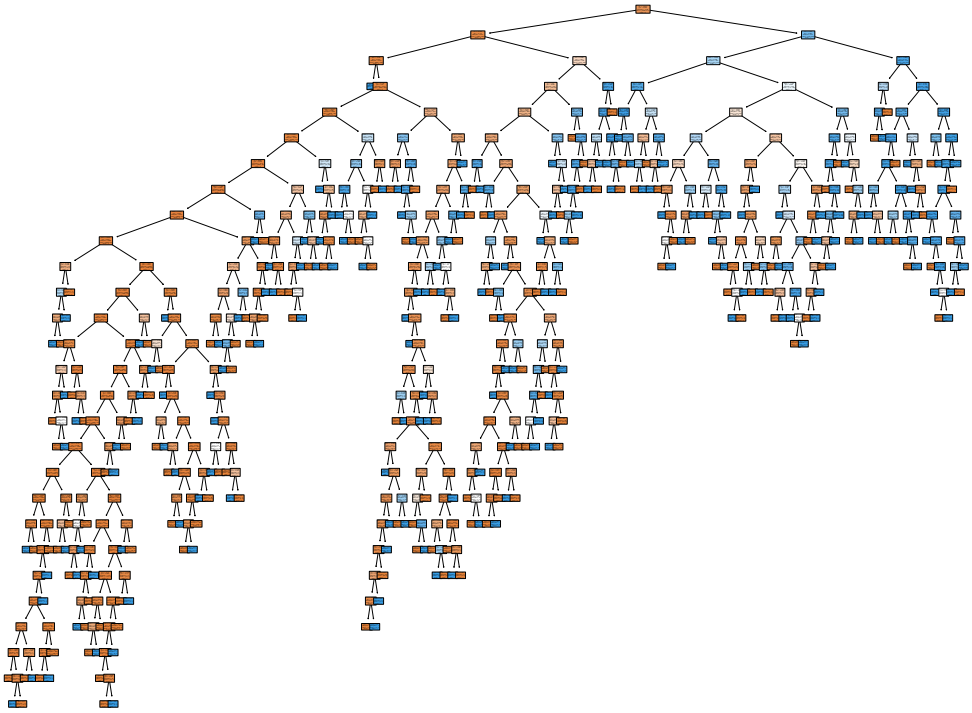


Figure 3.1: Scheme of a full decision tree

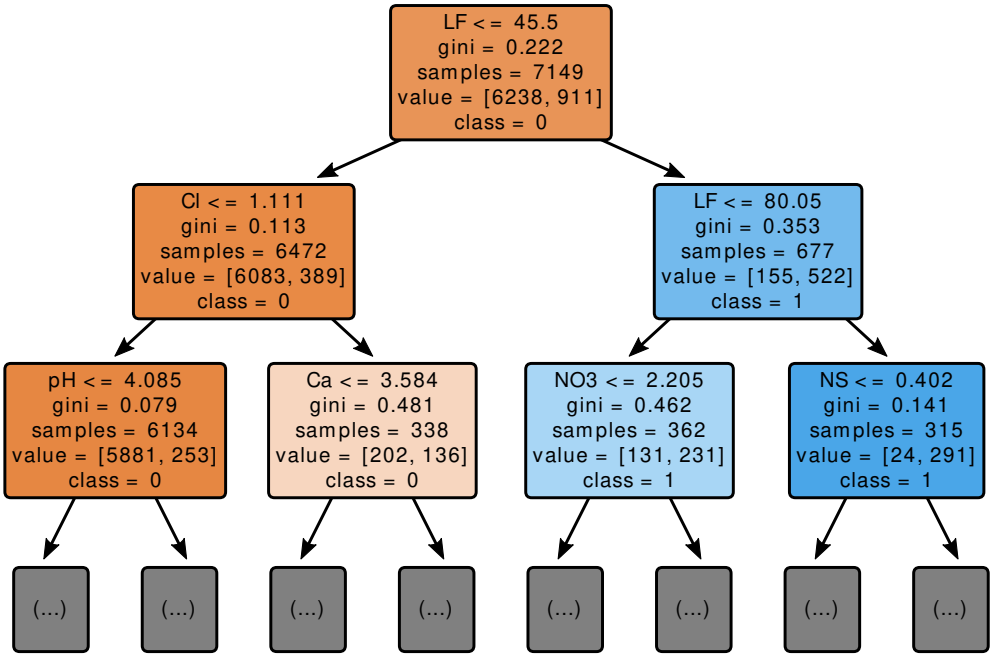


Figure 3.2: Exemplary first nodes of a decision tree

The data analysis in this section was performed with Python. Essential packages were pandas (McKinney, 2010) for data preprocessing, scikit-learn (Pedregosa et al., 2011) for all tasks concerning predictive analysis and matplotlib (Hunter, 2007) for visualization.

3.1.1. Selection of training and test data

At the time of this work the precipitation database includes almost 80 000 samples with 29 features (see table 2.2). Nevertheless, neither all samples nor all features can be used for the given task. Fully validated data for training and testing is available only after October 1, 2014 and only stations that have a sufficiently long record before and after this date can be used. Many stations in the database were only in operation for a couple of years and were shut down long before 2014 (see fig. 1.2). Therefore, only 14 stations given in table 3.1 were chosen out of a total of 46 stations (see tab. 2.1). This reduces the number of available samples to about 56 000. Furthermore, all incomplete records (e.g. with missing measurements) were excluded, reducing their count to 47 916. 39 397 of those are dated prior to October 1, 2014 leaving only 8519 complete and fully validated samples.

Table 3.1: Stations used for RF classification

ID	name	federal state	runtime
HF	Höfen	Tyrol	Nov 83 - Dec 20
NB	Niederndorferberg	Tyrol	Oct 83 - Dec 20
IV	Innervillgraten	Tyrol	Aug 84 - Dec 20
NH	Haunsberg	Salzburg	Oct 83 - Dec 20
WW	Werfenweng	Salzburg	Oct 83 - Sep 18
SO	Sonnblick	Salzburg	Oct 87 - Dec 20
LI	Litschau	Upper Austria	Oct 89 - Dec 20
LU	Lunz	Upper Austria	Apr 87 - Dec 20
OS	Ostrong	Upper Austria	Apr 91 - Dec 20
DR	Drasenhofen	Upper Austria	Oct 03 - Nov 17
MB	Masenberg	Styria	Mar 90 - Dec 20
HG	Hochgößnitz	Styria	Feb 90 - Dec 20
GS	Grundlsee	Styria	Feb 90 - Dec 20
AF	Arnfels	Styria	Oct 97 - Dec 20

In most cases contamination affects all components. However, there are exceptions, like elevated sodium and chlorid concentrations that can be caused by road salt. In these and other clear cases only affected ions are flagged accordingly. To simplify classification all flag columns were united to a single class column, which means that in above mentioned cases the whole sample was considered contaminated. The class is set to 1 (= contaminated) if at least one of the components is considered

contaminated. If all components are valid the class is set to 0. The column for text-based notes was discarded as well as the Pb^{2+} and Cd^{2+} concentrations, which are only available for a small amount of stations. Likewise date and station location were not used for the classifier creation, although there is a slight correlation between them and the target class. This decision was made to prevent a discrimination of samples taken in certain seasons and places. But the exclusion of the station location also eliminates the possibility to account for locations with overall higher concentrations. Therefore, the fact, that samples can be valid for one location, but most likely indicate a contamination for another location, cannot be taken into account. One way to circumvent this would be the creation of independent classifiers for every station but the database is not big enough to support this approach. Only around 2.5 % of the samples are usually marked as invalid and therefore some locations only have a very limited amount of contaminated data points.

First exploration showed that size and class composition in the training dataset are crucial for the properties of the resulting classifier. Classifiers trained on bigger datasets usually perform better on test data then classifiers trained on smaller datasets. In this application the ratio of valid and contaminated samples in the training set influences the ratio of the predicted classes. Above mentioned restrictions reduced the usable dataset to 8519 samples but only 194 of those were contaminated. For initial tests the dataset was randomly split (`random_state=42`) into quarters, three for the training set and one for the test set. After model training the created classifier was tested. The results of such testing are best to be shown in a confusion matrix, which is illustrated in table 3.2. A confusion matrix compares the original classes of the samples in the test dataset to the newly assigned classes. True positives (TP) and true negatives (TN) have been assign correctly. False positives (FP) are valid samples that were incorrectly classified as contaminated. Conversely, false negatives (FN) are contaminated samples that were considered valid by the classifier.

Table 3.2: Scheme of a confusion matrix

		validated	
		valid	contaminated
RF	valid	TP	FN
	contaminated	FP	TN

The random forest algorithm randomly separates the training dataset in in-bag (used) and out-of-bag (not used) samples for each tree of the forest. The confusion matrix on the left side of table 3.3 shows the results for one (`random_state=43`) of the many possible classifiers that are based on the aforementioned trainings set. This trainings set consists out of 6243 valid and 146 contaminated samples. As the training data represents three quarters of the total data, 2130 samples are available for testing.

Table 3.3: Confusion matrix for unbalanced (left) und balanced data (right)

		validated		
		valid	cont.	total
RF	valid	2066	5	2071
	cont.	39	20	59
	total	2105	25	2130

		validated		
		valid	cont.	total
RF	valid	50	7	57
	cont.	3	37	40
	total	53	44	97

At first sight it shows an exceptional accuracy of 98 %. But at second sight the number of predicted contaminations is more than twice as high (59) as the actual number of contaminated samples (25) in the test set. The high number of false positives certainly limits the usefulness of this classifier. This is especially true for this task because missing a small amount of true negatives in the border region, where the distinction is not clear, is not as important as avoiding a large amount of false positives. However, this classifier missed one fifth of the validated contaminations anyway.

Considering the poor results a new approach was taken. For that the amount of valid samples was balanced with the amount of contaminated samples in a random (`random_state=44`) downsampling process. It reduced the number of valid samples significantly to 194. The confusion matrix on the right side of table 3.3 shows the corresponding results. Although the matrix is now much more balanced, the percentage of overall false positives is actually higher than before. The first unbalanced classifier had a very high accuracy by putting most of the values in the valid class, which succeeded in many cases. This was not an option for the balanced classifier, which achieved an accuracy around 90 %. Therefore, further options were explored to improve the results.

One possibility was the inclusion of contaminated samples dated prior to October 1, 2014. This should be possible because the validation process applied today is considered to be stricter, which means all samples that before were considered contaminated are now also considered to be contaminated. This increases the number of available contaminated samples to 1208. Once again a test was performed with an unbalanced (valid: 8325, contaminated: 1208) and a balanced (valid: 1208, contaminated: 1208) dataset. The results are shown in the confusion matrices left and right in table 3.4.

The accuracy of the balanced classifier based on the extended dataset is almost exactly the same as with the one based on the smaller balanced dataset. This is a disappointing result considering the training set was six times larger than before. Nevertheless, the unbalanced classifier based on the extended database performed better than its smaller brother in several metrics like the true positive rate (see eq. 4 in section 3.1.2).

Table 3.4: Confusion matrix for unbalanced (left) and balanced extended datasets (right)

		validated		
		valid	cont.	total
RF	valid	2051	36	2087
	cont.	59	238	297
	Total	2110	274	2384

		validated		
		valid	cont.	total
RF	valid	265	37	302
	cont.	25	277	302
	total	290	314	604

3.1.2. Cross-validation

Section 3.1.1 discusses picked out results that heavily depend on random mechanisms (see random states). In this section a generalized approach is used to evaluate the up to now gained information. For that purpose a k-fold cross-validation is used to check how representative results in table 3.3 and 3.4 were. k-fold cross-validation means that the dataset is split into k parts. One part is used as test set whereas all others are used as training set. This procedure is repeated until every part was used for testing once. As this quickly gets computationally expensive ten k-values from 4 to 100 were tested. All datasets that were discussed in section 3.1.1 were reevaluated. They are summarized in table 3.5.

Table 3.5: Datasets for cross-validation

dataset	count valid	count contaminated	sum
unbalanced	8325	194	8519
balanced	194	194	388
extended unbalanced	8325	1321	9646
extended balanced	1321	1321	2642
upsampled balanced	8325	1321 ups. in training folds	

Due to the poor results in section 3.1.1 one additional approach was tested. Borderline-SMOTE is a Synthetic Minority Oversampling Technique (SMOTE) introduced by Han et al. (2005). It creates artificial new samples between nearest neighbors at the borderline of the two classes. Thereby the classes are evened out. To prevent training data leakage into test data, borderline-SMOTE is only used within training folds.

For the balanced datasets that were created by downsampling, the downsampling process was re-randomized for every k-value to account for instabilities caused by random exclusion and inclusion of important valid data points.

For every k-value k confusion matrices were calculated. For characterization purposes several confusion matrix related metrics were used. True positive rate (eq. 4), true negative rate (eq. 5), precision (eq. 6), negative predictive value (eq. 7), accuracy (eq. 8) and Matthews correlation coefficient (eq. 9) were calculated for every

pass according to Fawcett (2006). Matthews Correlation Coefficient (Matthews, 1975) was added because it is a great performance metric for unbalanced datasets (Boughorbel et al., 2017).

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (4) \qquad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (5) \qquad \mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (6)$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \quad (7) \qquad \mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (8)$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (9)$$

Mean and standard deviation were calculated for every dataset at each k-value. For better visibility the results are slightly shifted in figure 3.3 and 3.4. They

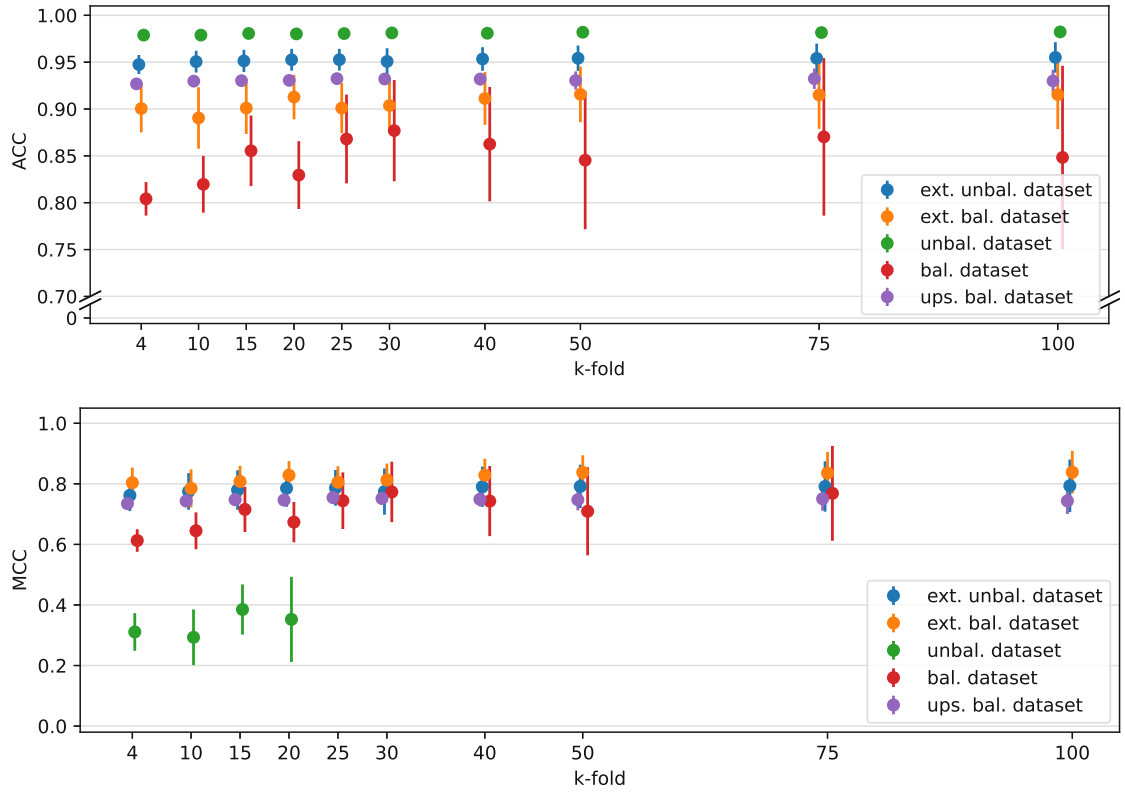


Figure 3.3: Mean and standard deviation of accuracy and MCC at different folds for different datasets

show that the qualitative information gathered in section 3.1 is correct. Unbalanced classifiers have higher accuracy but the balanced ones exceed in certain metrics like true negative rate or negative predictive value. In these two metrics the bigger

unbalanced classifier surpasses the performance of the smaller unbalanced one easily and is almost en par with the balanced classifiers. The upsampling approach with Borderline-SMOTE shows some interesting properties. It is more stable than the other balanced classifiers. It shows good performance in almost all metrics, especially in TNR and PPV. But regarding NPV it is almost as bad as the unbalanced small dataset. Which proves that further research in future work is needed to create the best possible classifier. Nevertheless, the classifier based on the extended unbalanced dataset seems to be the the most promising candidate for further testing although it is not the best in any metric.

Although some metrics show rather strong deviations figure 3.3 and 3.4 prove that the performance order of the different classifiers mostly stays the same as long as a k-value of at least 4 is used. Unbalanced classifiers show less deviation in accuracy, TPR and PPV, whereas balanced ones are more stable at TNR and NPV. Most standard deviations are increasing with the amount of folds used. This means that smaller test sets lead to stronger deviations although more values are used for the calculation. The mean metrics of the balanced classifiers are not as stable because of the randomized downsampling process, which introduces different valid samples at each k-value. Upsampling leads to very stable results across all metrics.

The fast data exploration approach in section 3.1.1 led to correct assumptions but only a k-fold cross-validation with a k-value of 4 or more enables a quantitative assessment of the different classifiers. It also shows that using bigger test datasets and therefore less training data hurts the classifier performance. This is noticeable for the lower k-values like k = 4, which means that only three quarters of the data are used for training. On the other hand increasing the k-value above 25 is computationally expensive and does not add much information.

To sum up the classifier based on the extended unbalanced dataset is most promising for the given task. Although it is weaker in some metrics it clearly outperforms the classifiers based on the smaller dataset and has the benefit of predicting a smaller and therefore more realistic number of contaminations.

3.1.3. Classifier parameters

When the above mentioned classifiers are applied to the not fully validated target dataset, they predict vastly differing amounts of contaminated samples. The balanced classifiers mark around one third of the old samples as contaminated, whereas the unbalanced ones found 1 % (small set) and 14 % (extended set). All but the small unbalanced classifier found more contaminated samples than expected as in recent years only around 2.5 % of the data were flagged as contaminated. The result further proves the strong correlation between class distribution in the training data

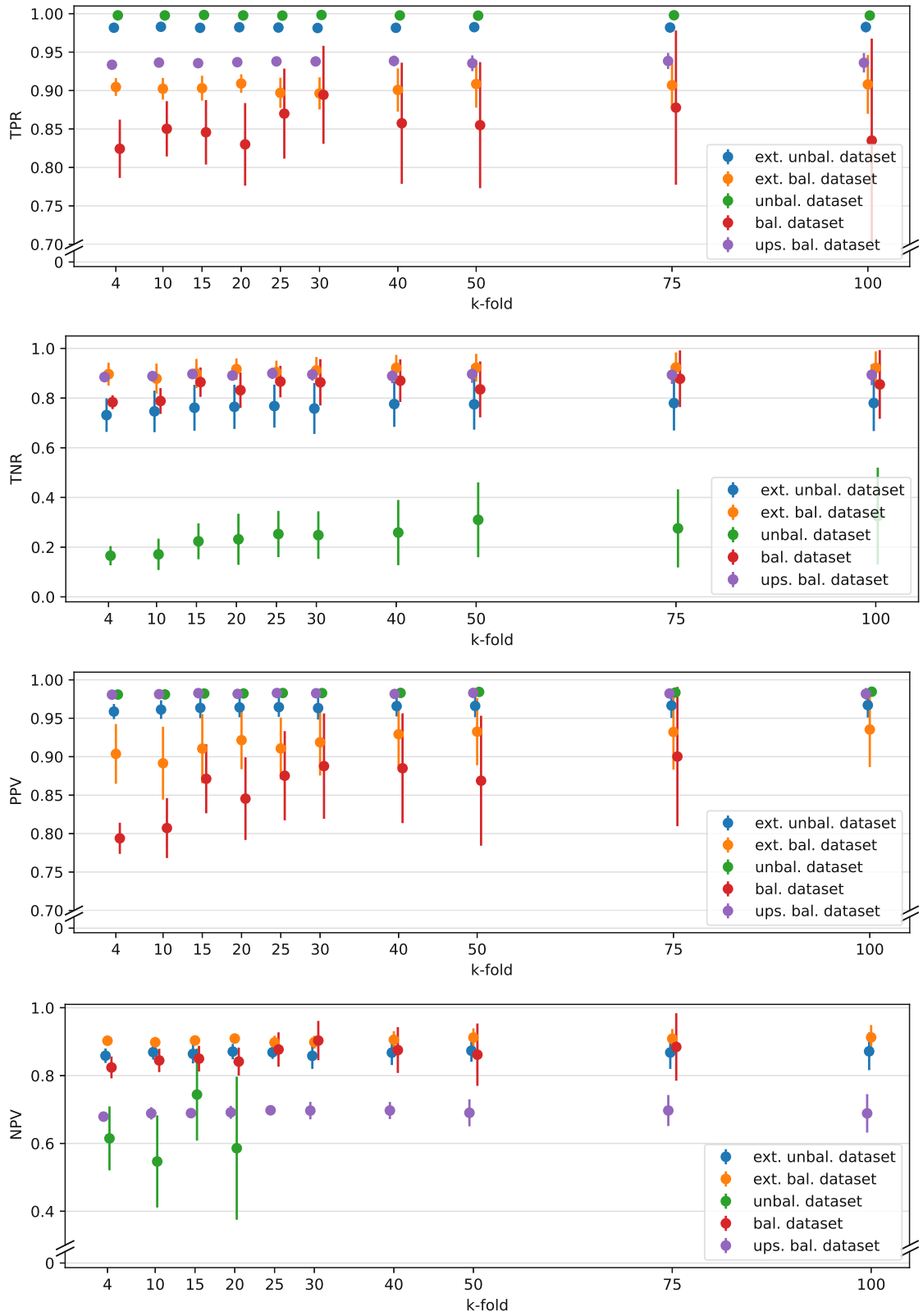


Figure 3.4: Mean and standard deviation of true positive rate, true negative rate, precision and negative predictive value at different folds

and predicted class distribution. Sections 3.1.1 and 3.1.2 showed that the small unbalanced classifier is not very good at finding contaminated samples, plus there is a good chance that the amount of contaminated samples is actually higher in the older part of the database, because of improving sampling routines. However, the other classifiers find more contaminations than expected. Therefore, a parameter to reduce the amount of found contaminations would be beneficial.

One simple possibility to do this is to adjust the majority vote of the random forest. By default more than 50 % of the trees have to vote for a contamination to achieve this classification. To change this behavior the random forest implementation of scikit-learn has the option to calculate the class probability instead of the plain result of the majority vote. When all ending nodes (leaves) are pure (= contain only samples of one class) the class probability is equal to the percentage of trees that voted for that class. Using these values the requirement can be changed and therefore the amount of found contaminated samples can be reduced.

Another important parameter for random forests is the used tree count. This parameter has no direct influence on aforementioned topics but it is important for prediction quality. Higher counts improve the classifier but are computationally more expensive. This is not as important for the relatively small dataset in this work because high tree numbers can be trained without any problems. But at a certain point no additional benefits can be gained and therefore the determination of a useable value is reasonable. To do this the out-of-bag error rate was calculated for several tree counts. The out-of-bag values can be used to test decision trees because they are per definition not used for their training. In contrast to cross-validation no separate test dataset is needed as every tree is tested with samples that were not used for its individual training. After that a majority voting is held and compared with the true class. Figure 3.5 shows the OOB error rate based on the unbalanced

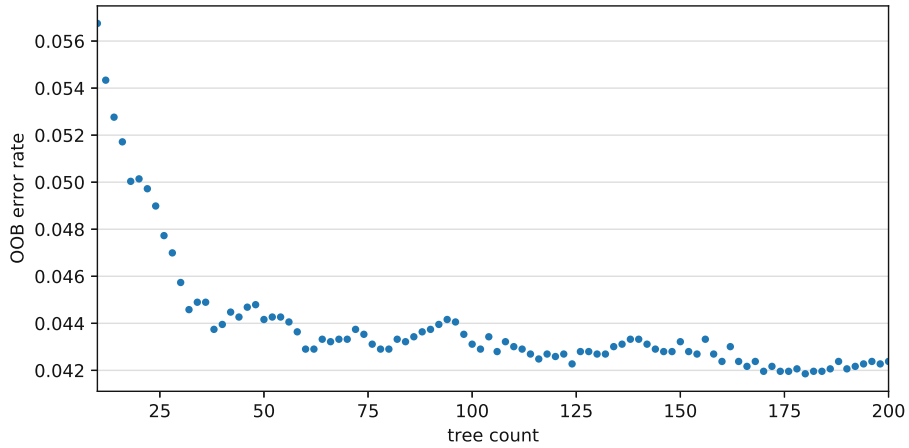


Figure 3.5: Out-of-bag error rate at different tree counts

extended dataset, which as one of the largest datasets profited more from higher tree counts. After a rapid decrease the error rate falls only slightly after 40 trees. To be safe 100 trees were used for all calculations.

Along the tree count the maximal branch length in all trees can be chosen. This can be done by setting a minimal number of samples in a node to be a leaf node. By default this is set to 1, which means that every branch is fully differentiated. Additionally this ensures that every leaf is pure, which naturally means that massive overfitting is applied. Nevertheless, overfitting should not be a problem as the trees in a random forest have different in-bag and out-of-bag datasets. However, to avoid underfitting the parameter was left at default.

3.1.4. Feature importance

There are several methods to determine the importance of every used feature for the classifier. Gini importance is calculated as the normalized total reduction in Gini impurity, that is caused by one feature. Features with higher Gini importance are more important for the classifier. Gini importance, which is also referred to as Mean Decrease in Impurity (MDI), is shown in figure 3.6. (Pedregosa et al., 2011)

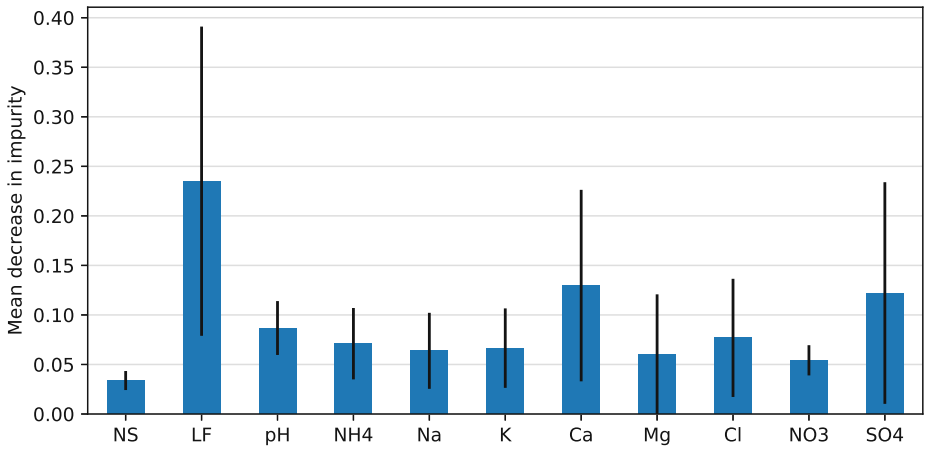


Figure 3.6: Gini importance

Data with hight cardinality (many unique values) can lead to misleading results in MDI. Subsequently, permutation importance was calculated and plotted in figure 3.7. Permutation importance uses a test set to determine the impact of the removal of each feature on classification. (Pedregosa et al., 2011)

For both methods calculations were based on the unbalanced extended dataset. They show that conductivity, calcium and sulfate concentrations are important for

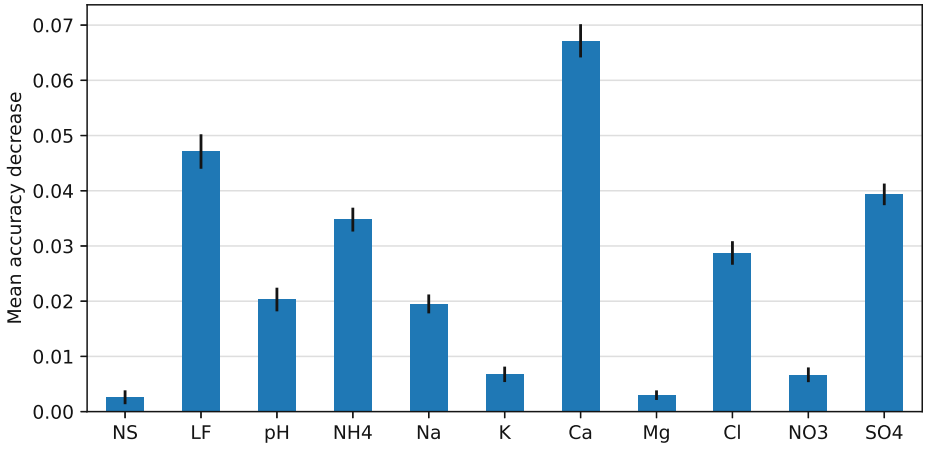


Figure 3.7: Permutation importance

the classifier. The high importance of conductivity was expected, but the low importance of precipitation amount is interesting. Usually samples of smaller precipitation events show higher concentrations. This ion enrichment is expected and the manual validation accounts for it by allowing higher values before flagging them as contaminated. For the classifier however the precipitation amount contributes very little.

3.2. Results

Finally the classifier based on the unbalanced extended dataset, that was extensively characterized in section 3.1, was applied to all samples. The period after 2014 that was used for training was included as a sanity check. Naturally the classifier did not change much on that part of data. No change was observed in periods where not all components were measured because classification could not be performed on these samples. The classifier was applied twice with two different majority vote settings. Once with the default 50 % limit (called RF50) and once with a more strict 90 % limit (RF90). The first one marked 12 % of all samples as invalid, while the second one marked only 2.5 %. The true amount of contaminated samples lies probably somewhere in between. It has to be noted that changing the majority vote limit, also changes the classifier characteristics that were investigated in section 3.1. Technically the results were temporarily integrated in the database as two new flag columns that state if a sample was considered contaminated by one or both of the classifiers. Naturally all invalid samples marked by RF90 are also marked by RF50. Due to the fact that all contaminated samples were used for training, all of them were marked by RF50. This is however not completely true for RF90

because 0.7 % of the manually marked contaminated samples were considered valid by RF90.

A common way to show trends in precipitation data is to calculate annual depositions as shown in equation 2 on page 16. The depositions are then plotted against time as points or in this work as columns. As shown by Schreiner (2017) the robust Theil-Sen Regression (95 % confidence) is well suited to uncover trends in precipitation data. In this work the implementation by Virtanen et al. (2020) after Sen (1968) and Conover (1980) was used. In contrast to the usual procedure, statistical significance of trends was not verified by Mann-Kendall testing. The reason for this is that the focus of this chapter is to highlight the impact of random forest classification, rather than analyzing the trends themselves.

Another way to present precipitation data is the Miles and Yost diagram, which plots ion balances against conductivity balances for individual samples (Miles and Yost, 1982). The diagram separates the samples in four sections, which are explained in figure 3.8. It is usually not used for data reporting but it is a valuable tool for the manual data validation process. The implementation in this work is based on the thesis by Firmkranz (2019). Samples are plotted in yellow if RF50 classified them as contaminated but they appear orange if they are also found by RF90. Manually flagged contaminations are displayed in red regardless of their random forest classification.

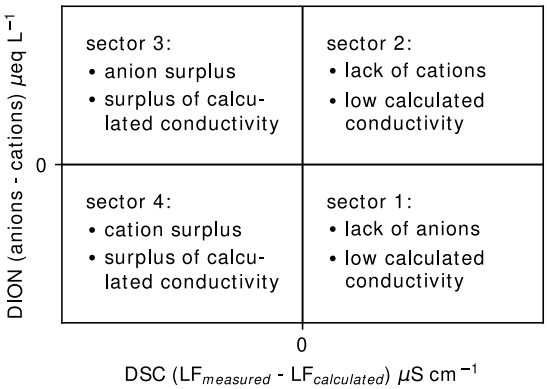


Figure 3.8: Scheme of a Miles and Yost diagram; after Firmkranz (2019)

Because 14 stations and 9 components were analyzed, a total of 126 time series plots were made. Not all of them can be discussed in detail. Therefore, one plot for each component was chosen based on the presence of interesting features and station diversity. A full set of graphs is given in appendix B. Miles and Yost diagrams are used to better understand which samples were marked by the classifiers. For ammonium, nitrate and sulfate deposition, data from the EMEP MSC-W model by Simpson et al. (2012) is given to provide context.

Figure 3.9, which is the first one to be described, is different from the other examples. The y-axis is no deposition value, but refers to the average H^+ concentration, which is given as pH value. Apart from that, this plot shows minimal change caused by random forest classification. The extent of change observed in Höfen is average and about the same as for the other stations. Most visible differences are between 1988 and 1996. The classifier predominantly marked samples with below average pH as contaminated, leading to higher average pH values. Although this is mostly also true for the other stations, there are only few exceptions within the whole database, where the identifications via the random forest classification leads to lower average pH values. In all of the Höfen station data only 13 samples were marked as invalid by more than 90 % of the trees (RF90). Almost all of those samples are centered between the third and the fourth sector of the Miles and Yost diagram (see left side of fig. 3.10), which means that the measured ion concentration is higher than expected. For most of the identified samples no distinct mismatch between anions

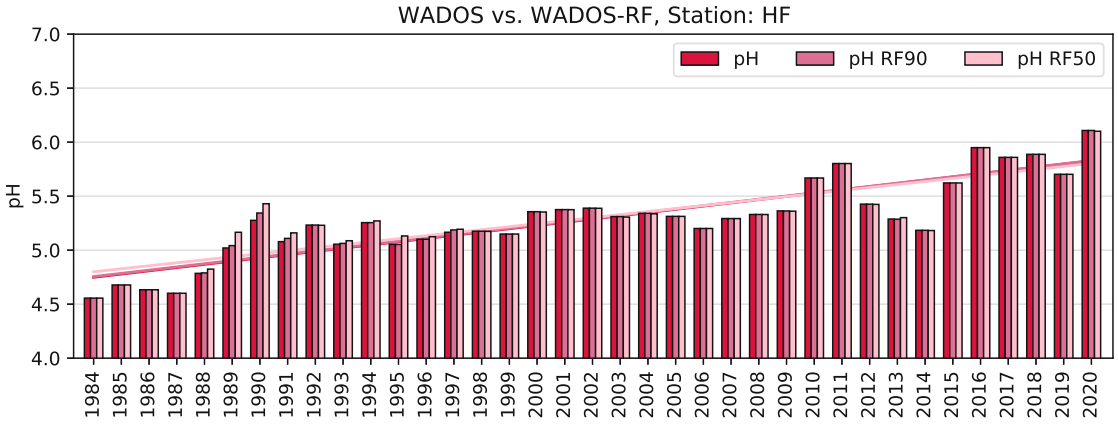


Figure 3.9: pH values based on original and RF adjusted data in Höfen

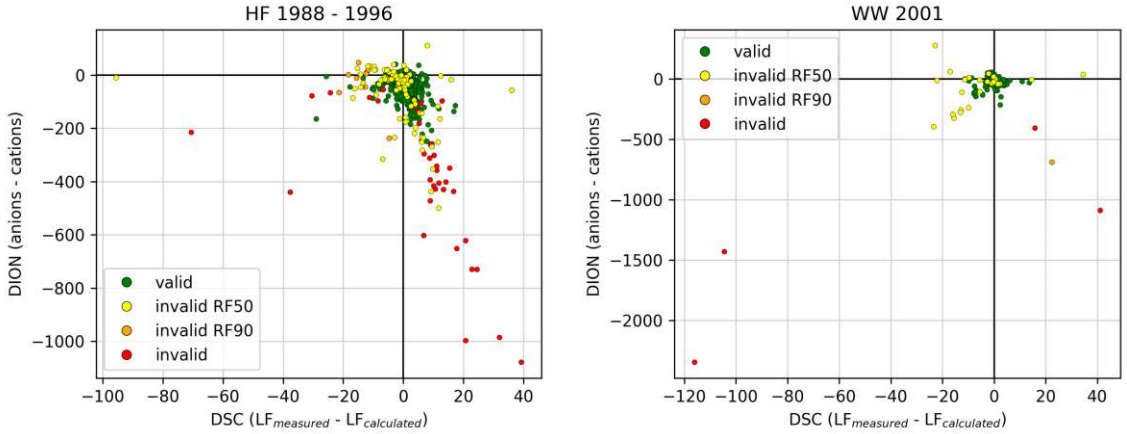


Figure 3.10: Miles and Yost diagrams for Höfen (left) and Werfenweng (right)

and cations was found. In two cases a marked excess of cations was found. Excluding these samples would lead to higher pH values. Still, the evaluation of the impact of single data points was not the focus of this work and needs to be addressed in subsequent work. If the usual 50 % or more are declared as the majority of the trees (RF50), 108 samples are marked as invalid. These are distributed across the whole data set given in the Miles and Yost diagram (fig. 3.10). It must be noted that both given numbers are contaminations that were additionally found by the classifiers. That is that the manually marked samples, that were part of the trainings dataset, were mostly classified correctly. More precisely, all of the invalid samples in Höfen were also marked as invalid by RF50 and more than 97 % of those were classified as invalid by RF90. All things considered, overlooked contaminations have probably no impact on the pH trend in Höfen. However, in the first four years magnesium and potassium ions were not measured and therefore the classifier could not be applied in those years.

Figure 3.11 shows the sodium deposition in Werfenweng. This time series was chosen because, like at most other stations, it features more or less pronounced outliers. It has at least one clear outlier in 2001 and a longer period of rising and then falling depositions, peaking in 2012. The elongated period was left almost untouched by the classifier regardless of the set majority vote limit. In 2001 RF90 found one

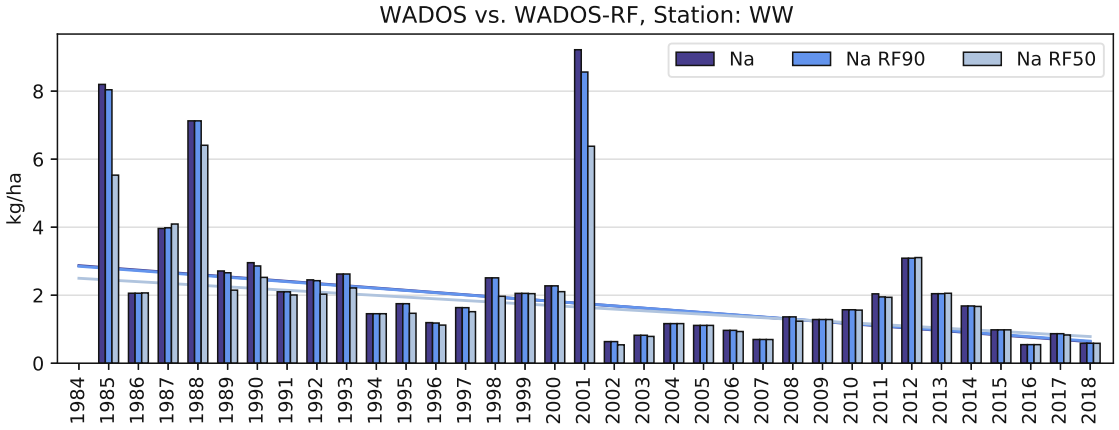


Figure 3.11: Sodium deposition based on original and RF adjusted data in Werfenweng

sample that, due to its position in the first sector of the Miles and Yost diagram (right side of fig. 3.10), next to two contaminated samples, can also be considered as contaminated. By using the stricter classifier the sodium deposition in 2001 is further reduced. However, a look in the Miles and Yost diagram shows that the yellow samples, which were excluded by RF50, are scattered in and around the center. Still, the majority of the samples gives a negative conductivity balance, indicating that either the measured conductivity is incorrect and too low, or the analysis of several or some ions is wrong, i.e. too high. As mentioned before, the

evaluation of single samples is beyond the scope of this work and has to be carried out independently. Most probably for the majority of the samples analyses would be repeated according to the current procedures. Still, 2001 remains an outlier. This is to say that the boundary to the green valid samples is fluid and probably many of these yellow samples would not have been marked in a manual review. Therefore, 2001 is an outlier regardless of the validation process.

An even greater disparity is found in the data of station Haunsberg. Figure 3.12 shows the chlorid time series, which features one period (1994 - 1996) with elevated depositions that are only coupled with calcium depositions (see appendix B.4). This time RF90 as well as RF50 drastically reduce depositions in that period, although the period stands out even after the removal. The according Miles and Yost diagram (left side of fig. 3.13) shows that the classified samples scatter over all sectors but

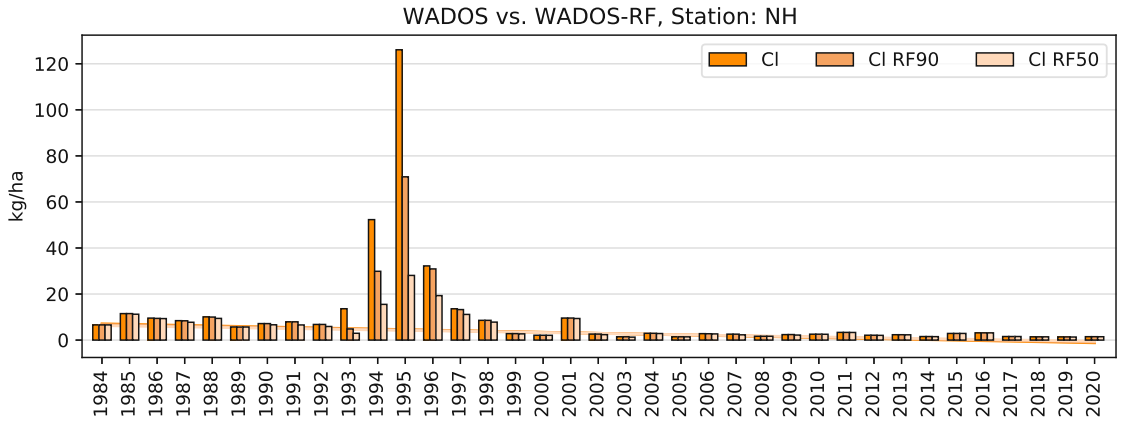


Figure 3.12: Chlorid depositions based on original and RF adjusted data at Haunsberg

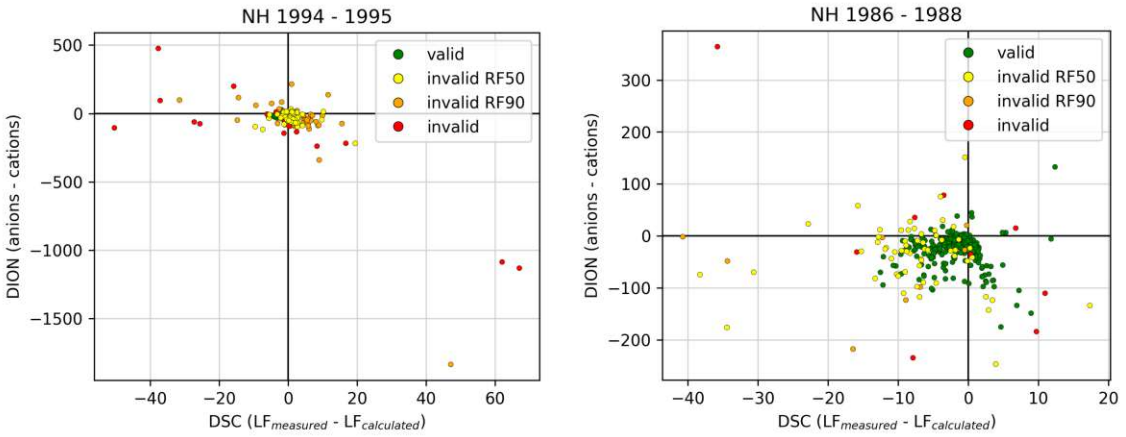


Figure 3.13: Miles and Yost diagrams for Haunsberg in 1994 - 1995 (left) and in 1986 - 1988 (right)

most of them are in the first and third sector, which means that they either have a surplus or a lack of anions.

To illustrate the impact of this elevated chlorid and calcium deposition period on other components, the ammonium depositions in Haunsberg are given in figure 3.14. However, the impact is not as high as anticipated, especially not for the period of 1994 to 1996. For the ammonium deposition, reference values from the EMEP MSC-W model are available. They are not measured but calculated for a $0.1° \times 0.1°$ longitude-latitude grid, which roughly translates to 7.5×11.1 km patches in Austria. Therefore, they cannot be considered equal to the measurements that are carried out at the WADOS station locations under specific local conditions. Nevertheless, the comparison of modeled and measured data is a starting point for further evaluations and allows, for example, to identify whether a trend measured at the sampling site is a local or regional phenomenon. In case of figure 3.14 the MSC-W model gives considerably higher depositions. Still, it becomes visible that neither the measurements nor the model gives an increasing or decreasing trend of the deposition loads for the last 20 years, when this comparison is possible. Independent of the comparison between model and measurement the original data set was compared to the results obtained with the RF classifiers. RF90 detected 97 additional contaminations in the dataset. Many of these appear during the first decade of operation and slightly flatten the decreasing trend in ammonium deposition. This flattening is even stronger when additionally 386 samples, marked by RF50, are excluded. To investigate this period in more detail, the Miles and Yost diagram for the samples between 1986 and 1988 (right side of fig. 3.13) was created. The classified samples scatter over all except the second sector, which indicates that the classifications are probably not traceable to a single cause. The reductions in the first decade naturally lead to better alignment with the modeled slope, which lacks data in that period. Nevertheless, this would mean that almost 20 % of all samples are contaminated, which needs to be evaluated in more detail in future work.

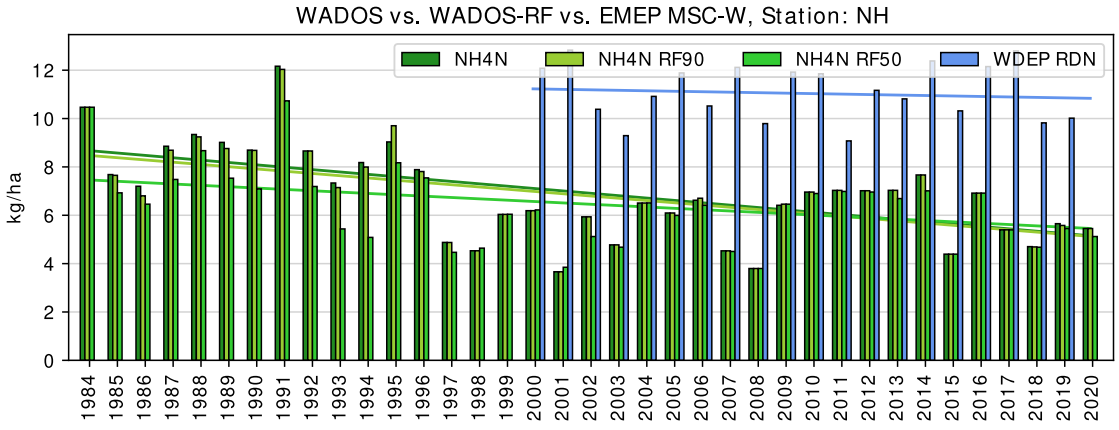


Figure 3.14: Reduced nitrogen deposition based on original, RF adjusted and modeled data at Haunsberg

Data of station Haunsberg is used one more time in figure 3.15, which depicts sulfur depositions. Once again Haunsberg was chosen to check the random forest impact caused by an elevated chloride and calcium period on other components but also to enable a comparison with the previously given ammonium depositions. Again, the exclusions around 1995 do not impact the time series as much as expected. The changes in trends are comparable to the ones observed in figure 3.14. The

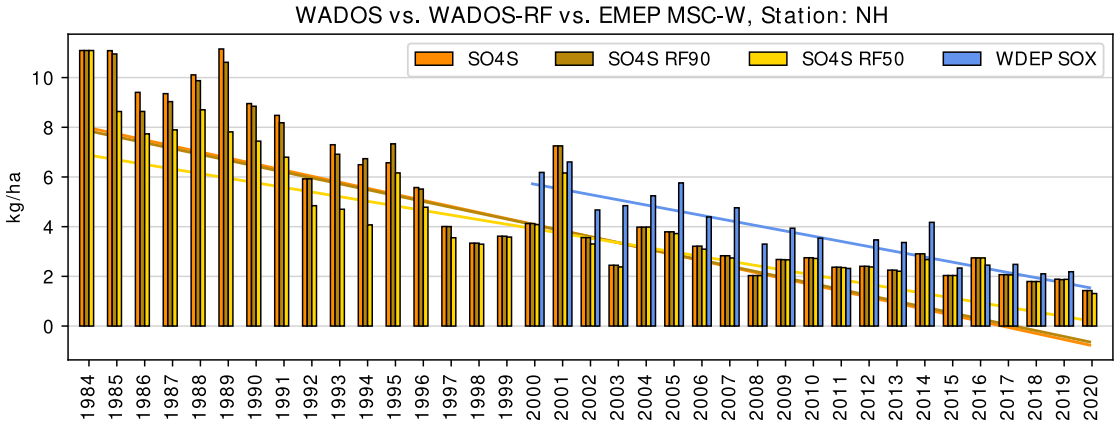


Figure 3.15: Sulfur depositions based on original, RF adjusted and modeled data at Haunsberg

modeled sulfur depositions (WDEP SOX) in figure 3.15 are slightly higher than the measured depositions. The situation corresponds qualitatively to the results of the ammonium example, although all observed slopes are steeper for sulfur. At Haunsberg 2.4 % of the data were manually classified as invalid. When RF90 is used the share of excluded samples increases to 4.8 %. Moreover, RF50 increases the proportion to 14 %. But it becomes apparent that even a sizeable number of possible contaminations has a smaller impact than changing the range selection of the time series. For example removing the first two years results in roughly the same slope than using RF50 on the full dataset. The graph shows that, due to a flattening of the sulfur deposition decrease, the linear model is not ideal. The reduction in sulfur deposition has slowed down compared to the last century, which results in the fact that the fit suggests negative values by now. This topic will receive further discussion in section 4.3.

Figure 3.16 shows the annual potassium depositions in Litschau. This dataset was chosen because it features several outliers before as well as after 2014. Under those 2008 clearly stands out as RF90 increases the deposition whereas RF50 almost bisects the value. A look in the Miles and Yost diagram on the left side of figure 3.17 reveals that this interesting situation is caused by samples with a negative conductivity balance and a negative ion balance pointing to an excess of cations. As the data points are closely following a line, the respective ion can be determined. Almost all samples taken in July and September 2008 were manually marked as invalid (red).

But four samples (yellow) in October still show very high potassium concentrations. It is interesting that these were exclusively detected by RF50.

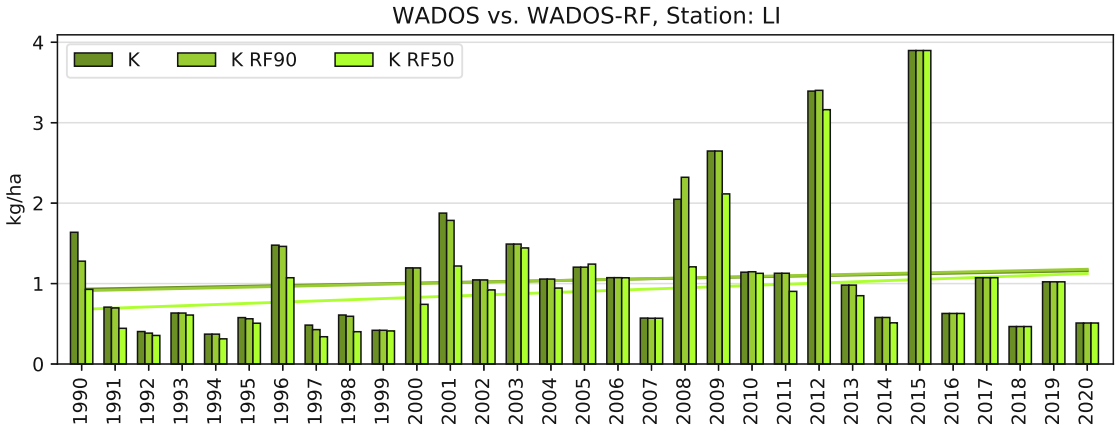


Figure 3.16: Potassium deposition based on original and RF adjusted data in Litschau

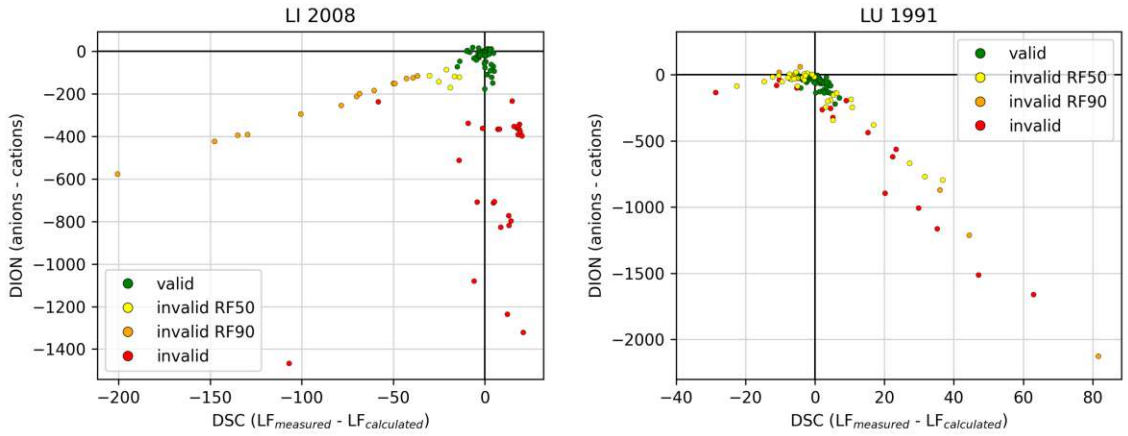


Figure 3.17: Miles and Yost diagrams for Litschau (left) and Lunz (right)

Figure 3.18 presents the nitrogen depositions introduced by nitrate in Lunz. Oxidized nitrogen depositions (WDEP OXN) from the EMEP MSC-W model are given for reference. This time the values match more closely. Nevertheless, the observed trends differ. In this dataset the manual validation marked 1.6 % of the samples as contaminated. The use of RF90 increases this to 2.6 % and RF50 to almost 10 %. The latter one leads to a much flatter trend, which differs even more from the EMEP reference trend. Many of the classified samples are far out in the first sector of the Miles and Yost diagram. As an example the diagram is plotted for all samples of the year 1991 on the right side of figure 3.17.

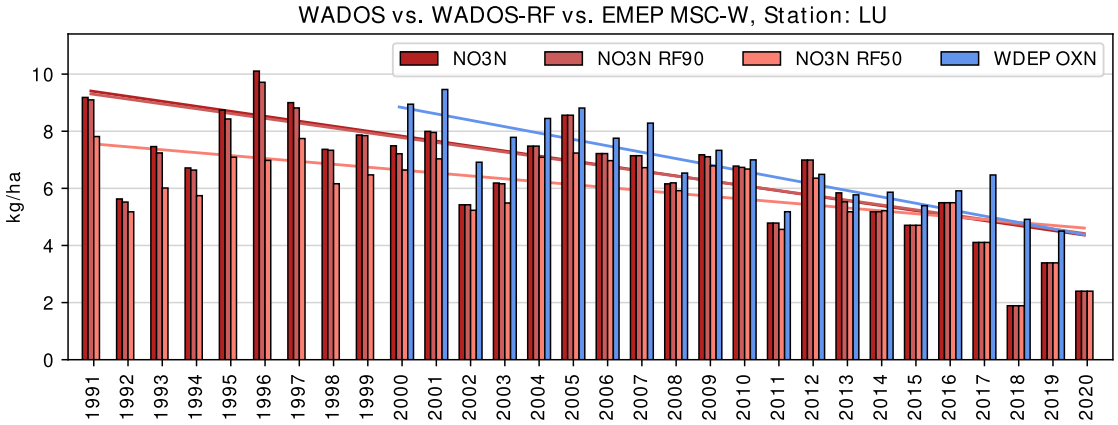


Figure 3.18: Oxidized nitrogen depositions based on original, RF adjusted and modeled data in Lunz

Figure 3.19 depicts the calcium deposition in Arnfels. Like figure 3.12 it features a single elevated period, which abruptly starts in 2002 and then declines over the next two years. In 2002 manual validation marked no samples as contaminated, whereas RF90 marked 58 % and RF50 89 % of the data. Even after this drastic reduction the elevated period is still visible. Due to their large number the marked samples are distributed across all sectors of the Miles and Yost diagram (left side of fig. 3.20). However, most of the data and the biggest outliers are situated in the first and fourth sector, which means that either a lack of anions or a surplus of cations is present. Since it is unlikely that the measured conductivity is wrong by 20 $\mu S\,cm^{-1}$, a cations surplus is more probable.

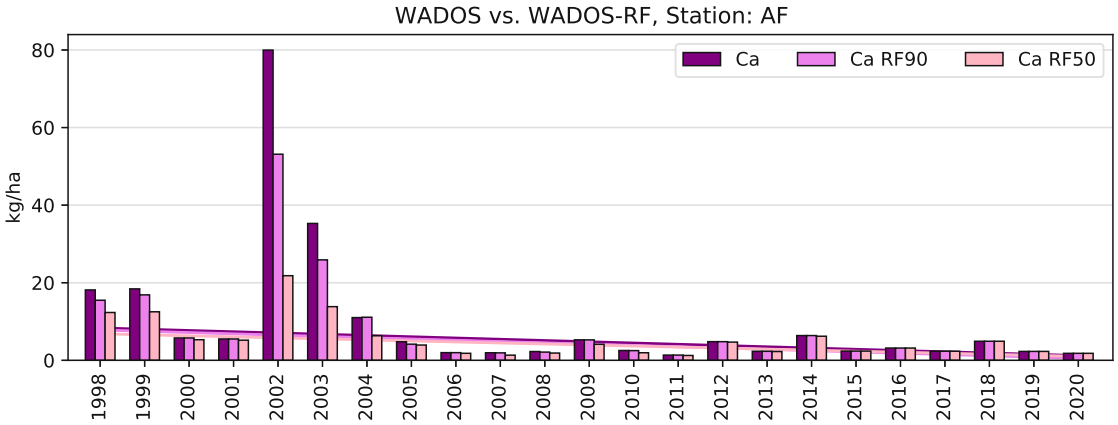


Figure 3.19: Calcium depositions based on original and RF adjusted data in Arnfels

The magnesium deposition in Innervillgraten is given in figure 3.21. This series was chosen because there is no distinct increasing or decreasing trend visible. However, the mean deposition changes noticeably after the random forest classification. The

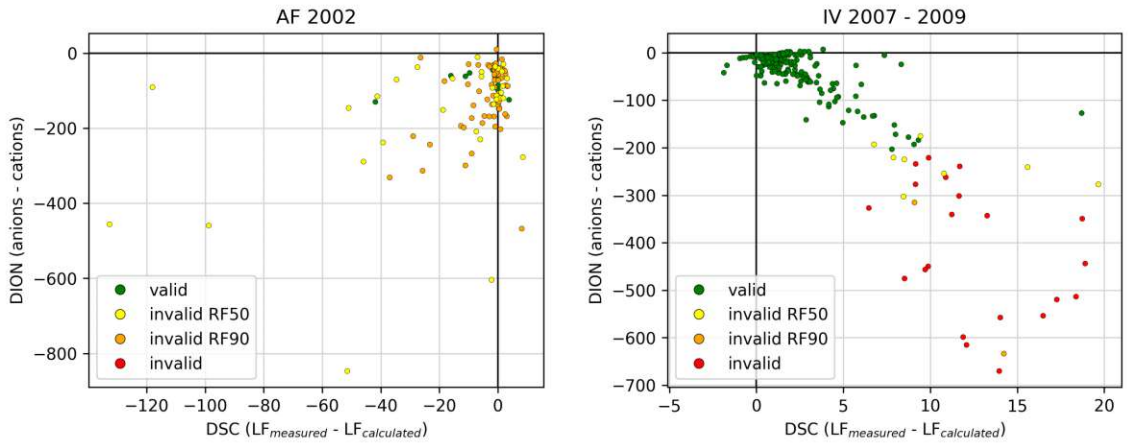


Figure 3.20: Miles and Yost diagrams for Arnfels (left) and Innervillgraten (right)

main part of this is caused by the reductions between 2007 and 2009. The right side of figure 3.20 shows that samples that are situated in the first sector of the Miles and Yost diagram are the reason for this. Although 23 samples in this period were already marked by the manual validation, RF90 added two samples that clearly have a high impact on the magnesium deposition. The stricter RF50 added eight more samples to the class of invalid samples, further decreasing the depositions. These samples, together with the red dots already marked as invalid in the original data set, show the characteristic features of an influence of mineral dust. Further research is needed to determine whether this dust has local (construction work or gravel applied during the winter period) or regional origin. The classifiers give the necessary tools to identify periods of interest.

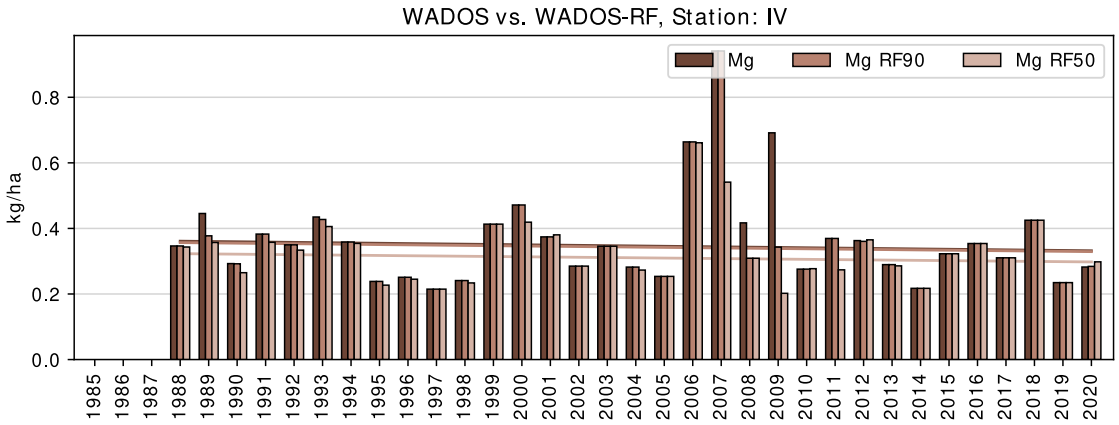


Figure 3.21: Magnesium depositions based on original and RF adjusted data in Innervillgraten

3.3. Summary

Section 3.1 showed that using all contaminated samples before 2014 in addition to all samples after 2014 as training data results in a compromise on several metrics. A classifier (RF50) based on that data classifies 12 % of the target data as contaminated, which is probably above the number that would be marked if a manual validation would be carried out. However, this can be adjusted by fine tuning the majority vote of the random forest (RF90). Therefore, two classifications were used to mark out an upper and a lower limit on how overlooked contaminated samples could influence the data. For the upper limit 50 % and for the lower limit 90 % of the trees have to agree on a contamination of a sample. The lower limit leads to 2.5 % of the samples being classified, which matches the amount that is usually found by manual validation.

In section 3.2 the impact of these two classifications on the data was shown by plotting annual depositions and trends. Outstanding features in the results were highlighted and investigated with Miles and Yost diagrams. Thereby varying results were revealed for different stations. The classifiers removed several isolated contaminations as shown in figure 3.21. Longer periods of elevated depositions as in figure 3.12 and 3.19 were flattened although not completely removed. The data of some stations like Höfen were changed only slightly. But for several stations like Haunsberg or Lunz at least the stricter classifier caused a noticeable flattening in trends because many classified samples are dated at the beginning of decreasing time series. Although in no case a trend reversal was observed. The strict classifier identified 12 % of all samples as contaminated. This can certainly be considered as a worst case and therefore the question raised at the beginning of section 3 can be answered: Overlooked contaminated samples allow only minor trend changes in some stations and can certainly not cause trend reversals.

Random forest classification is a promising tool to identify irregularities in data. But because the border between contaminated and valid samples is fluid and the manual validation process is subject to fluctuations, a completely accurate classification is not possible. Figure 3.17 and 3.13 show the capabilities and problems of the random forest in a condensed way. For Litschau the classifiers precisely marked some clearly outstanding samples. Whereas for Haunsberg a multitude of samples were found that can hardly be separated from valid samples. Many of those would probably not be marked in a manual validation. Therefore, the final decision and the fine tuning for individual stations still have to remain in human responsibility.

Future work may include the implementation of random forest classification in the data review process to highlight samples suspicious of contamination. At the moment the workflow is not yet ready for unsupervised routine usage. However, some flags were already set based on information gained through the classifier. Further flagging adds to the training data set, which may improve the performance of the

classifier. In addition further research is needed on data upsampling to evaluate its full potential and to understand its impact on classification.

4. Trend analysis

The new database not only enables per-sample methods like the random forest classification in section 3, it also facilitates analysis on aggregated data. The unification of data from all stations over the complete time series greatly accelerates the data retrieval process for all kinds of analysis. The following sections present illustrative analysis examples that profit heavily from the taken approach.

4.1. Seasonal variance

As single extreme weather events can easily dominate single-year datasets, weak seasonal trends can only be uncovered with data from long-term measurement campaigns. Therefore, the database containing the complete record of the Austrian Precipitation Sampling Network data is an excellent source for seasonal trend analysis.

Schreiner (2017) investigates seasonal trends based on fourier analysis because aforementioned extreme events have a disproportionate effect on monthly mean values. In this work seasonal trends are identified by plotting time series on top of each other for every individual location and species. This way the impact of outliers on the mean value can be assessed. A complete collection of all these plots is given in appendix C. The plots are based on monthly aggregated values because daily values or weekly aggregation would create too much noise due to the limited number of precipitation events. In this section some remarkable examples are highlighted and a summary on seasonal trends will be given in comparison with Schreiners work to show the way for future work.

The reason why these plots are created separately for all stations lies in different local conditions. For example all stations exhibit more or less pronounced seasonal precipitation amount trends (see appendix C.1). The seasonal precipitation amounts for Werfenweng and Sonnblick are highlighted in figure 4.1. Although these stations are separated by only 50 km, both show very different trends. Werfenweng like

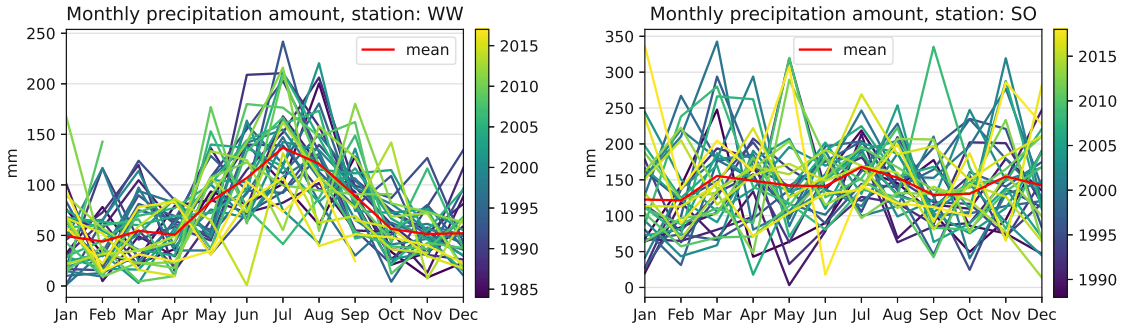


Figure 4.1: Monthly precipitation amounts in Werfenweng and at Sonnblick

most stations registers more precipitation in summer, whereas the station at mount Sonnblick does not follow any obvious trend. This can be explained by its high-altitude position at 3106 m.a.s.l. and its exposed location that allows unrestricted incident air flow.

By contrast only few stations feature a seasonal trend in Na^+ concentrations (see appendix C.2). Still it has to be mentioned that weak seasonalities can be disguised in the graphs due to the scaling, adjusted to single data points showing elevated monthly averages. The most prominent example for a seasonality is the station in Drasenhofen, where winter concentrations are a multiple of the summer concentrations. In a less pronounced form seasonality is also visible in Lunz. Similarly most stations show no seasonality for chlorid concentrations (see appendix C.7). Exceptions are the previously mentioned stations in Drasenhofen and Lunz, which display a Cl^- concentration increase in winter. As Na^+ and Cl^- concentrations accompany each other, road salt in the near vicinity may be one possible explanation for this behavior. In Werfenweng, which also gives a seasonality for chloride, the respective trend is less pronounced for sodium.

As can be seen in appendix C.3, ammonium concentrations show one of the strongest seasonal trends of all ions. They peak in April at almost all stations. Only the station at Ostrong is one month ahead with an NH_4^+ peak in March, while Sonnblick shows the maximum in May.

Potassium and the divalent cations calcium and magnesium show weak or no trends at all (see appendix C.4, C.5 and C.6). Exceptions are the stations Innervillgraten, Niederndorferberg and Ostrong that exhibit elevated Ca^{2+} and Mg^{2+} concentrations in spring. Overall a more detailed investigation is needed for these ions, including the retrospective evaluation of single samples.

Like ammonium, nitrate has a clear spring elevation with a maximum in April for most stations (see appendix C.8). Although the trend appears to start earlier. For some stations like Ostrong, Drasenhofen and Masenberg the NO_3^- concentrations rise even before the turn of the year.

Sulfate is another component with a clear seasonal trend. An example is given in figure 4.2. It shows the monthly SO_4^{-2}-S concentrations at Niederndorferberg. The concentrations quickly rise during springtime with a maximum in April. Then they slowly decrease till winter. The seasonal trend is still present in more recent years although it is hard to see because sulfate concentrations have dropped considerably over the years. Plots for the other stations are given in appendix C.9.

The pH value does not show a clear seasonal trend (see appendix C.10). Although many stations seem to have minima in February and/or in September with a slightly elevated period in summer. Historically hydrogen concentrations correlated closely with the sulfate concentrations (Hornbeck et al., 1976). However, in more recent history it was recognized that due to the decrease of strong mineral acids like sulfuric, nitric and hydrochloric acid, weak organic acids have increased their impact on

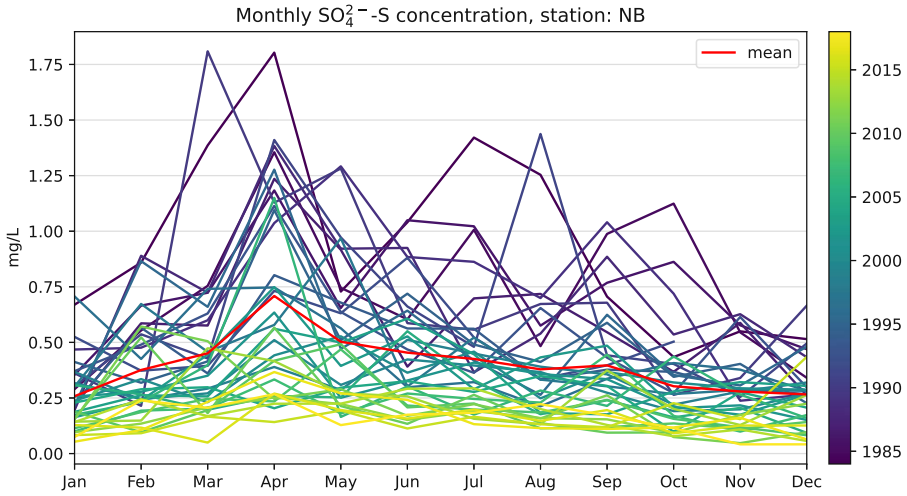


Figure 4.2: Monthly $SO_4{}^{-2}$-S concentration at Niederndorferberg

precipitation acidity in Europe (Vet et al., 2014). Nevertheless, there is no clear connection regarding seasonality, found in this work.

Table 4.1 summarizes all findings of an optic inspection of the seasonal trend plots presented in this section and in appendix C. The given months mark the perceived top of the seasonal increase, which often matches but is not necessarily the absolute max of the mean value. Exceptions are created by random accumulations or concentrated outliers in certain months that influence the mean in a way that it does not follow the seasonal trend of the majority of years. Greyed out month names are used

Table 4.1: Maxima of seasonal variation in concentration (visual inspection); grey font indicates less pronounced trends

	NS	Na^+	$NH_4{}^+$	K^+	Ca^{2+}	Mg^{2+}	Cl^-	$NO_3{}^-$	$SO_4{}^{2-}$	pH
HF	Jul	-	Apr	May	Apr	Apr	-	Apr	Apr	May
NB	Jul	-	Apr	May	Apr	Apr	-	Mar	Apr	Jun
IV	Jul	Mar	Apr	-	Apr	May	Apr	Apr	May	-
NH	Jul	Mar	Apr	-	-	-	Mar	Apr	Apr	-
WW	Jul	Mar	Apr	-	Apr	-	Feb	Apr	Apr	-
SO	-	-	May	May	Jun	-	-	Apr	May	-
LI	Jul	Dec	Apr	Sep	-	-	-	Mar	Apr	Jul
LU	Jul	Dec	Apr	Nov	-	-	Dec	Mar	Apr	Jul
OS	Jul	Jan	Mar	Nov	Apr	Apr	-	Mar	Mar	Apr
DR	Jul	Dec	Apr	-	-	-	Dec	Feb	Apr	Jul
MB	Jul	Feb	Apr	-	Feb	-	Nov	Mar	Mar	Apr
HG	Jul	-	Apr	Dec	-	-	Dec	Mar	Mar	-
GS	Jul	Feb	Apr	-	-	-	-	Mar	Apr	Apr
AF	Jul	-	Apr	-	-	-	-	Feb	-	May

when the seasonal trend is weak or barely visible. Although optic inspection seems to be a less scientific approach than the fourier analysis performed by Schreiner, the results are in good agreement. Furthermore, some fourier analysis peaks were misinterpreted as maxima although they actually represent minima. Therefore, human pattern recognition is still a viable tool for seasonal analysis. Future work may test the usage of the median instead of the mean for a more stable seasonality inspection as well as a split of the dataset in older (e.g. the 1990s) and more recent data. The impact of random forest classification, discussed in section 3, on the observed seasonal trends would be another interesting topic.

4.2. Temporal changes in seasonal trends

Section 4.1 showed that some components exhibit a clear seasonal trend. Now it can be determined if said trends have changed over the years. Therefore, data was aggregated by season (Jan - Mar, Apr - Jun, Jul - Sep, Oct - Dec), which is easily possible by using `1Q-DEC` as aggregation parameter in listing 2.

The seasons can then be plotted as individual columns in a column plot. In addition Mann-Kendall testing (two-sided p-value < 0.05) and Theil-Sen approximation (95 % confidence) can be applied to investigate the individual trends. Slopes are only plotted if the MK test indicates a trend. All plots are given in appendix D. Figure 4.3 is an example for these plots. It shows that all seasons exhibit a sta-

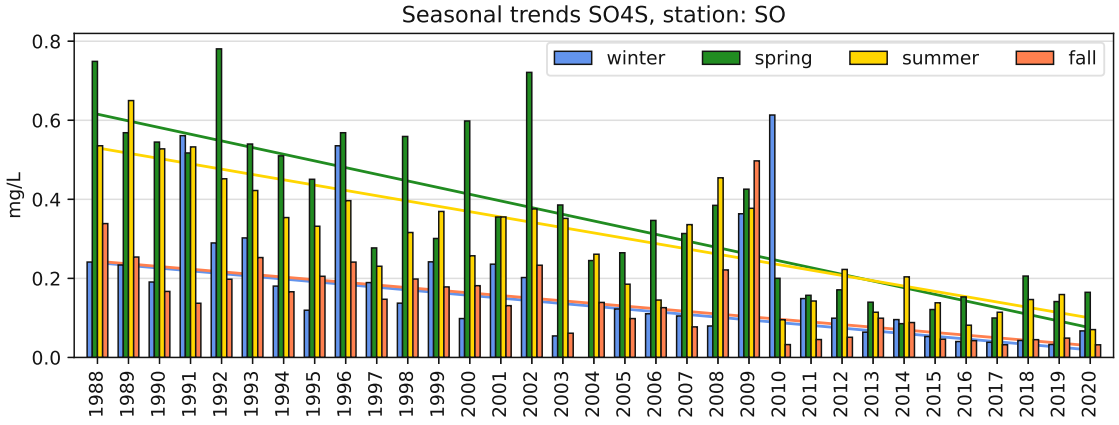


Figure 4.3: Seasonally separated $SO_4{}^{-2}$-S trends (Theil-Sen estimator) at Sonnblick

tistically significant decreasing sulfate trend at Sonnblick. In addition it reveals that concentrations in spring and summer are not only higher but also decrease faster than in fall and winter. As seen in appendix D.9, this behavior is also found in Höfen, Niederndorferberg, Innervillgraten, Haunsberg and Werfenweng although with varying degrees of clarity. This indicates that in western Austria or at least in

the inner-alpine region sulfur depositions have declined faster in spring and summer than in fall and winter.

Table 4.2 summarizes all trends. If a trend is significant for a component at a station, a colored triangle is given. The colors resemble the seasons of the year. Filled downward looking triangles are used for decreasing trends and upright triangles are used for increasing trends. The latter ones are not filled for better readability. The table illustrates that with the exception of Haunsberg and Ostrong no station shows a temporal trend in precipitation amount for any season. In contrast almost

Table 4.2: Concentration trends for seasons (winter - blue, spring - green, summer - yellow, fall - orange; filled triangle down - significant decrease, triangle up - significant increase)

	NS	Na^+	NH_4^+	K^+	Ca^{2+}	Mg^{2+}	Cl^-	NO_3^-	SO_4^{2-}	pH
HF		▼ ▼	▼ ▼ ▼ ▼		▼ ▼		▼ ▼ ▼ ▼	▼ ▼ ▼	▼ ▼ ▼ ▼	△ △ △ △
NB		△ 	▼ ▼ ▼ ▼		 ▼ ▼	▼ ▼	 ▼	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	△ △ △ △
IV		 ▼ ▼		△ 	▼ ▼	▼ ▼	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	△ △ △ △
NH	▼ ▼			△ △ △	▼ ▼ ▼ ▼		▼ ▼ ▼ ▼	▼ ▼	▼ ▼ ▼ ▼	△ △ △ △
WW		▼ ▼ ▼		△ △ △	▼ 		▼ ▼ ▼	 	▼ ▼ ▼ ▼	△ △ △ △
SO		 ▼	▼ ▼ ▼ ▼	▼ ▼	▼ ▼	▼ ▼ ▼	 ▼ ▼	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	△ △ △
LI			▼ ▼ ▼		▼ ▼ 	▼ 	▼ ▼ ▼ ▼	▼ 	▼ ▼ ▼ ▼	△
LU			▼ ▼ ▼	 ▼	 	 	 ▼ ▼	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	△ △ △ △
OS	△ 	▼ ▼	 ▼ ▼	▼ △ 	 	 	▼ ▼ ▼	 ▼	▼ ▼ ▼ ▼	△ △ △
DR			△ 	 	 	 	 	△ 	 	
MB			 ▼	 	▼ 	 	 	 	▼ ▼ ▼ ▼	△ △ △
HG			 	 	 	 	 	 	▼ ▼ ▼	△ △ △ △
GS		▼ ▼	 	 	▼ ▼ ▼	▼ ▼ ▼	▼ ▼ ▼ ▼	▼ ▼	▼ ▼ ▼	△ △ △
AF		▼ ▼ ▼ ▼	 	▼ ▼ ▼ ▼	▼ ▼ ▼ ▼	▼ ▼ ▼	▼ ▼ ▼ ▼	 	▼ ▼ ▼ ▼	△ △

all stations have decreasing SO_4^{2-} concentrations and increasing pH values. The only exception is Drasenhofen where the time series only spans from 2004 to 2017. Due to this short period significant trends are only found for nitrogen containing components, which increase in winter. Although this likely may be a coincidence as NH_4^+ and NO_3^- concentrations decrease for all other stations. In addition, results obtained for potassium should be regarded cautiously, as concentrations are generally quite low and still might be influenced by single events. One half of the significant K^+ trends is increasing, while the other is decreasing. K^+ is also the only component that has both increasing and decreasing trends for different seasons, although this is only the case for Ostrong.

Overall it does not seem to be necessary to observe trends separately per season. Nevertheless, some components like sulfate and the pH value exhibit vastly differing slopes for different seasons. Figure 4.3 is one example for this behavior. All inner-alpine stations exhibit faster decreasing sulfate concentrations in spring and summer compared to fall and winter. As a matter of fact for the stations Niederndorferberg and Haunsberg the steep slopes for spring suggest negative concentrations by now. This is caused by the very strong sulfur reductions in the last century and the linear Theil-Sen trend line that cannot account for the flattening in recent years. This leads to a possible segmentation of the time intervals which will be discussed in the next section.

4.3. Trend segmentation and EMEP emission data

In this section it will be investigated if a trend separation in two parts is feasible to reflect the changing conditions for sulfur and oxidized/reduced nitrogen in a better way. In addition, a comparison with emission data is performed.

All European countries submit emission data to the EMEP Centre on Emission Inventories and Projections (CEIP). One way to check which countries need to be considered for a proper comparison is to reverse model the Austrian immission data. As this is beyond the scope of this thesis a less resourceful way was chosen. By combining gridded European emission data with deposition data from EMEP MSC-W model a general idea on species transport can be obtained.

For sulfur this is done in figure 4.4. Please mind that due to scaling some emission hotspots appear black instead of red on the map. These points align nicely with regions of high deposition. Natural sources like volcanoes and sea salt seem to play a major part for sulfur depositions. But also larger cities and heavily industrialized regions exhibit plumes, which influence deposition over several hundred kilometers. Examples are the northern coast of Spain or eastern Ukraine. Therefore, it is not sufficient to exclusively use Austrian emission data. At least all neighboring countries have to be considered. A special feature of Austria is its topography, which strongly influences transport of air masses and pollutants. In figure 4.5 and 4.6 this is even

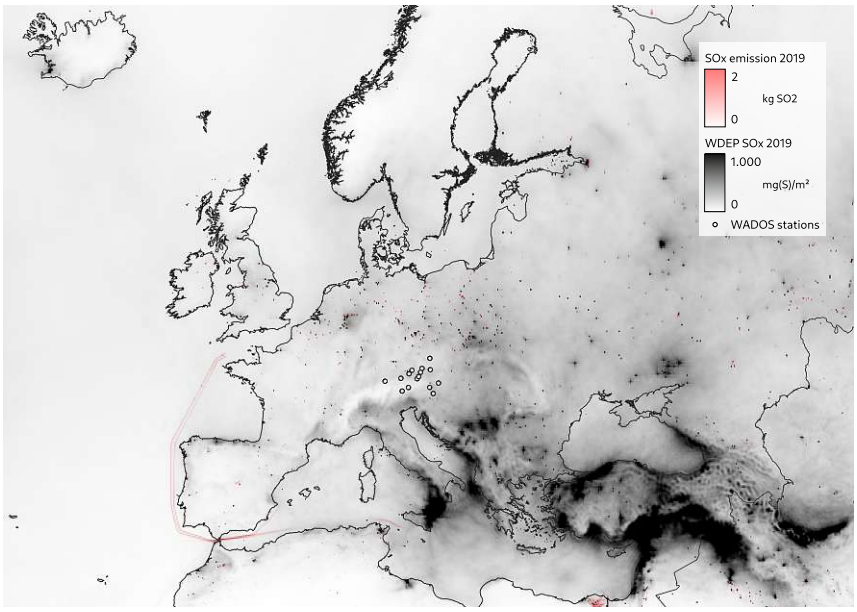


Figure 4.4: Map showing gridded SOx emissions ($0.1° \times 0.1°$) and wet deposition of sulfur in 2019; data provided by EMEP/MSC-W and EMEP/CEIP (2021), coastline map by European Environment Agency

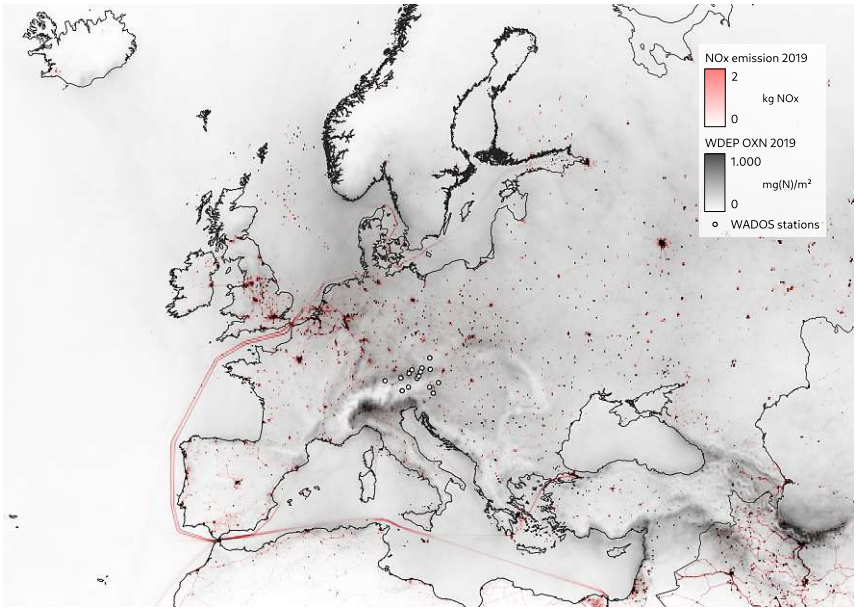


Figure 4.5: Map showing gridded NOx emission ($0.1° \times 0.1°$) and wet deposition of oxidized nitrogen in 2019; data provided by EMEP/MSC-W and EMEP/CEIP (2021), coastline map by European Environment Agency

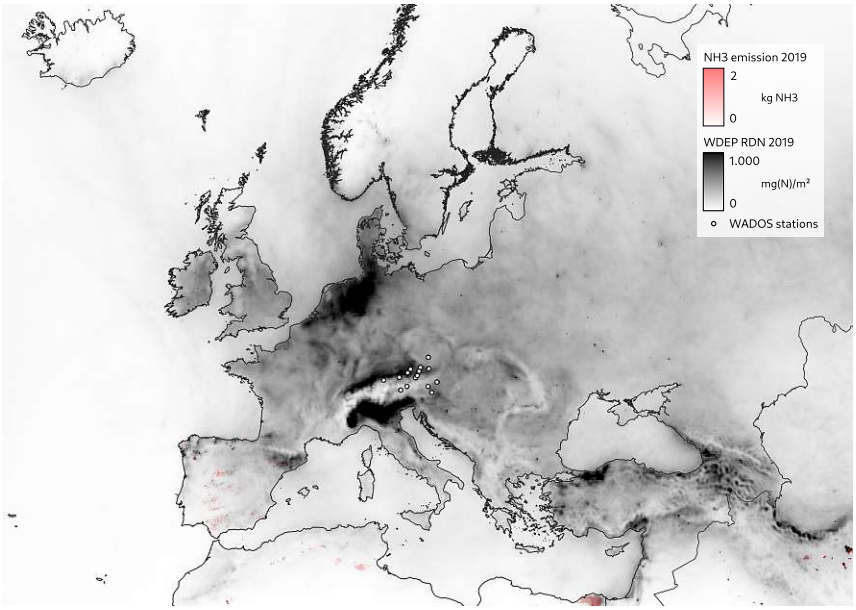


Figure 4.6: Map showing gridded NH_3 emission ($0.1° \times 0.1°$) and wet deposition of reduced nitrogen in 2019; data provided by EMEP/MSC-W and EMEP/-CEIP (2021), coastline map by European Environment Agency

more obvious because of intense nitrogen emission in the Po valley caused by road transport, industry and agriculture (Larsen et al., 2012). The maps indicate that stations are influenced differently by different countries. Deviating results at the various stations are therefore expected. However, a detailed elaboration on local specialties is beyond the scope of this work. Therefore, stations will be compared with the sum of the reported emissions from Germany, Czech Republic, Poland, Slovakia, Hungary, Slovenia, Croatia, Italy, Switzerland, France and Austria. This area is 1 945 492 km^2 in size (CIA, 2021).

A breakdown of the SOx, NOx and NH_3 emissions per country is given in figure 4.7. Emission data was obtained from EMEP/CEIP (2021), although this time not the gridded but the annual total sum for each country was used. Please mind that this data is claimed to be "inconsistent and/or incomplete" in comparison with the data that is used for modeling by EMEP. But the latter one is not available for years before 1990. Indeed some countries lacked some data points before 1990. To prevent distortion of total emissions, backward filling was applied. This means that some missing emission values from Hungary, Czech Republic, Slovakia and Poland were replaced with their next existing value. As expected by their size Germany, Italy, Poland and France are responsible for a majority of the total emissions in the chosen region. All three components show decreasing emissions but at different speeds. Sulfur emissions have decreased by an order of magnitude in the last 40 years. Meanwhile oxidized and reduced nitrogen have been reduced by 60 % and 40 % respectively. In addition the reduction speeds differ between the countries.

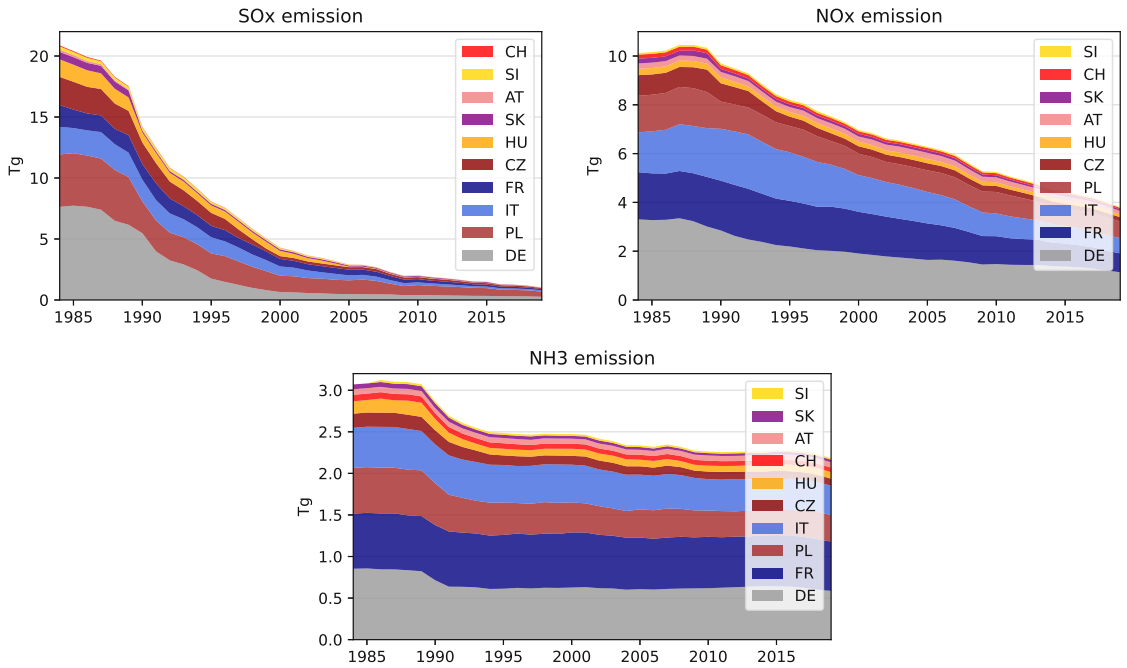


Figure 4.7: SOx, NOx and NH_3 emissions stacked by country; data source: EMEP/-CEIP (2021)

Regardless of the scaling, aforementioned emissions are now used to give context to observed deposition trends in the Austrian Precipitation Sampling Network. As a first example the sulfur depositions at Haunsberg are compared with the SOx emissions. Figure 4.8 clearly shows steeper slopes for the sulfur emission as well as the deposition in the last century. In addition to the Theil-Sen estimator, the deposition was modeled with a Gaussian filter, which essentially is a weighted moving average. A more detailed explanation on this smoothing algorithm is for example given by Regmi (2021). The filter is not used as a competitor to the Theil-Sen/Mann-Kendall procedure because it does not make any statement about the significance of the modeled trend. Instead it is given to simplify the trend tracking for the viewer. This allows to compare the reported emissions with a linear and a more flexible model of the measured depositions. Obviously the Theil-Sen estimator is not able to reflect the changing trend due to its linearity. In contrast to that, the smoothed line fits emission data better. This, however, is not true for all stations and components.

For example if the same plot is prepared for oxidized nitrogen. This is done in figure 4.9, which compares nitrogen deposition introduced by nitrate and NOx emissions. Here smoothing and Theil-Sen estimation fit the data equally well. The flatter NOx emissions in the 1980s are better represented by the Gaussian filter. This is also true if only German emissions are considered, which according to figure 4.7 plateaued in this period. However, the flattening of the smoothed curve after 2010

is not explained by emission data. Here the Sen's slope gives a more appropriate representation of the trend.

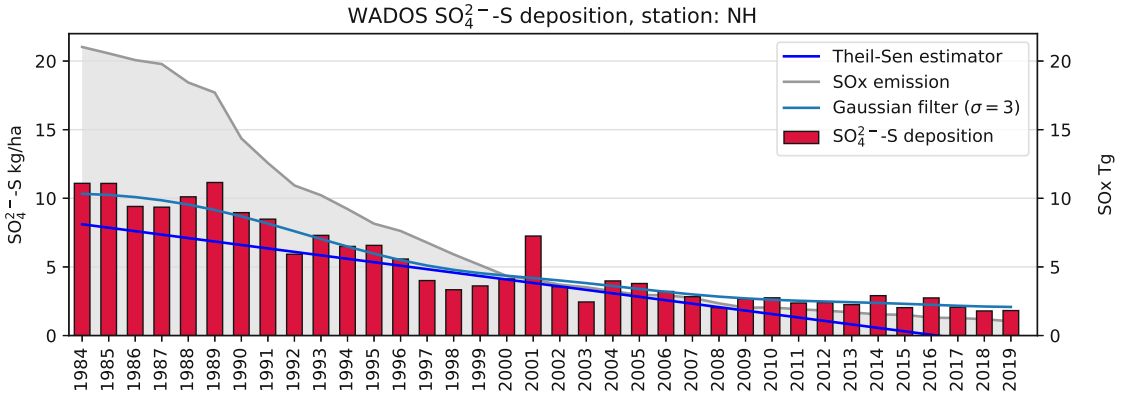


Figure 4.8: $SO_4{}^{-2}$-S deposition at Haunsberg with corresponding SOx emission data from EMEP/CEIP (2021)

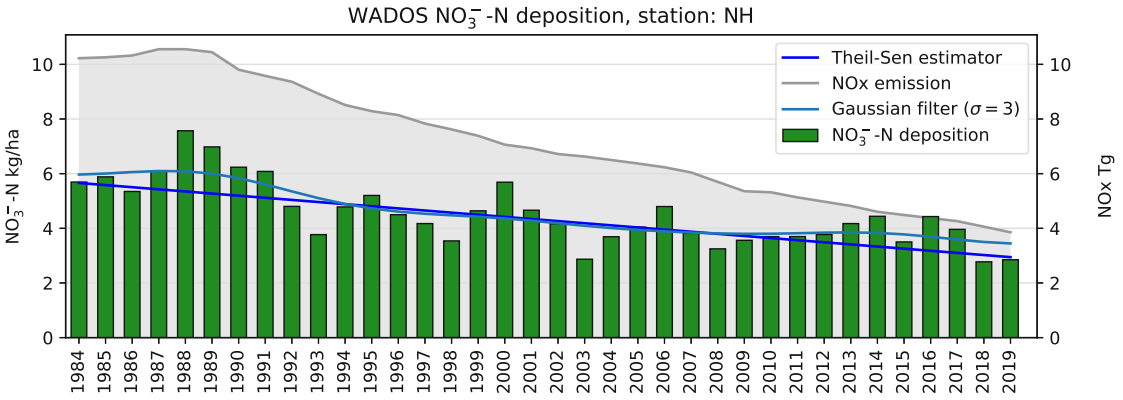


Figure 4.9: $NO_3{}^{-}$-N deposition at Haunsberg with corresponding NOx emission data from EMEP/CEIP (2021)

A completely different picture emerges when reduced nitrogen is investigated. Figure 4.10 compares nitrogen depositions, caused by ammonium, and NH_3 emissions. In contrast to SOx and NOx, the total ammonia emissions are much lower and more stable. The $\curlyvee$-shaped course of the smoothed line can certainly not be explained with the emission sum of the chosen countries. Reasons for the line shape have to be more local or may even be coincidental. However, the Theil-Sen estimator shows a falling trend that is backed up by Mann-Kendall testing. Therefore, it is better suited to depict the actual supra-regional trend.

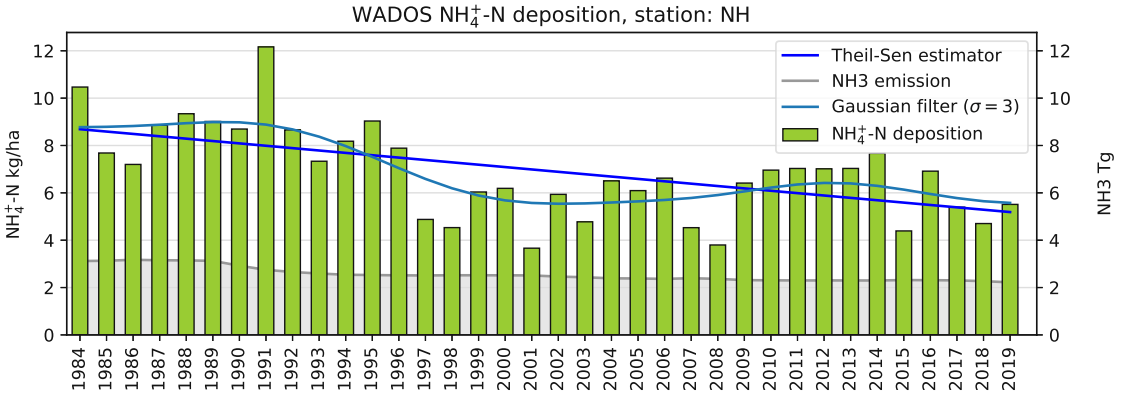


Figure 4.10: $NH_4{}^+$-N deposition at Haunsberg with corresponding NH_3 emission data from EMEP/CEIP (2021)

The previous examples show that despite the ever increasing time series length, Theil-Sen estimation is still a valuable tool to investigate trends. However, due to the extreme reductions in SOx emissions in Central Europe in the 1990s a linear model does not fit the data properly. To account for the changing sulfur deposition trend a split of the data is proposed.

For Haunsberg this is done in figure 4.11. Based on visual inspection several splits were tested. For this time series a split at the millennium border seems to be the most reasonable. This approach cannot account for the more stable period before 1990 but it matches the data better than the Theil-Sen estimator based on the full time series. In comparison with the smoothed line in figure 4.8 the split approach offers two advantages. Firstly the Mann-Kendall test is used to attest the significance of both trend parts. And secondly the Theil-Sen estimator is more robust to outliers like the elevated value in 2001.

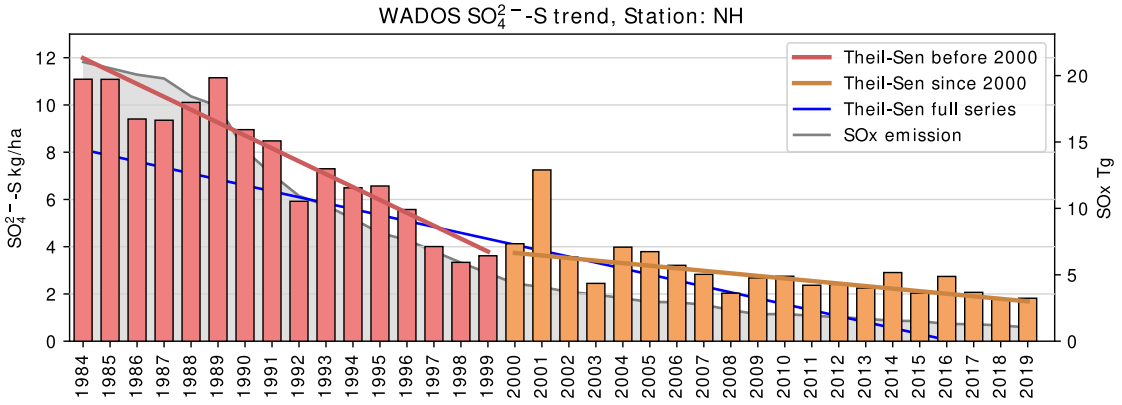


Figure 4.11: $SO_4{}^{-2}$-S deposition at Haunsberg with a dataset split in 2000

Now the question arises if the split method is applicable at all stations. Investigations with the Gaussian filter showed that 8 out of 12 stations with long time series have a steeper slope in the first half of the dataset. A successful application is defined conservatively with two assumptions. Both split trends have to be significant and the second part has to have a lower slope than the first part. Even with these quite general requirements only 4 time series passed. Since it was Niederndorferberg, Haunsberg, Litschau and Innervillgraten, there is no local connection between the stations where the split can be applied. There were two main reasons for the failing of the method at the other stations. The stations in Styria and at Ostrong started operation in 1990. Ten or less years are not sufficient for the Mann-Kendall test to detect a separate trend until 2000. The second main reason lies in the start period of the Austrian Precipitation Sampling Network in the mid 1980s when strong variations occurred that in most cases do not match the rapidly falling trend in the following decade. At the one hand this behavior is covered by emission data, especially since Germany reported stagnating sulfur emission in that period. But on the other hand the results are not very stable, which may be caused by local conditions.

One example where the split method failed was the global background station at mount Sonnblick. Strong fluctuations in sulfur deposition are observed in figure 4.12, although there are no local sulfur sources to be expected near the observatory. Furthermore, the decrease in the 1990s does not seem to be more pronounced than in the rest of the dataset. It is not possible to split the dataset in parts with different trends that are both significant according to Mann-Kendall testing. This is interesting as the station at an elevation of 3106 m.a.s.l. definitely reflects the background conditions above Europe. Obviously it was less impacted by the maximum emissions occurring during the 1980s and 1990s. This corresponds to the results obtained for the seasonal trends in section 4.2. The decrease is less pronounced during the cold season, when emissions are highest, but most likely will not reach the 3 km level.

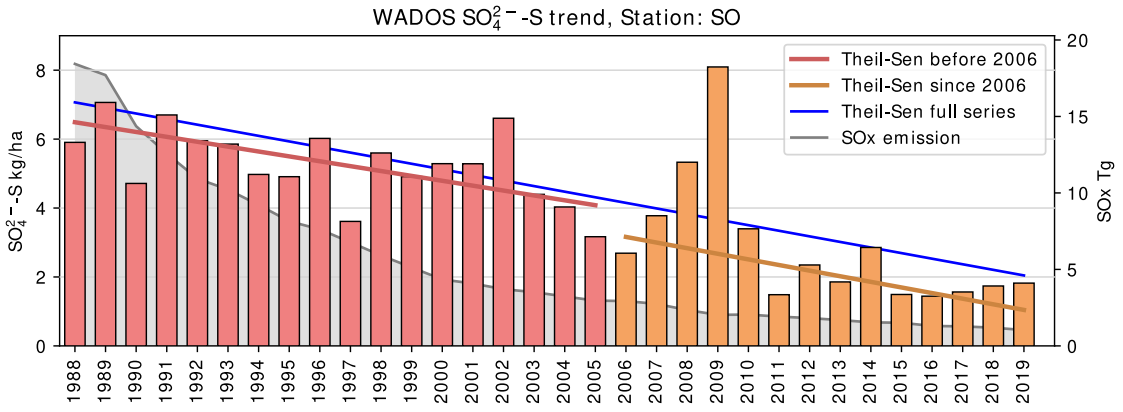


Figure 4.12: SO_4^{-2}-S deposition at Sonnblick with a dataset split in 2006

Another possibility to avoid the influence of the period of the rapidly decreasing emissions is to exclude that period from the analysis and to state which time period the current trend is reflecting. This way the data from varying stations becomes more comparable because stations started operation in different years and therefore include different amounts of said period. However, the question arises when to set this new starting point. The emissions in figure 4.7 indicate differing situations for different components. For sulfur the situation is most obvious as emissions decreased rather fast until 2000. Therefore emission data suggests to set the starting point in the range of 1990 to 2000. One possible starting point is 1991 as the Styrian stations Masenberg, Hochgößnitz and Grundlsee started operation in 1990. Regarding the measured depositions the situation is too varied to give one distinct proposition. For some stations and components the starting point changes only little, whereas for others every removed year has a noticeable impact on the observed trend. For ammonia and NOx the emissions decrease more gradually and the measured depositions do not require a split to avoid trend lines in the negative concentration range.

To sum up, dataset splitting for a better reflection of changing deposition trends is only suitable in some cases, but then it makes sens and is necessary to avoid negative values. Although the drastic reduction in SOx emissions in Central Europe are reflected in the deposition values at almost every station (see table 4.2), it is, however, not obvious that the trend lines are not uniform. While Haunsberg and three other stations show a slowdown of the decline, the other stations do not share this trend or at least it cannot be statistically proven with the methods applied within this work. It is especially interesting that the GAW background station Sonnblick does not show a slowdown of the decreasing trend. However, as the sulfur depositions are stable since 2011, the analysis has to be repeated regularly.

5. Summary

5.1. Precipitation database

Within the scope of this work a database including all available measurements of the Austrian Precipitation Sampling Network was created. The dataset fulfills the requirements of level 1 data according to the GAW definition given in section 1.4. This means that all daily samples including contaminated ones are present. In addition a flag system was developed and implemented to enable auto removal of contaminated and other invalid samples. The results of all previous data review processes were migrated into this new system.

To achieve said objectives a comprehensive data analysis and clean-up process was performed. Five problems were identified as the main sources for inhomogeneities in the old datasets.

1. Date offsets led to misalignment of the datasets and were caused by ambiguous definitions of the sampling date. For the database the day before sample extraction was chosen as it better reflects the date of the actual precipitation event.

2. Another reason for deviations from the old system was rounding. As interim results were not passed on, this restricted the review process for datasets prior to 1994. Regarding the database no additional rounding is applied above the limitations of double precision numbers (16 digits). For future entries the amount of decimal places is therefore only defined by the measurement output.

3. During data examination the problem of sample overflow was uncovered. Precipitation depth exhibits overflow at events beyond 31 mm at some stations. Whenever possible data from the next station of the Austrian hydrographic service was used to match the previously reported precipitation amounts. Nevertheless, some stations still exhibit an according accumulation in their precipitation depth distribution.

4. One minor issue was the presence of duplicated samples in the original files. These mostly occurred due to initial tests, the sampling of rinsing water and double bottling events. However, a maximum of one sample per day and station remained in the database.

5. Finally the factor that converts the ion concentrations of NO_3^-, NH_4^+ and SO_4^{2-} to the equivalent concentrations of nitrogen and sulfur was not uniformly applied. The database does not include this factor for any sample.

The report creation workflow was redesigned to incorporate and make the maximum out of the database. Due to the inline position of the new database, data integrity is preserved and data export can be performed from a central source. As

much steps as possible were automated within a Python script to facilitate and minimize error probability. Nevertheless, some steps of the data review process remained within the *ionenprüfen* files because of their interactive hands-on character.

The *ionenprüfen* files are therefore used to import new data to the database. Flags can be set within the *ionenprüfen* files during the data review process. If flags are added retroactively they are incorporated automatically at the next import run. Incorrect data formats raise errors during the process. Thus database integrity is maintained and no incorrect results can be exported.

Like the import script the script for data export was written as a Jupyter Notebook. It contains code cells that produce data files and plots that are necessary for report creation or that are generally useful for data analysis. The most important parts, like the flag system usage or the calculation of aggregated concentrations or depositions are useful for all users of the database. They were introduced in section 2.6.

5.2. Random forest classification

The random forest classification was mainly performed for two reasons. The first one is to show an application that makes use of the level 1 character of data. But more importantly it confirmed that contaminated samples that were overlooked in previous data review processes may only lead to minor trend changes in some stations and cannot cause trend reversals. In addition this work provides indications on the usage of random forest classification for data series homogenization.

In section 3.1 different possibilities for training data and classifier settings were investigated. It turned out that a compromise in several performance metrics was reached when all data after October 2014 and contaminated samples before this date were used for training. With regard to classes, this dataset is not balanced, which leads to a lower and therefore more realistic number of found contaminations. Although downsampling the valid class improved some metrics it was not considered to be a practical path. However, upsampling the contaminated class seems to be a promising idea for further research.

After an analysis of the out-of-bag error rate it was discovered that hundred trees are sufficient to not restrict classifier performance, without being too computationally expensive. An analysis of feature importance revealed conductivity, calcium and sulfate concentrations as the most important factors for classification. The least important ones are precipitation amount and depending on the typ of calculation magnesium, potassium or nitrate concentrations.

Furthermore, the possibility to influence classification results by tuning the majority vote of the trees was explored. The impact of RF classification was examined with

the default 50 and a 90 % decision limit. These result in a 12 and 2.5 % incidence of sample contamination respectively. The true amount is believed to be somewhere in between. The data sets that have been cleaned of the marked values therefore represent a best estimate on how a complete reevaluation of older entries may influence observed trends.

While providing valuable insights the random forest classifier is not ready to be used unsupervised or as a replacement for manual data evaluation. The distinction between valid and contaminated samples is unclear in many cases and requires expert knowledge and sometimes even individual investigations. However, further development may bring the classifier in a position where it can support the manual review process both for new and for old samples. First flags have already been set based on classifier input. Although, generally the results of the random forest classification, performed within this thesis, are not included in the database. Further flagging may improve classifier performance because of the growing training set. Still, this would need a completely new evaluation.

5.3. Trends

In section 4 further analysis examples based on the database are given. The first one was centered around the seasonal trends within years. By plotting monthly aggregated time series of every year on top of each other, the existence of seasonality was presented. All plots are given in appendix C. Table 4.1 on page 44 summarizes which components and at which stations exhibit seasonality. In addition the months that are the perceived maximum of each seasonal increase are given. A rough classification on how pronounced the trends are is indicated through font color.

The table reveals clear seasonal trends for the precipitation depth, ammonium, nitrate and sulfate. The precipitation amount peaks for all stations in July, except at the Sonnblick observatory where no trend is observed. Ammonium, nitrate and sulfate concentrations peak in spring with small deviations between the stations and the components. This corresponds with the results of the fourier analysis by Schreiner (2017).

After determining the existence of seasonal trends, their change over the course of time was investigated. For this purpose, the data was aggregated per season and for every season an individual time series was created. Theil-Sen estimation was used for trend representation and Mann-Kendall testing was applied to ensure statistical significance of each trend. Table 4.2 on page 46 summarized the trends for each individual season by depicting the presence of significant rising or falling trends for each station and component.

It was revealed that almost no trends regarding precipitation depth are observed. While pH values rise in almost all stations, sulfate concentrations fall in almost all

stations. Overall the trends within the different seasons are consistent. Interestingly some components like sulfate and the pH value exhibit differing slopes for different seasons. It was uncovered that sulfur depositions in the inner-alpine region have declined faster in spring and summer than in fall and winter.

In section 4.3 the usage of a dataset split to better describe possibly nonlinear sulfur and nitrogen deposition trends was investigated. In order to have comparison values for the depositions of the Austrian Precipitation Sampling Network, NOx, NH_3 and SOx emission data from EMEP/CEIP (2021) was evaluated. To check which countries emissions need to be taken into account, maps with reported emission and modeled deposition data were created. The maps led to two conclusions. Firstly transboundary air pollution transport necessitates the consideration of at least all neighboring countries. For this work France, Poland and Croatia were considered as well. Secondly due to Austria's position in the Alps and at its foothills, different areas of influence for the different stations are to be expected.

Emissions per country were given in figure 4.7 on page 50. Naturally the biggest countries Germany, France, Poland and Italy contribute the most in the considered region. The sum of all countries was used to give context to the observed depositions. The smoothing of a Gaussian filter was used to show a nonlinear model. Only for sulfur this smoothed lined resembled the reported emissions better than the linear Theil-Sen estimator. This behavior can probably be explained with the exponentially decreasing SOx emissions in the last 35 years. Therefore, the dataset splitting approach was tested on sulfur data. For four out of twelve stations it was possible to find a common split point that separates the data in two independent series that both contain significant and distinct trends. This was a surprising result, especially since these four stations have no local relation. Related to the idea of the split it was proposed to move the starting point of the series forward in time to better match the current situation. Additionally this would improve the comparability of the stations. However, there is no obvious choice for a unified starting point regarding the measured depositions. Therefore 1991 was proposed because three Styrian stations started operation at that time.

The given analysis examples prove the advantages of the newly formed database. The achieved speedup and the omission of lengthy data collection enables comprehensive analysis that does not need to be limited to small excerpts of the available data.

Bibliography

Allen, M. A., ed. *Manual for the GAW Precipitation Chemistry Programme*. Tech. rep. No. 160. World Meteorological Organization, 2004. https://qasac-americas.org/manual.

BMGU. *Richtlinie 11, Immissionsmessung des nassen Niederschlags und des sedimentierten Staubes, Luftverunreinigung - Immissionsmessung*. Vienna, Austria: Bundesministerium für Gesundheit und Umweltschutz, 1984.

Boughorbel, S., Jarray, F., and El-Anbari, M. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PLOS ONE* 12, no. 6 (June 2017): 1–17. https://doi.org/10.1371/journal.pone.0177678.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. *Classification and Regression Trees*. Taylor & Francis, 1984. https://books.google.at/books?id=JwQx-WOmSyQC.

Capellupo, D. "Data Science Trends Based on 4 Years of Kaggle Surveys". *towards data science* (Jan. 2, 2021). Visited on 11/02/2021. https://towardsdatascience.com/data-science-trends-based-on-4-years-of-kaggle-surveys-60878d68551f.

Cehak, K. and Chalupa, K. "Observations of various chemical contaminants of the precipitation at a BAPMoN station in the eastern pre-alpine region". *Arch. Met. Geoph. Biocl., Ser. B* 35 (1985): 307–322. https://doi.org/10.1007/BF02334487.

CIA. *The World Factbook - Europe*. Central Intelligence Agency, 2021. Visited on 12/07/2021. https://www.cia.gov/the-world-factbook/europe/.

Conover, W. *Practical nonparametric statistics*. 2nd ed. 493. New York: John Wiley / Sons, 1980.

EMEP/CEIP. *Present state of emission data*, 2021. https://www.ceip.at/webdab-emission-database/reported-emissiondata.

Erisman, J. W. and Draaijers, G. *Atmospheric deposition in relation to acidification and eutrophication*. Studies in environmental science ; Elsevier Science, 1995.

Fawcett, T. "An introduction to ROC analysis". *Pattern Recognition Letters* 27, no. 8 (2006): 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Finlayson-Pitts, B. J. and Pitts, J. N. *Chemistry of the Upper and Lower Atmosphere*. 294–348. San Diego: Academic Press, 2000. https://doi.org/10.1016/B978-012257060-5/50010-1.

Firmkranz, J. "Evaluation of the chemical composition of wet precipitation samples collected in Austria". Diploma thesis, TU Wien, 2019. https://repositum.tuwien.at/handle/20.500.12708/2061.

Fowler, D. et al. "A chronology of global air quality". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378, no. 2183 (2020): 20190314. https://doi.org/10.1098/rsta.2019.0314.

Han, H., Wang, W.-Y., and Mao, B.-H. "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning". In *Advances in Intelligent Computing*, ed. by D.-S. Huang, X.-P. Zhang, and G.-B. Huang, 878–887. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. https://doi.org/10.1007/11538059_91.

Hornbeck, J. W., Likens, G. E., and Eaton, J. S. "Seasonal patterns in acidity of precipitation and their implications for forest stream ecosystems". In *Proceedings of the first international symposium on acid precipitation and the forest ecosystem*, ed. by L. S. Dochinger and T. A. Seliga, 597–609. U.S. Department of Agriculture, 1976. https://www.fs.usda.gov/treesearch/pubs/11469.

Hunter, J. D. "Matplotlib: A 2D graphics environment". *Computing in Science & Engineering* 9, no. 3 (2007): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Larsen, B., Gilardoni, S., Stenström, K., Niedzialek, J., Jimenez, J., and Belis, C. "Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources". *Atmospheric Environment* 50 (2012): 203–213. https://doi.org/10.1016/j.atmosenv.2011.12.038.

Matthews, B. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405, no. 2 (1975): 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

McKinney, W. "Data Structures for Statistical Computing in Python". In *Proceedings of the 9th Python in Science Conference*, ed. by S. van der Walt and J. Millman, 56–61. 2010. https://doi.org/10.25080/Majora-92bf1922-00a.

Miles, L. J. and Yost, K. J. "Quality analysis of USGS precipitation chemistry data for New York". *Atmospheric Environment (1967)* 16, no. 12 (1982): 2889–2898. https://doi.org/10.1016/0004-6981(82)90039-7.

Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12 (2011): 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Kroneis GmbH. *Productinformation: WADOS Wet And Dry Only precipitation Sampler*. Iglaseegasse 30-32, A-1190 Vienna, Austria: Kroneis GmbH, 2005.

Puxbaum, H. et al. "Long-term assessment of the wet precipitation chemistry in Austria (1984–1999)". *Chemosphere* 48, no. 7 (2002): 733–747. https://doi.org/10.1016/S0045-6535(02)00125-X.

Regmi, S. "Gaussian Smoothing in Time Series Data". *towards data science* (May 30, 2021). Visited on 11/29/2021. https://towardsdatascience.com/gaussian-smoothing-in-time-series-data-c6801f8a4dc3.

Schreiner, E. "Analyse der Messungen zur Nassen Deposition seit 1983 - Zeitliche Trends und Saisonalität sowie Regionale Unterschiede". Diploma thesis, TU Wien, 2017.

Sen, P. K. "Estimates of the Regression Coefficient Based on Kendall's Tau". *Journal of the American Statistical Association* 63, no. 324 (1968): 1379–1389. https://doi.org/10.1080/01621459.1968.10480934.

Simpson, D. et al. "The EMEP MSC-W chemical transport model - technical description". *Atmospheric Chemistry and Physics* 12, no. 16 (2012): 7825–7865. https://doi.org/10.5194/acp-12-7825-2012.

Vet, R. et al. "A global assessment of precipitation chemistry and deposition of sulfur, nitrogen, sea salt, base cations, organic acids, acidity and pH, and phosphorus". *Atmospheric Environment* 93 (2014): 3–100. https://doi.org/10.1016/j.atmosenv.2013.10.060.

Virtanen, P. et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". *Nature Methods* 17 (2020): 261–272. https://doi.org/10.1038/s41592-019-0686-2.

List of figures

Appendix A Additional information

Table A.1: Laboratories responsible for sample analysis

ID	name	state	laboratory	start	end
TB	Thüringerberg	Vbg.	CTA, TU Wien	Apr 90	Mar 92
GA	Gaschurn	Vbg.		Apr 92	Apr 94
HD	Hard	Vbg.		May 94	Mar 98
BZ	Bizau	Vbg.		Apr 98	Mar 01
AM	Amerlügen	Vbg.		Apr 01	Aug 03
HF	Reutte	T	Chemisch-technische Umweltschutzanstalt, Tirol	Nov 83	
AK	Achenkirch	T		Oct 83	Aug 95
NB	Niederndorferberg	T		Oct 83	
IV	Innervillgraten	T		Aug 84	
SG	IBK-Seegrube	T		Dec 85	Apr 88
RA	IBK-Reichenau	T		Dec 85	Mar 88
NL	Nößlach	T		Oct 84	Sep 85
IS	Innerschmirn	T		Oct 85	Jun 88
NH	Haunsberg	Sbg.	CTA, TU Wien / Landeslabor Salzburg	Oct 83	
SF	Sbg. Flughafen	Sbg.		Oct 83	Sep 86
GB	Gaisberg	Sbg.		Jul 89	Nov 90
SK	St. Koloman	Sbg.		Oct 83	Dec 03
WW	Werfenweng	Sbg.		Oct 83	Sep 18
KS	Kolm Saigurn	Sbg.		Jul 89	Apr 95
SO	Sonnblick	Sbg.	CTA, TU Wien	Oct 87	
AS	Almsee	OOE	Umwelt Prüf- und Überwachungsstelle, Oberösterreich	Jan 86	
AP	Aspach	OOE		Jan 94	
KM	Kremsmünster	OOE		Jan 84	
LR	Linz-Römerberg	OOE		Jan 16	
MB	Masenberg	Stmk.	CTA, TU Wien / Fachabteilung 17a, Steiermark (1992 - 2009)	Mar 90	
HG	Hochgößnitz	Stmk.		Feb 90	
GS	Grundlsee	Stmk.		Feb 90	
WZ	Weiz	Stmk.		Apr 90	Sep 91
SA	Stolzalpe	Stmk.		Oct 93	Apr 97
ND	Niklasdorf	Stmk.		Oct 02	Sep 08
AF	Arnfels	Stmk.		Oct 97	
NF	Nassfeld	Ktn.	CTA, TU Wien	Oct 89	Mar 98
VH	Vorhegg	Ktn.		Oct 97	Dec 09
HB	Herzogberg	Ktn.		Oct 99	Sep 10
HW	Hirschwang	NOE	CTA, TU Wien	Apr 86	Mar 88
NW	Naßwald	NOE		May 88	Sep 07
LI	Litschau	NOE		Oct 89	
WD	Wolkersdorf	NOE		Oct 89	Sep 97

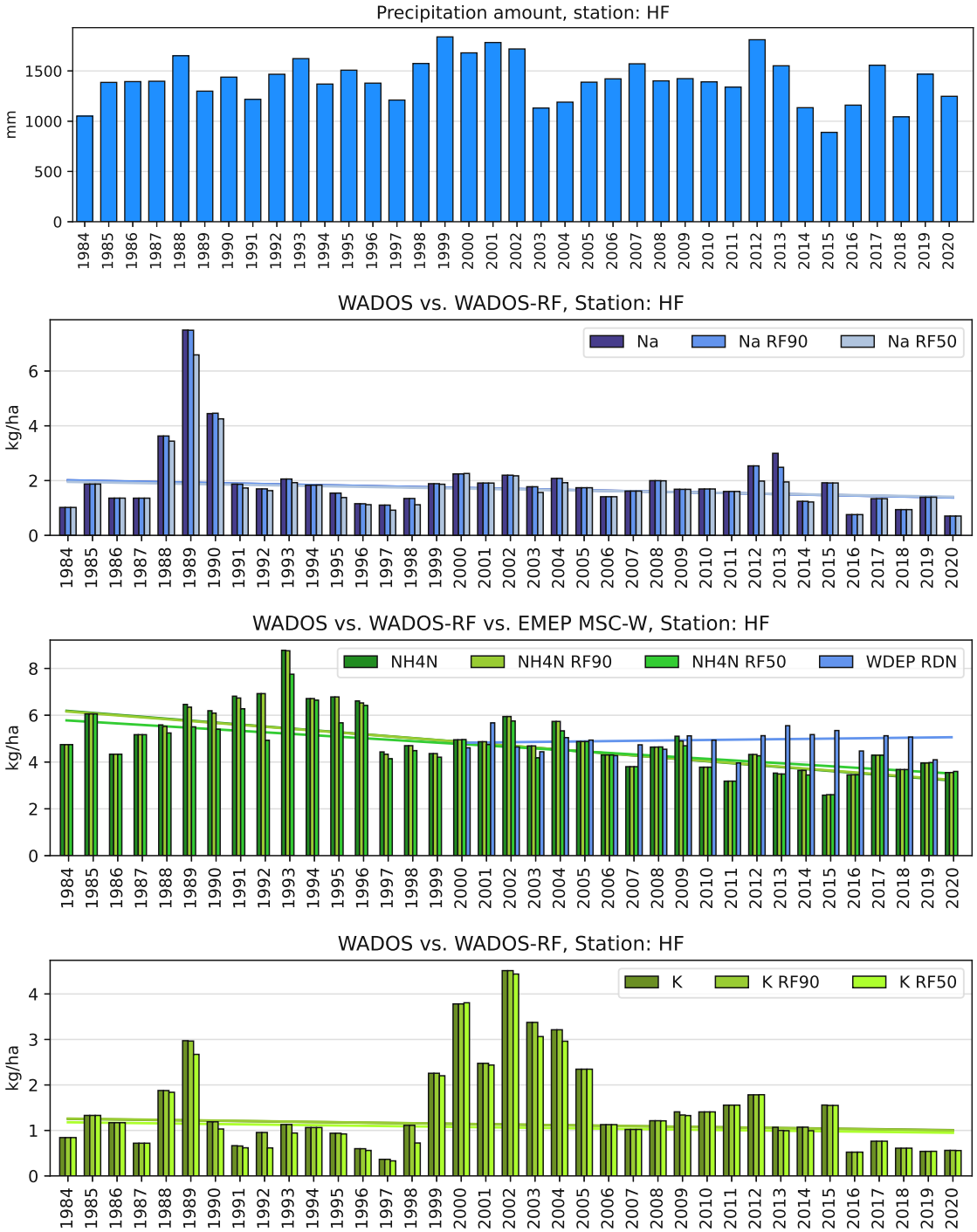
Table A.1: Laboratories responsible for sample analysis

ID	name	state	laboratory	start	end
MI	Mitterhof	NOE		May 98	Apr 03
DR	Drasenhofen	NOE		Oct 03	Nov 17
GK	Großkadolz	NOE		Mar 20	Jul 20
JB	Josefsberg	NOE	CTA, TU Wien	Oct 89	Aug 96
LU	Lunz	NOE		Apr 87	
OS	Ostrong	NOE		Apr 91	
KL	Kl.-Leopoldsdorf	NOE		Jul 91	Sep 97
LZ	Lainz	W		Apr 86	Sep 07
LB	Laaer Berg	W	CTA, TU Wien	Apr 86	Mar 90
LO	Lobau	W		Apr 86	Sep 07
BI	Bisamberg	W		Apr 90	Sep 07

Table A.2: Primarily responsible people at Institute of Chemical Technologies and Analytics

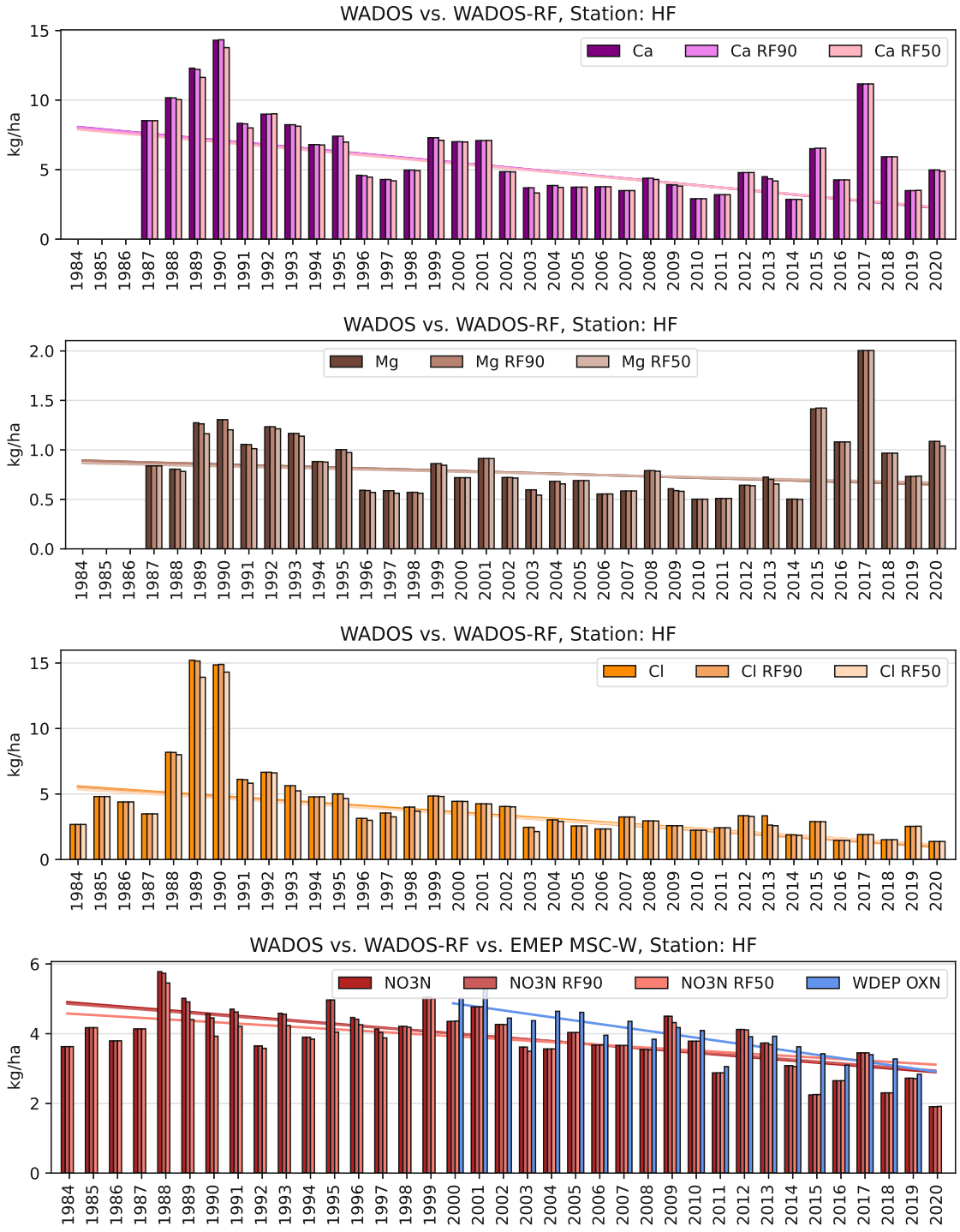
1983 - 1993	Andreas Kovar	Hans Puxbaum
1993 - 2002	Michael Kalina	
2002 - 2008	Klaus Leder	
2008 - 2010		Heidi Bauer
2010 - 2011	Elisabeth Schreiner	
2011 - 2014		Anne Kasper-Giebl
2014 - 2017	Julia Firmkranz	
2017 - 2019	Thomas Rosa-Steinkogler	
2019 -	Hong Huang	

Appendix B Random forest classification

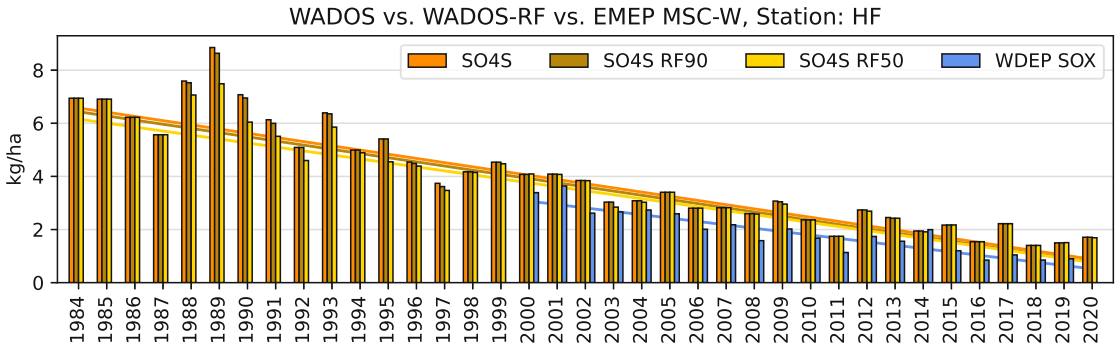


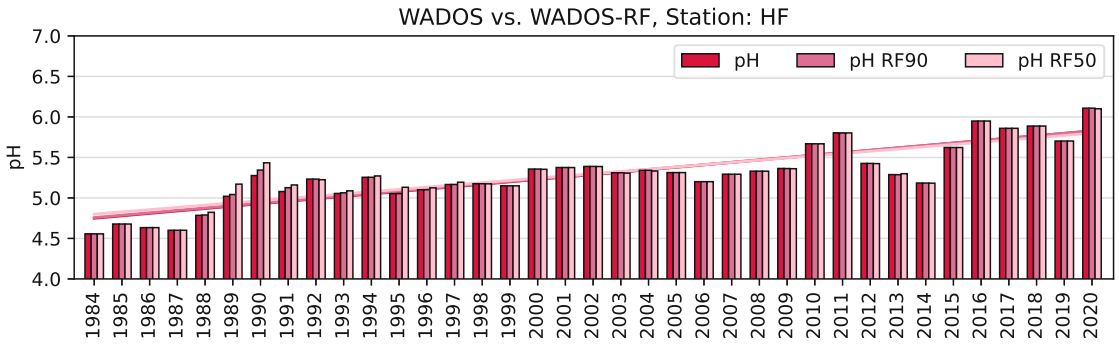
(a) Precipitation amount, sodium, reduced nitrogen, potassium

Figure B.1: Influence of RF classification on depositions in Höfen

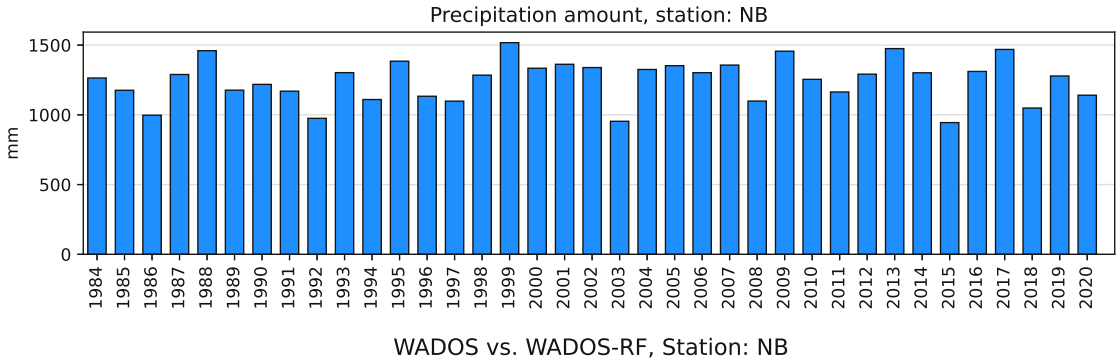


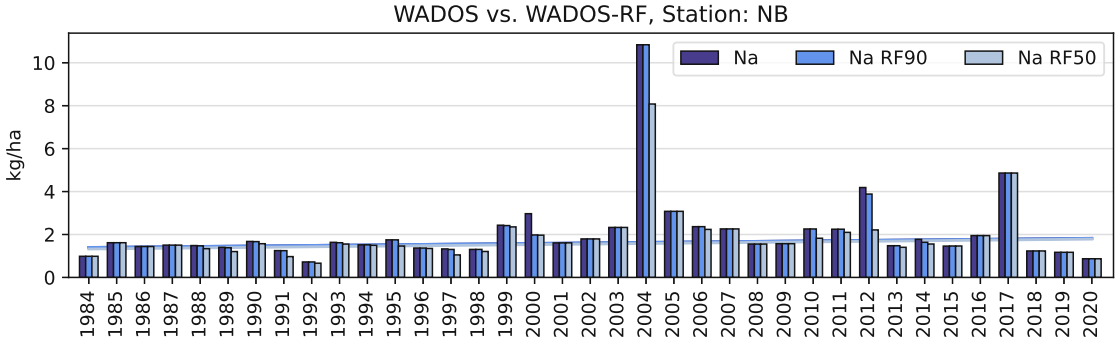
(b) Fig. B.1 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen





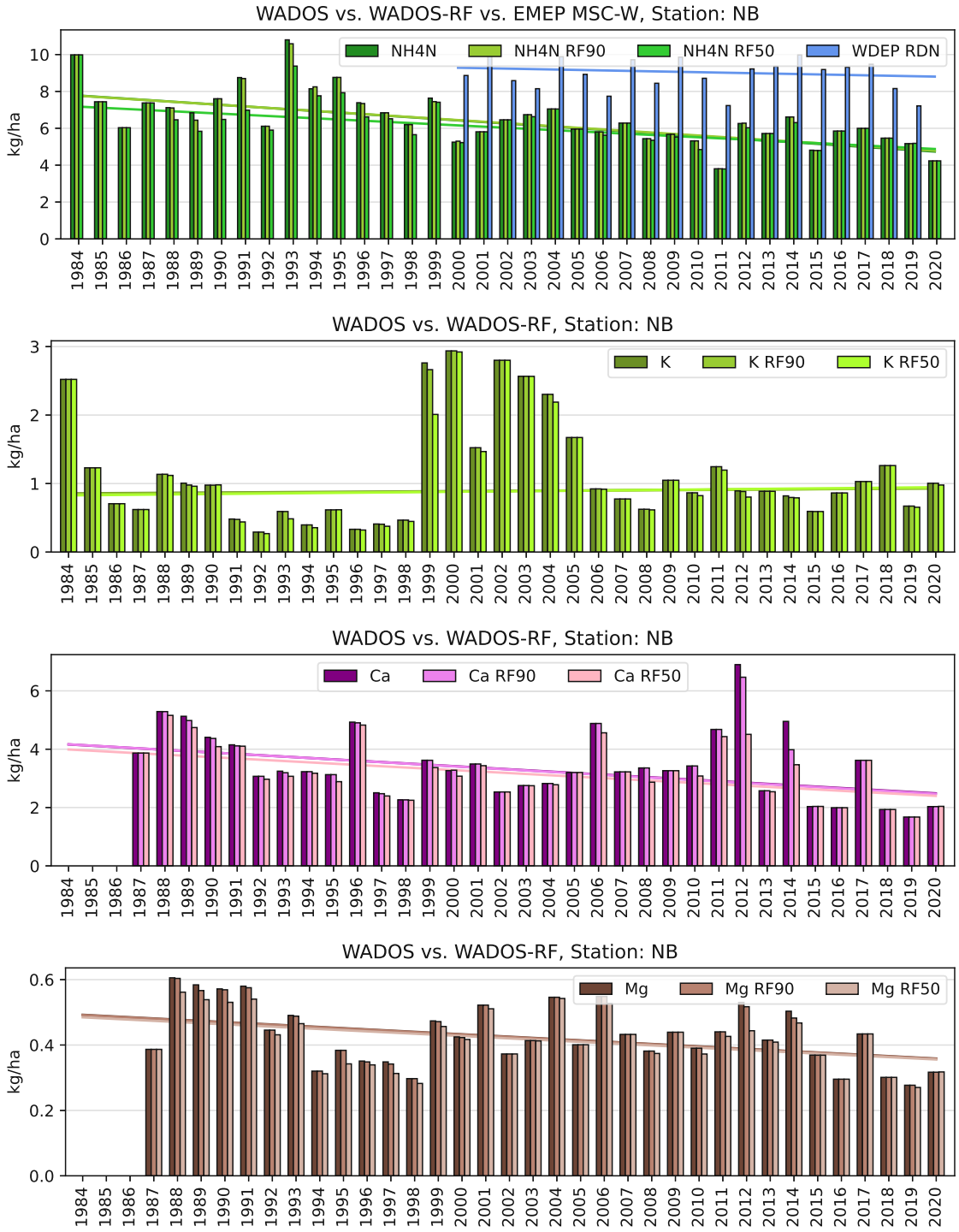
(c) Fig. B.1 (cont.): Sulfur, pH value



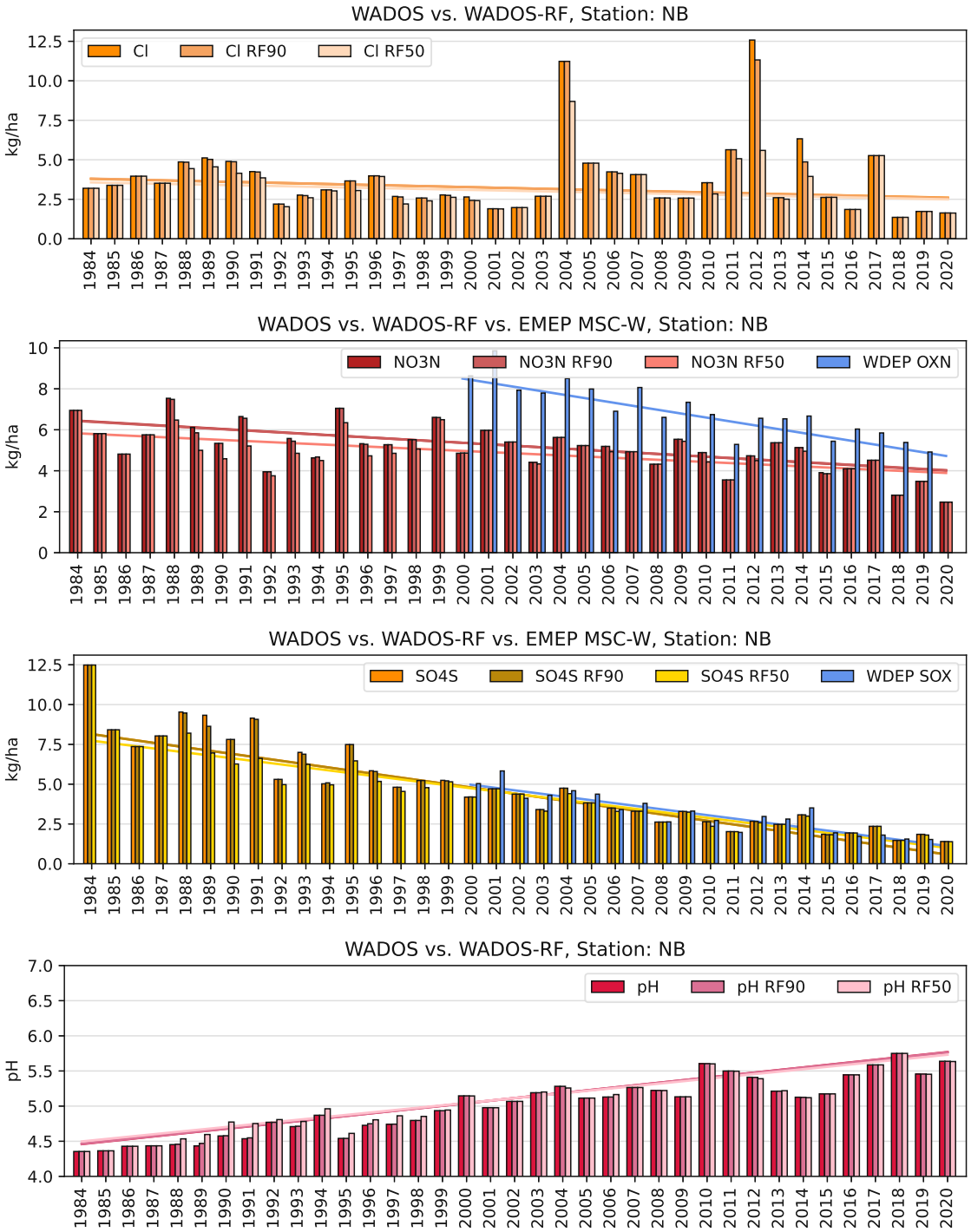


(a) Precipitation amount, sodium

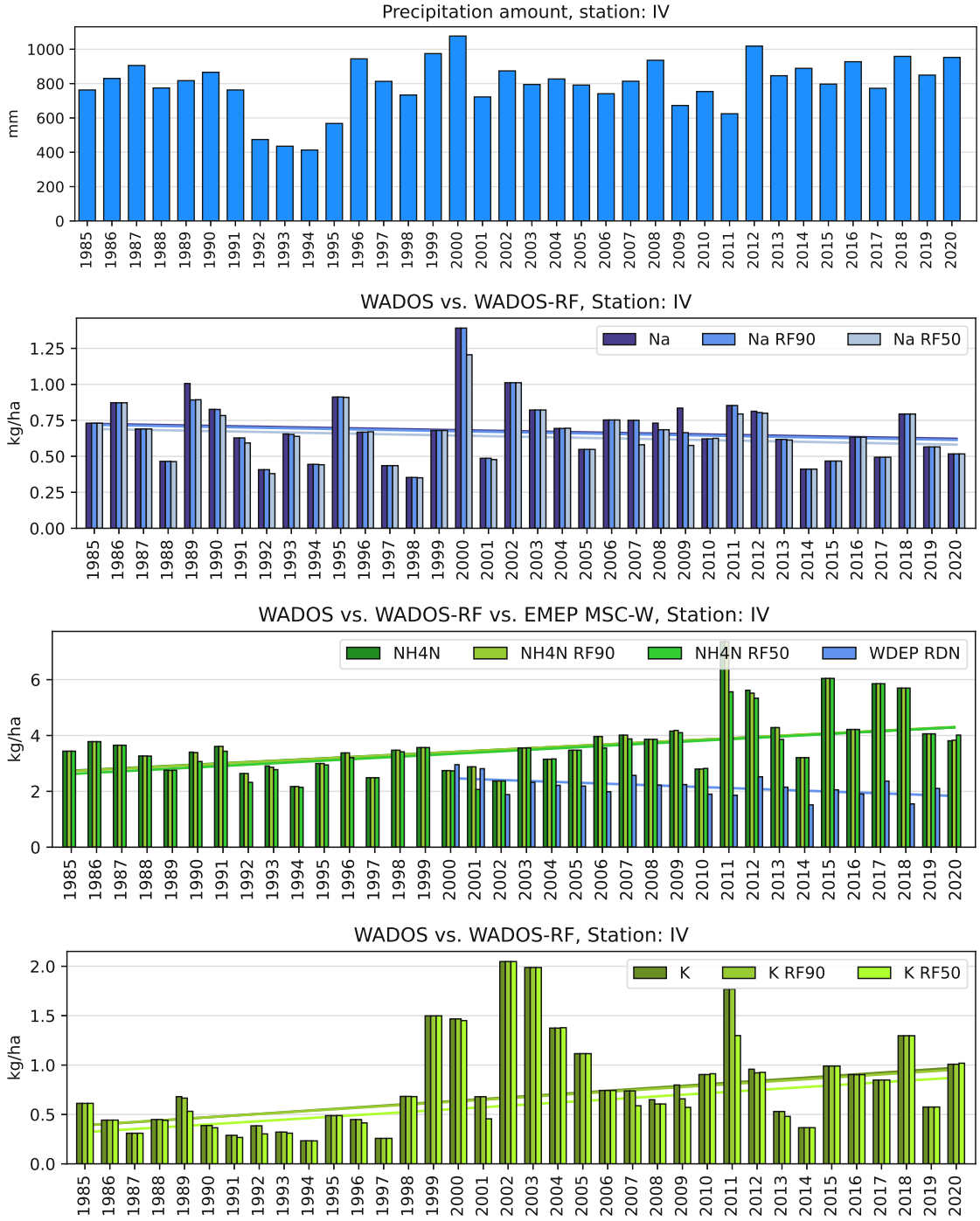
Figure B.2: Influence of RF classification on depositions at Niederndorferberg



(b) Fig. B.2 (cont.): Reduced nitrogen, potassium, calcium, magnesium

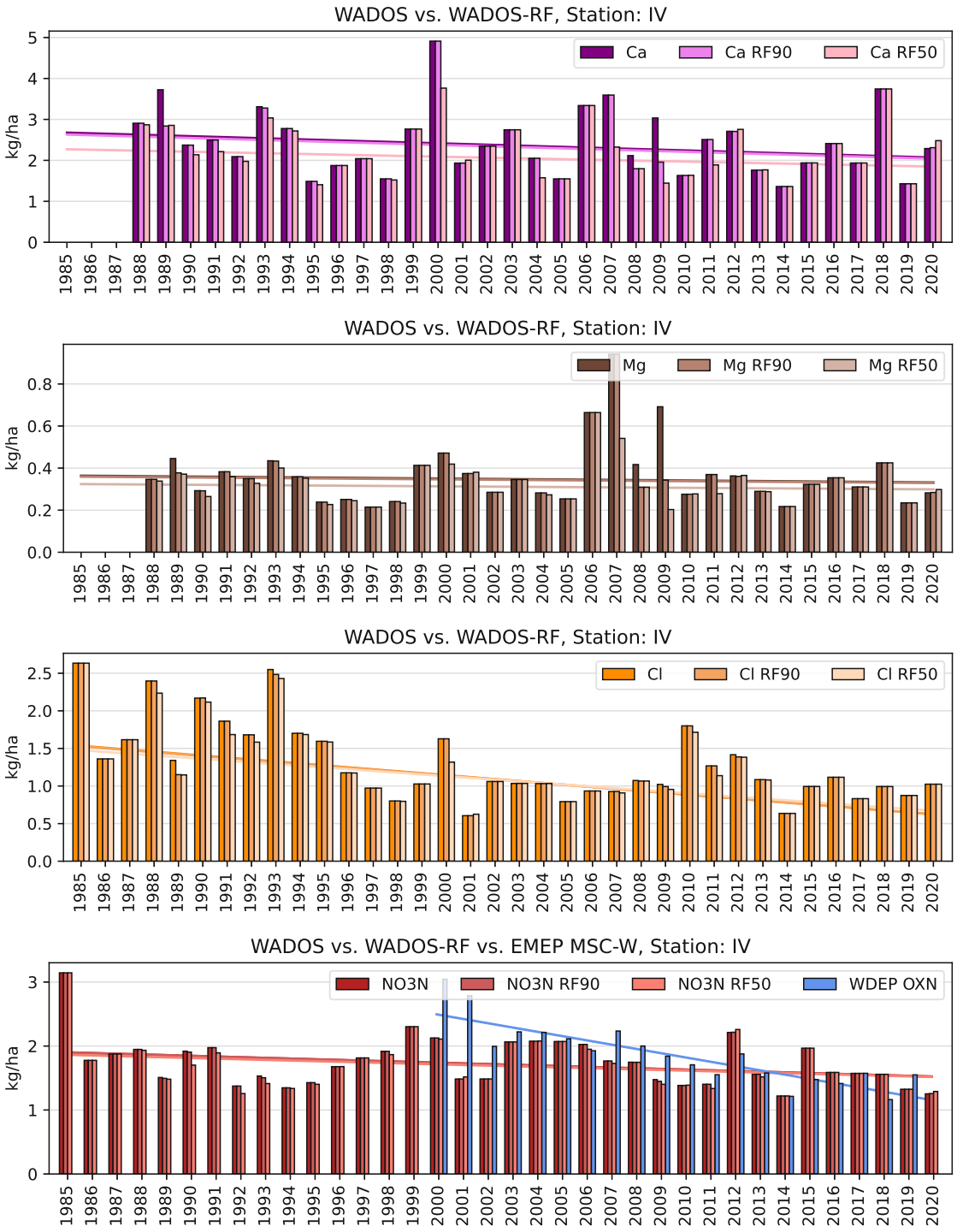


(c) Fig. B.2 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

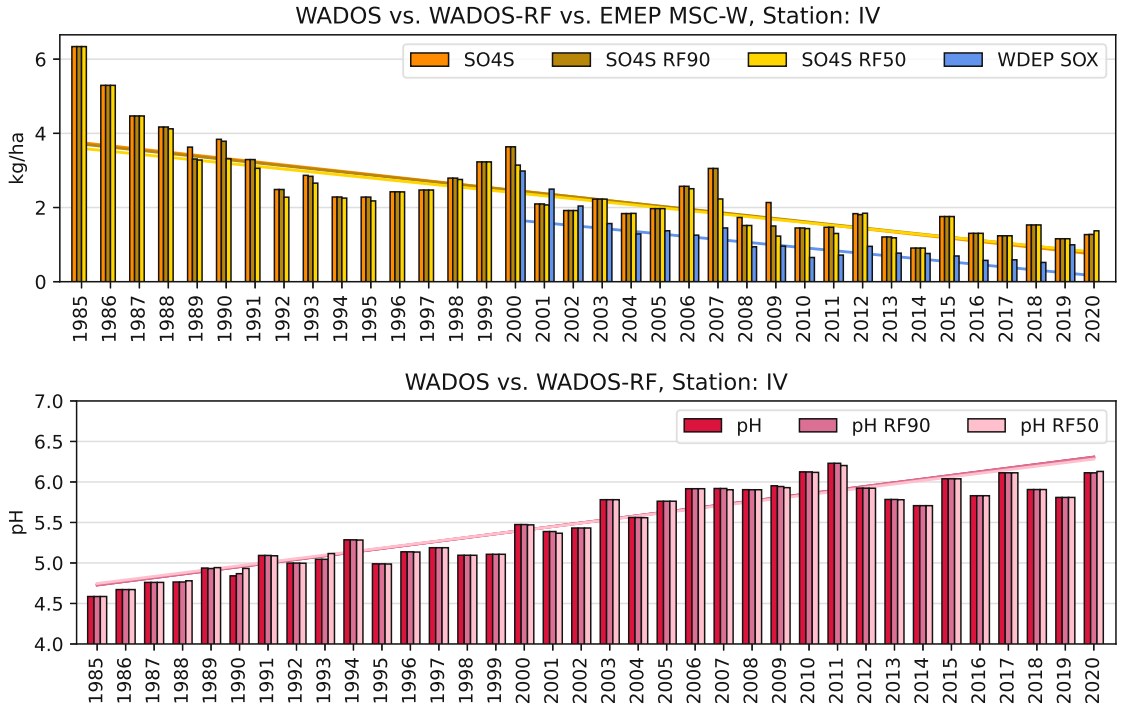


(a) Precipitation amount, sodium, reduced nitrogen, potassium

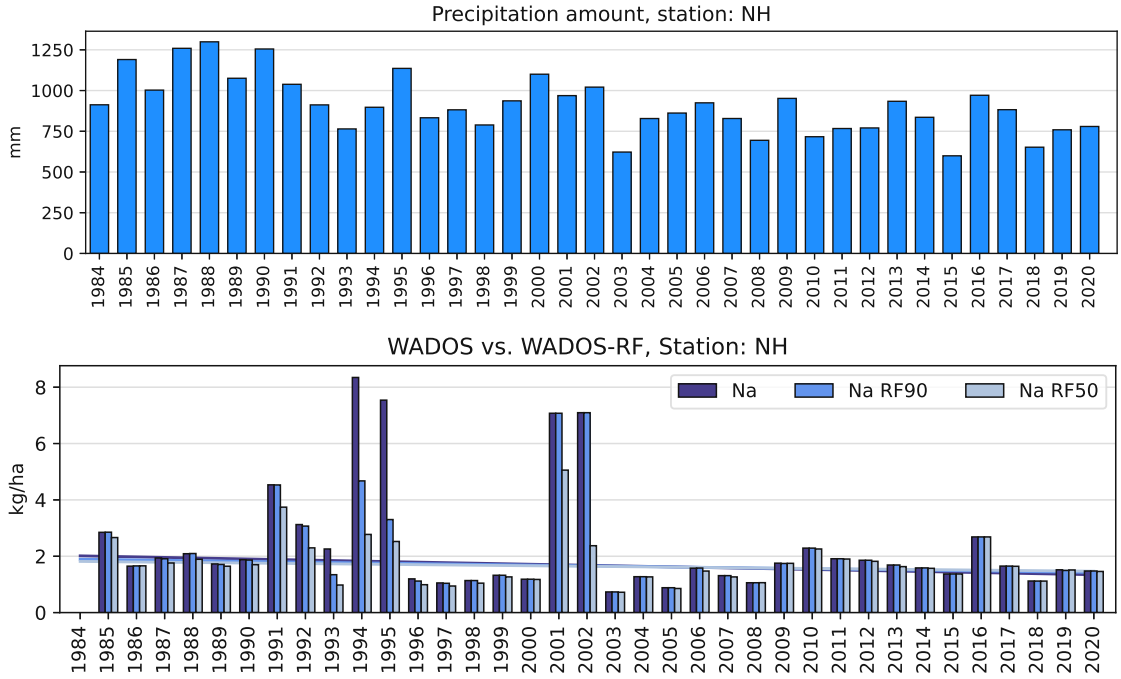
Figure B.3: Influence of RF classification on depositions in Innervillgraten



(b) Fig. B.3 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen

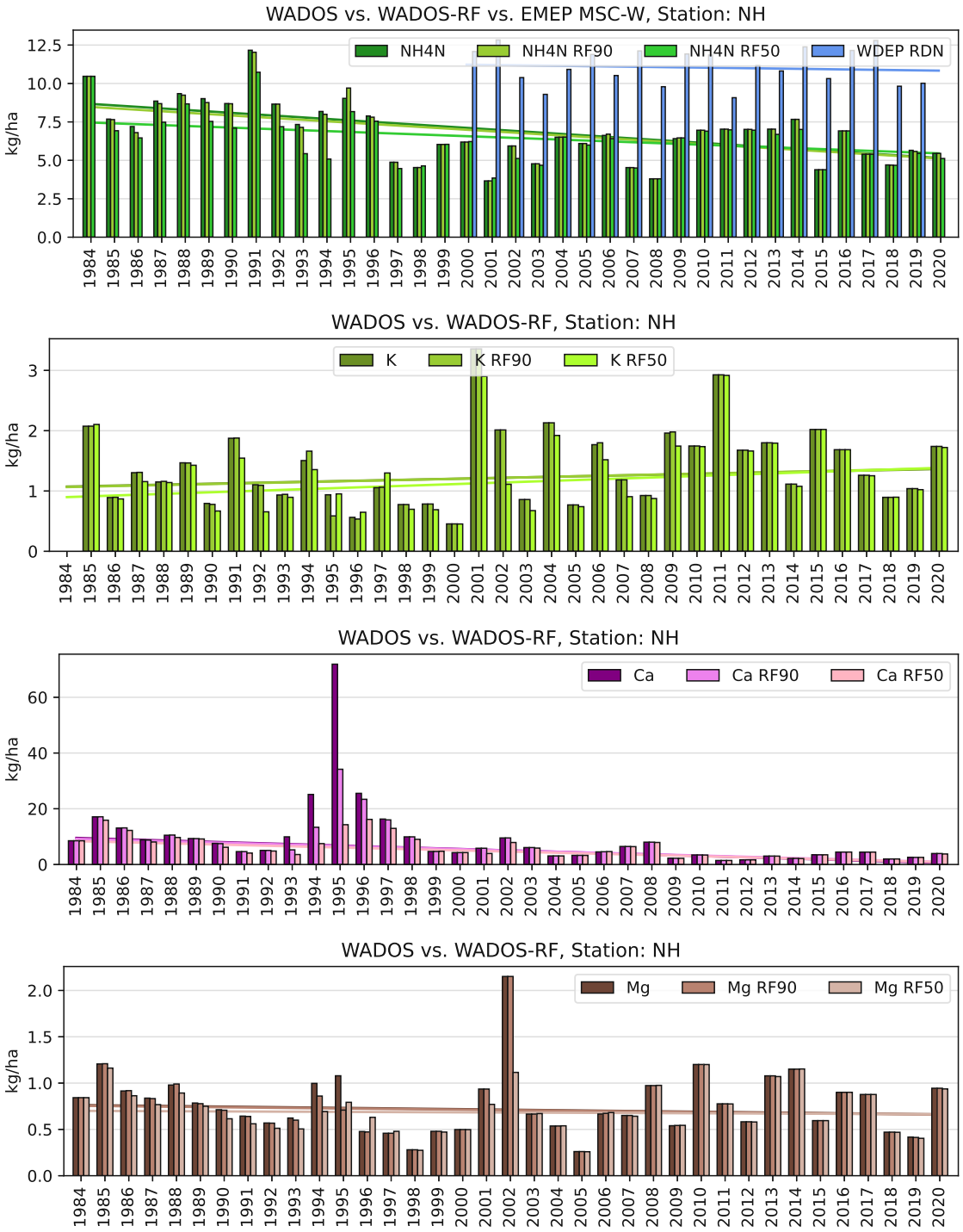


(c) Fig. B.3 (cont.): Sulfur, pH value

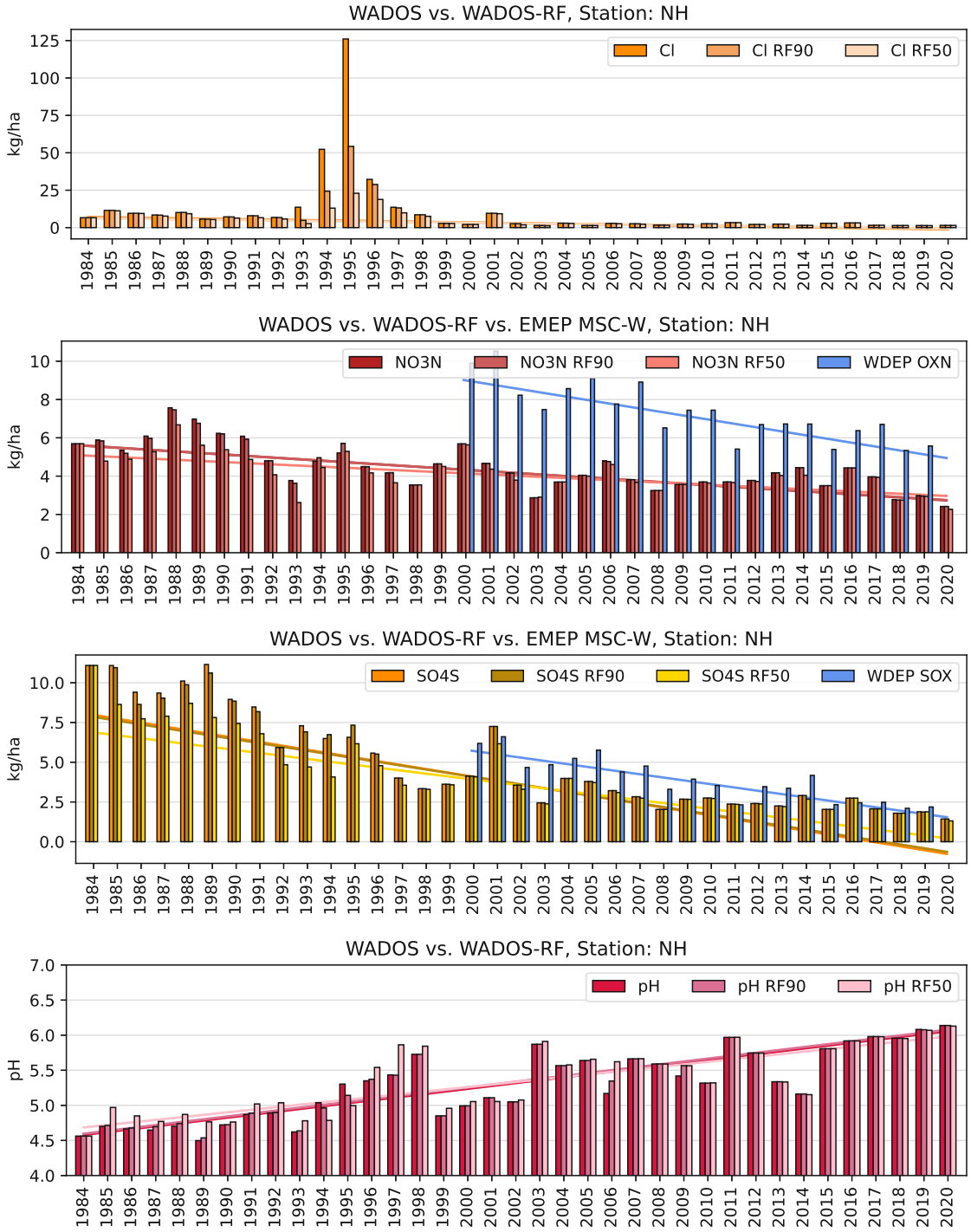


(a) Precipitation amount, sodium

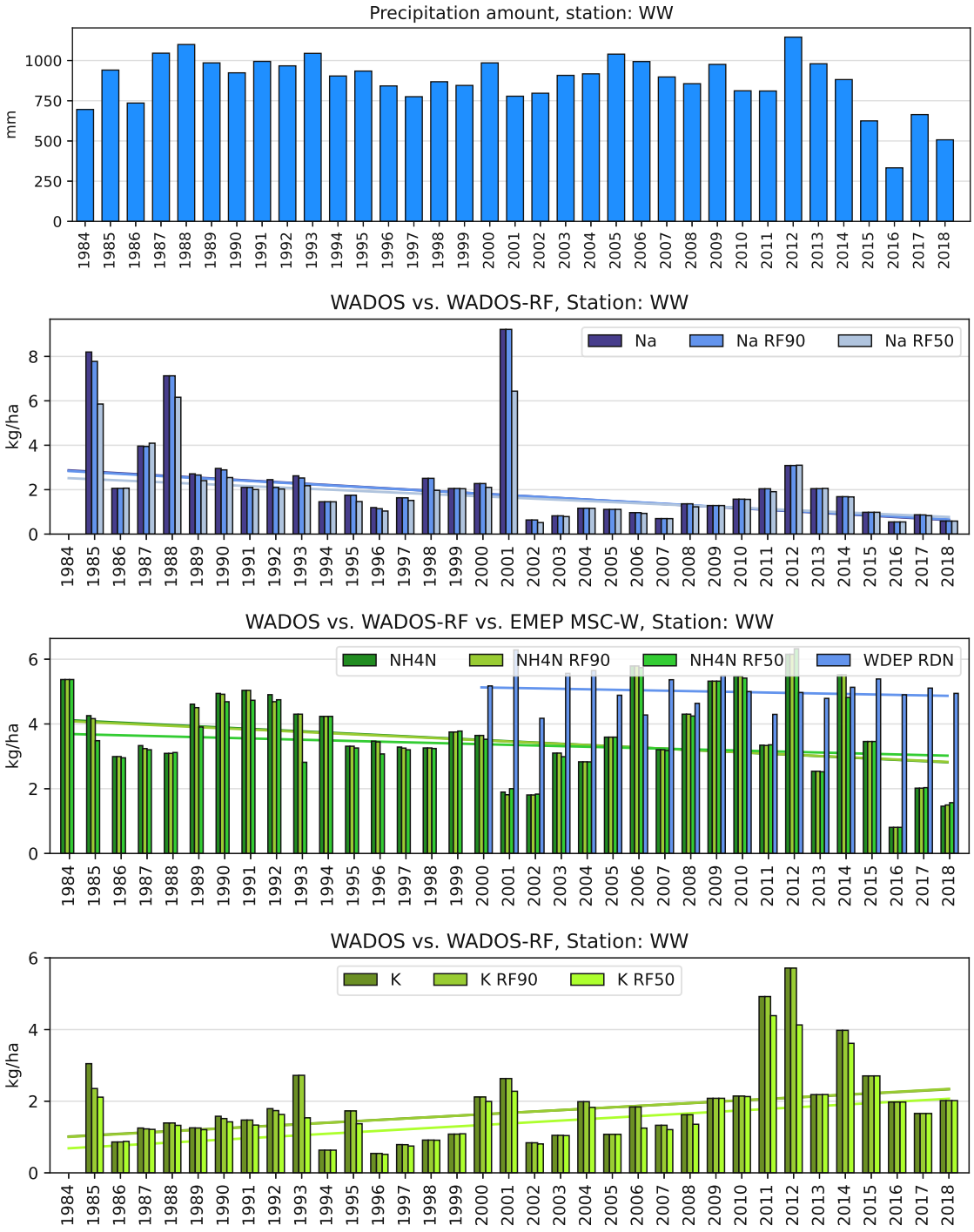
Figure B.4: Influence of RF classification on depositions at Haunsberg



(b) Fig. B.4 (cont.): Reduced nitrogen, potassium, calcium, magnesium

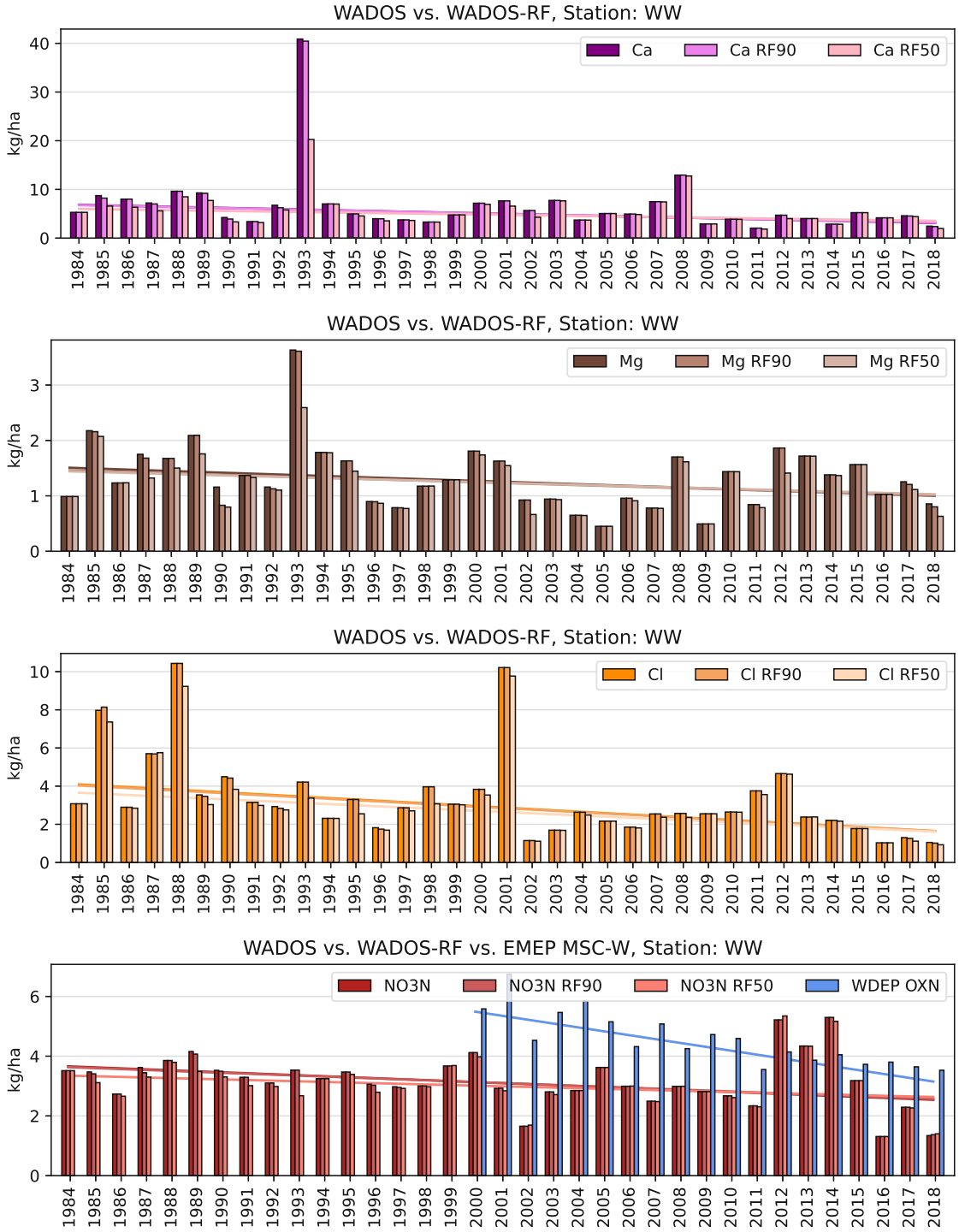


(c) Fig. B.4 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

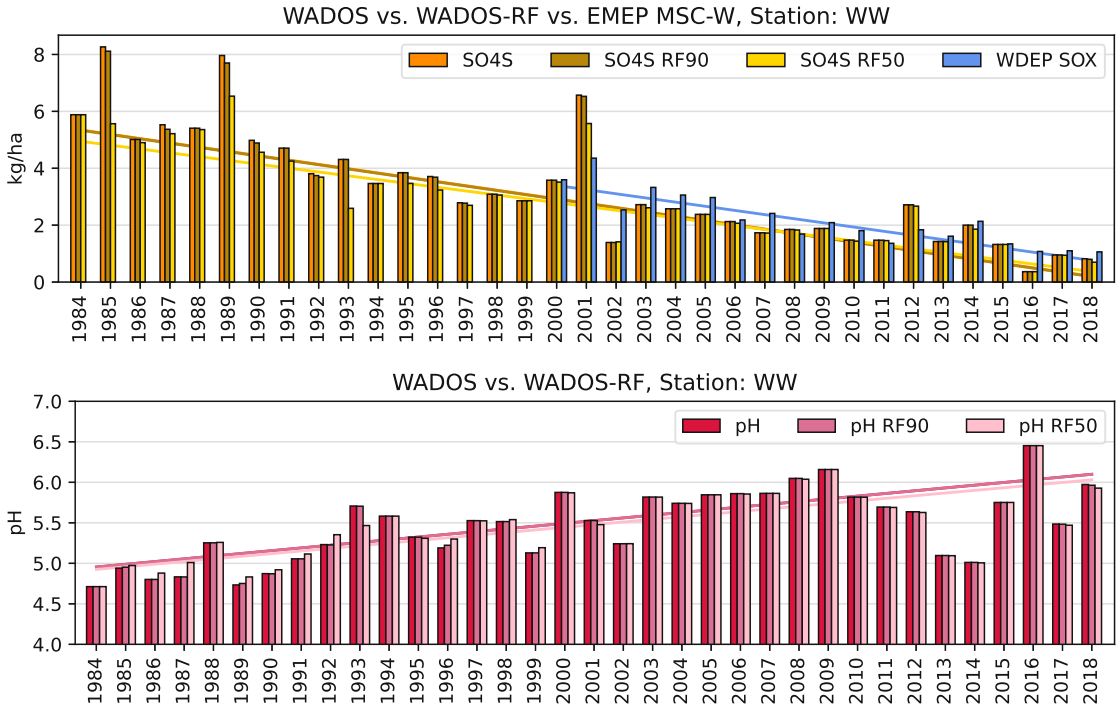


(a) Precipitation amount, sodium, reduced nitrogen, potassium

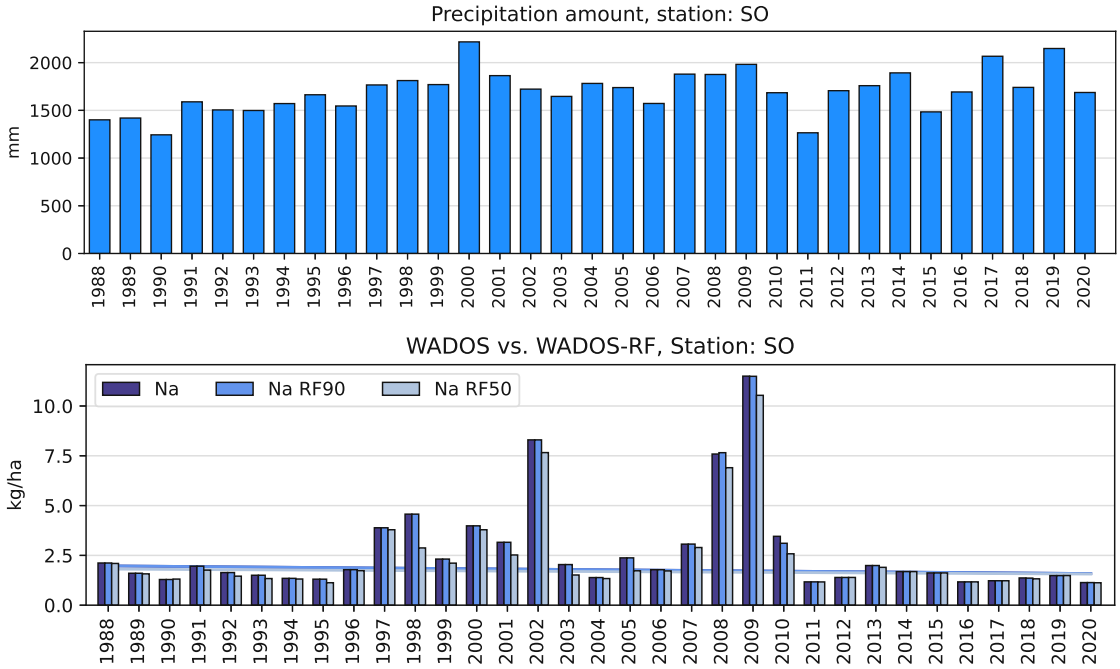
Figure B.5: Influence of RF classification on depositions in Werfenweng



(b) Fig. B.5 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen

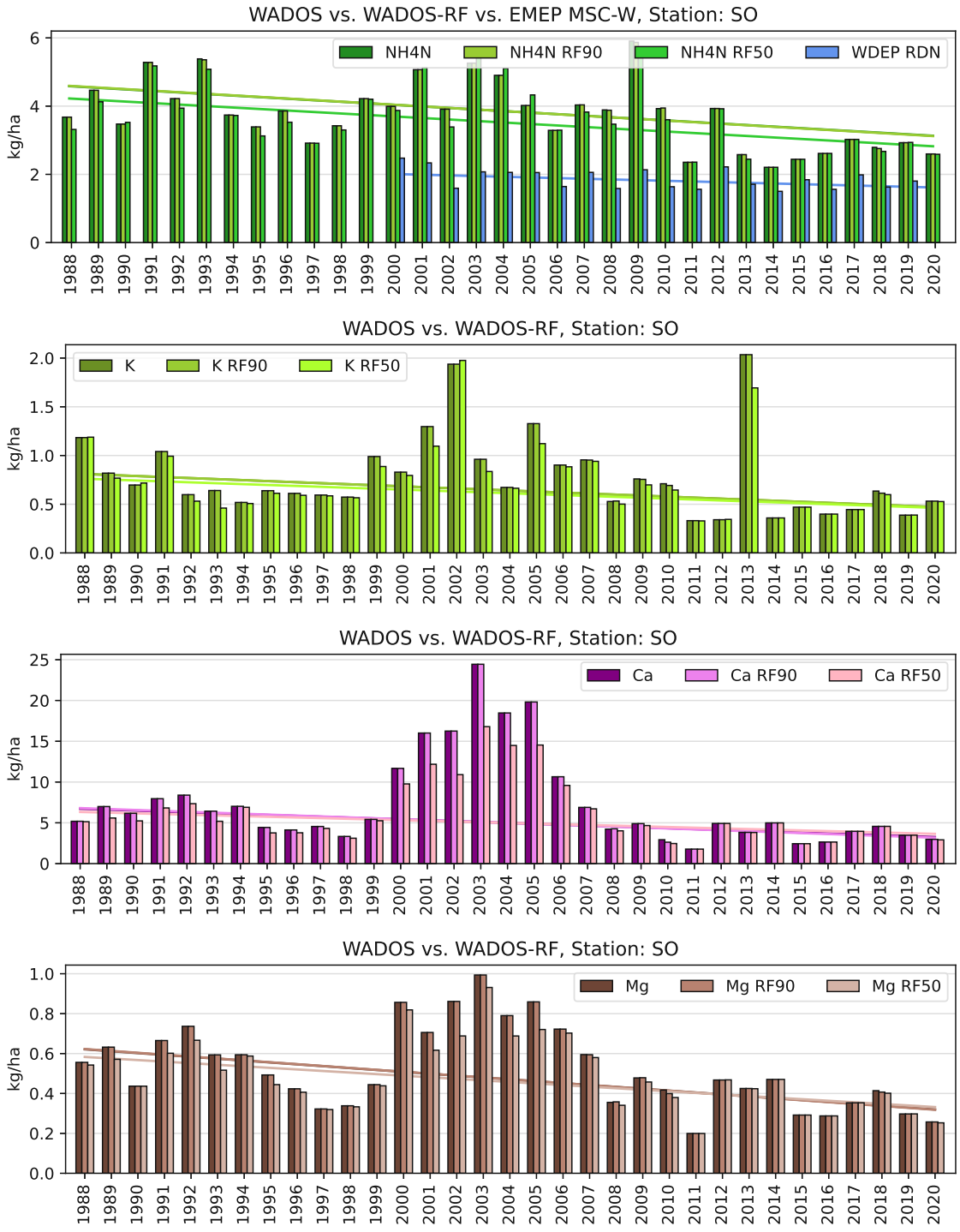


(c) Fig. B.5 (cont.): Sulfur, pH value

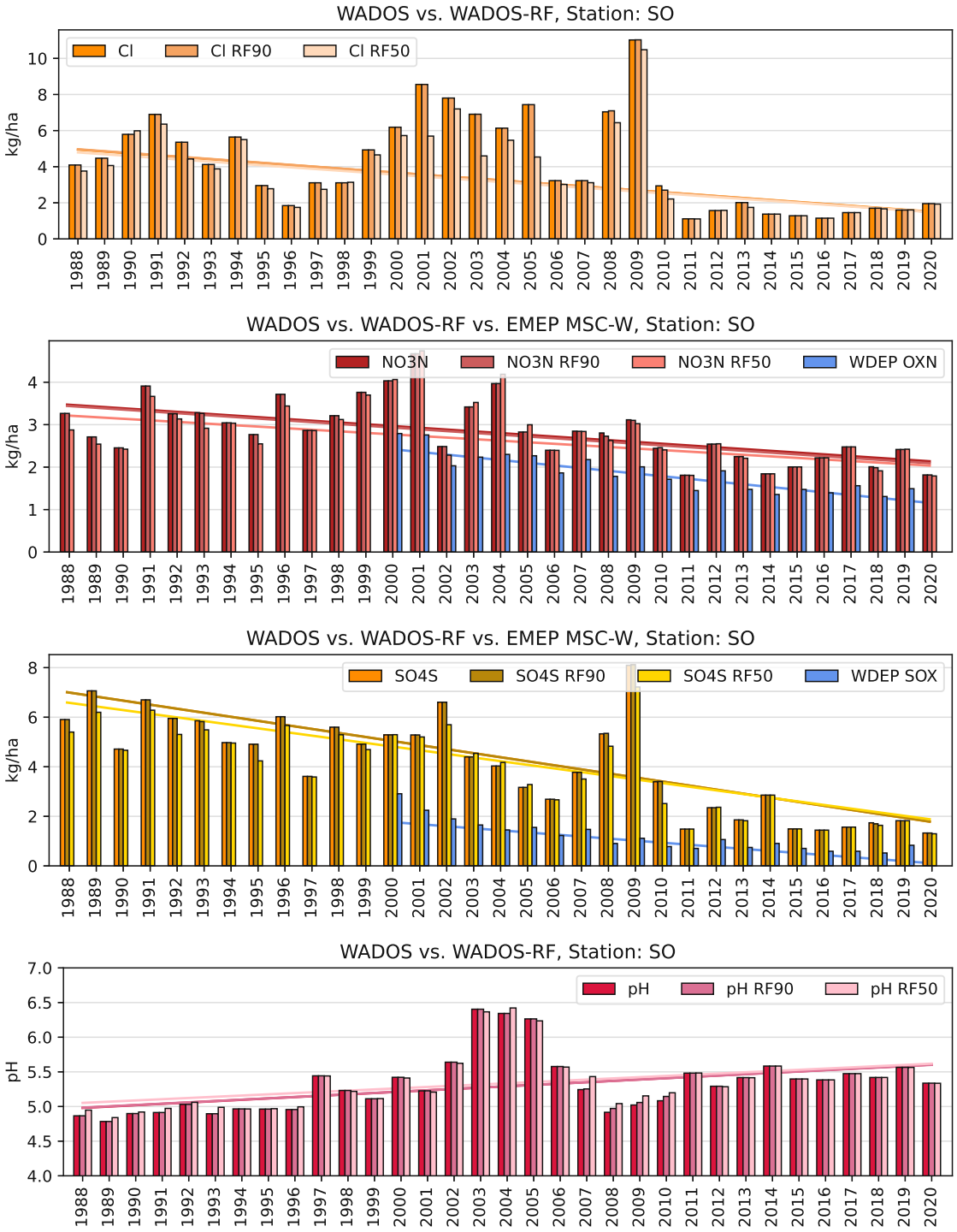


(a) Precipitation amount, sodium

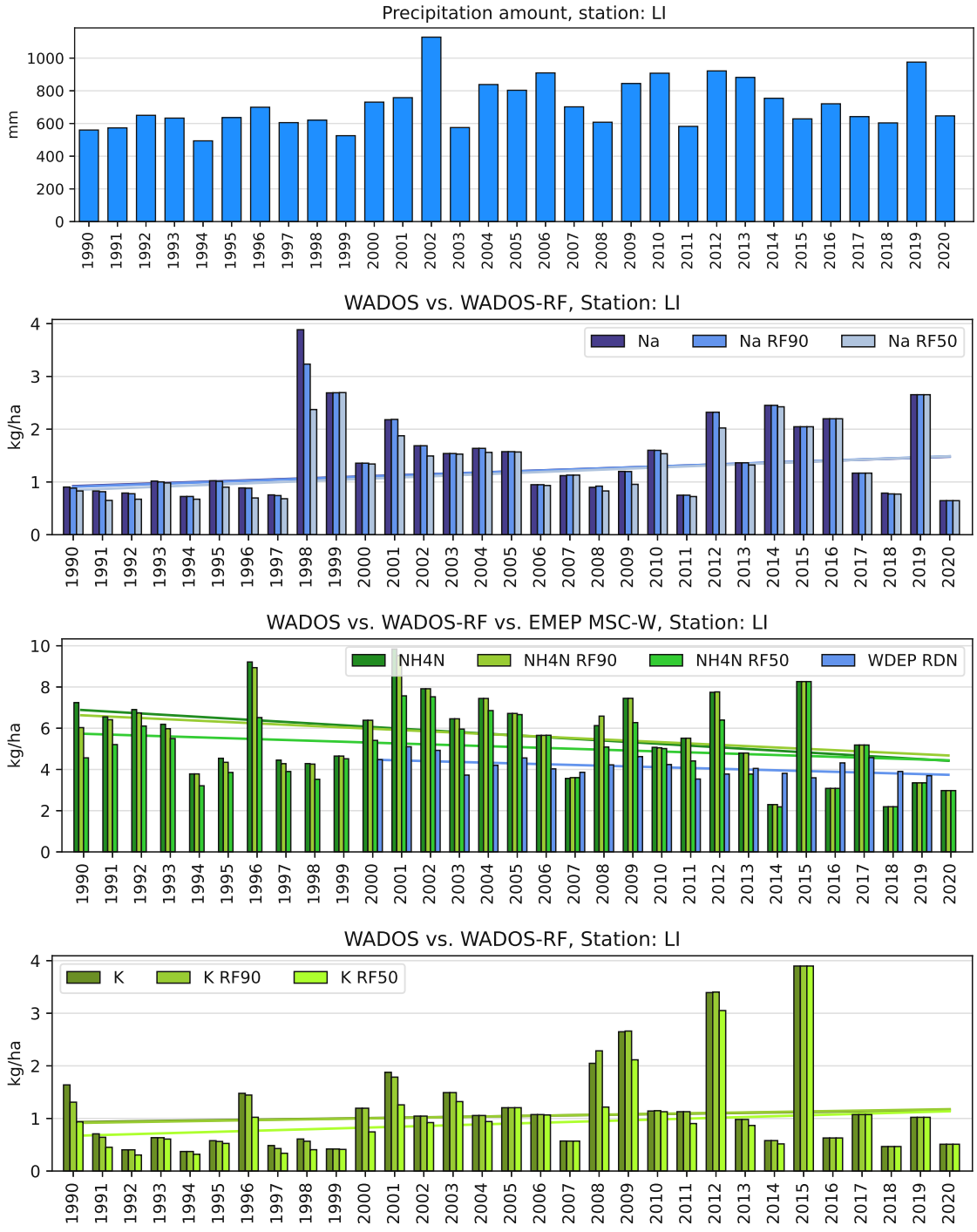
Figure B.6: Influence of RF classification on depositions at Sonnblick



(b) Fig. B.6 (cont.): Reduced nitrogen, potassium, calcium, magnesium

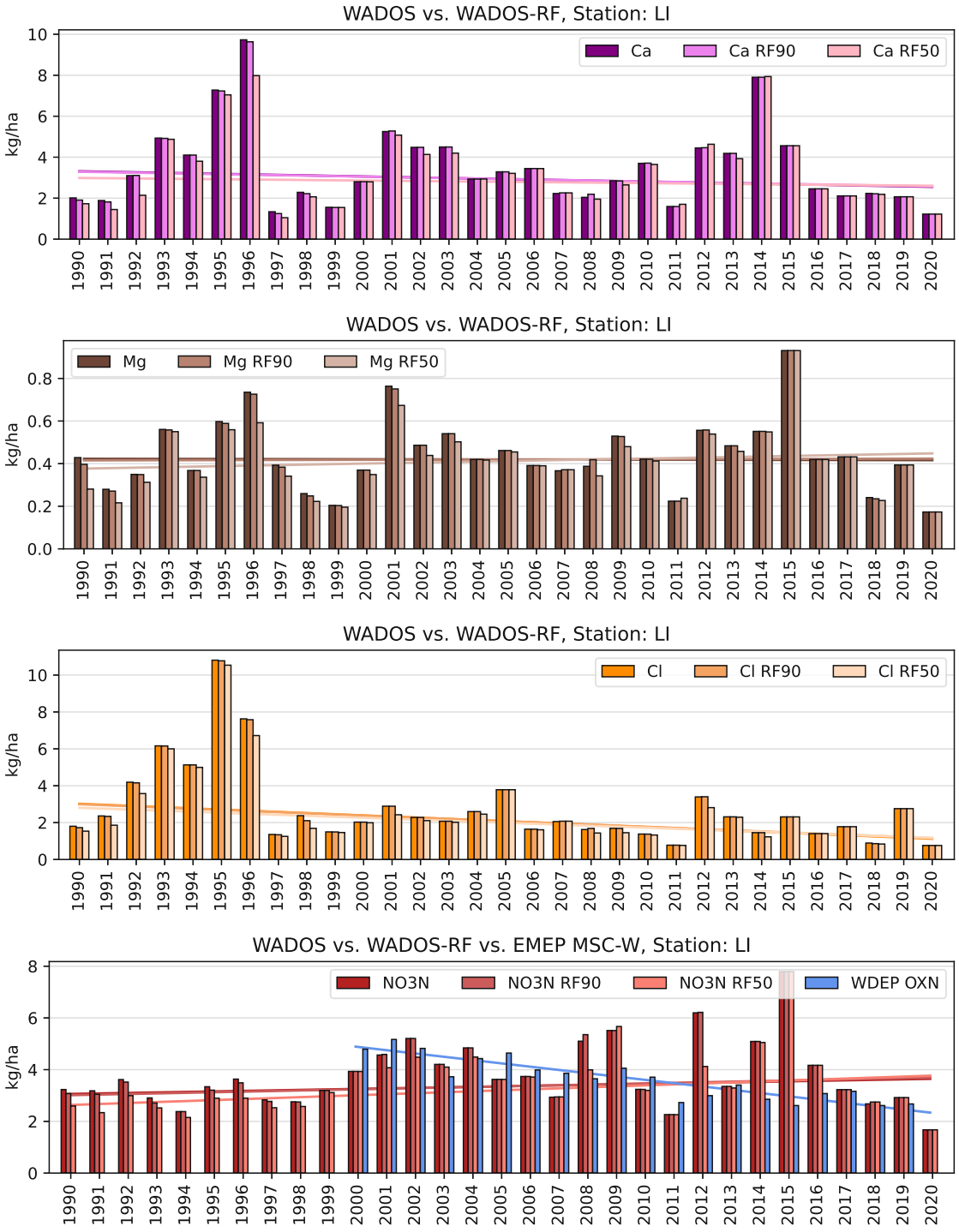


(c) Fig. B.6 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

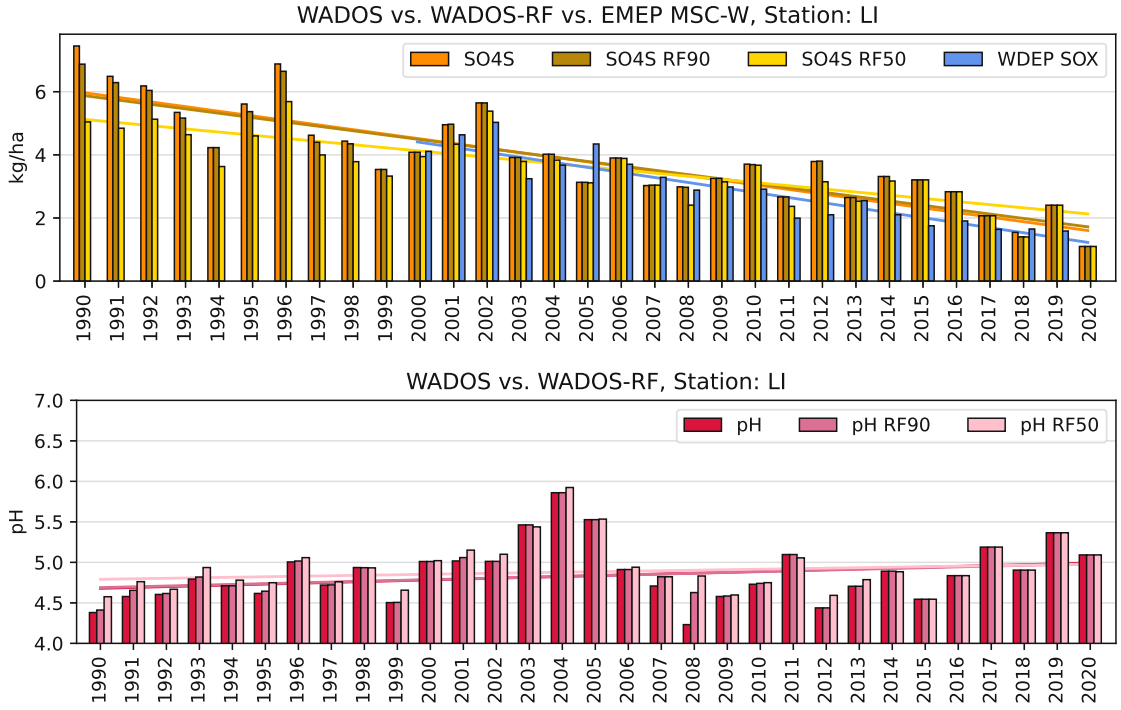


(a) Precipitation amount, sodium, reduced nitrogen, potassium

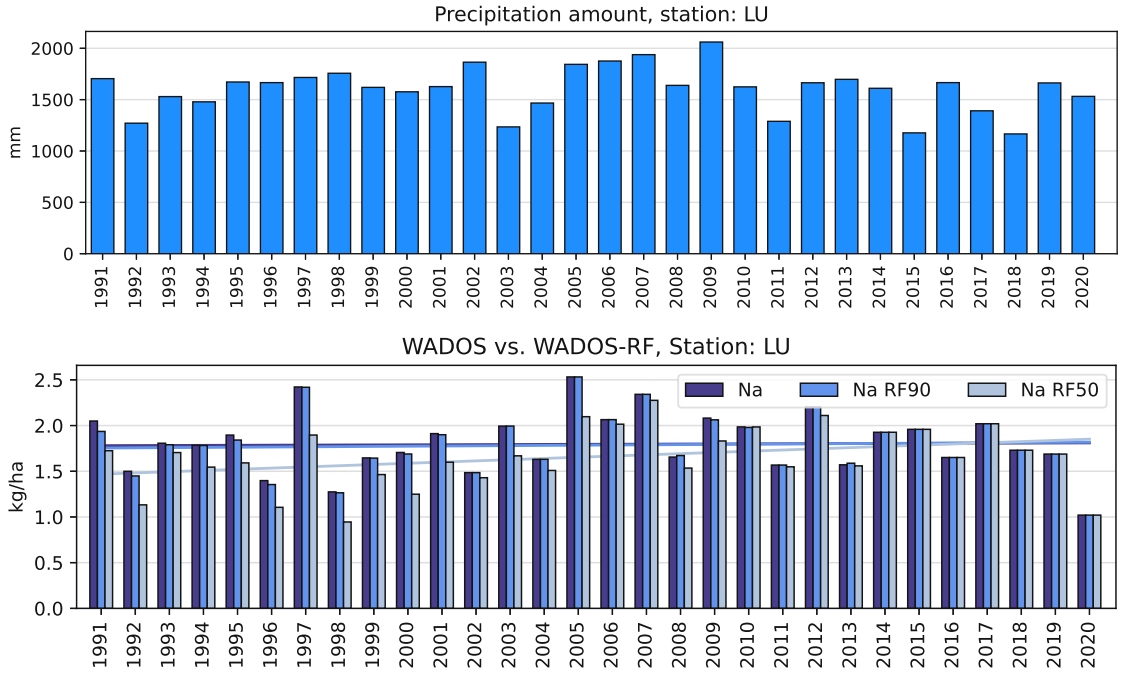
Figure B.7: Influence of RF classification on depositions in Litschau



(b) Fig. B.7 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen

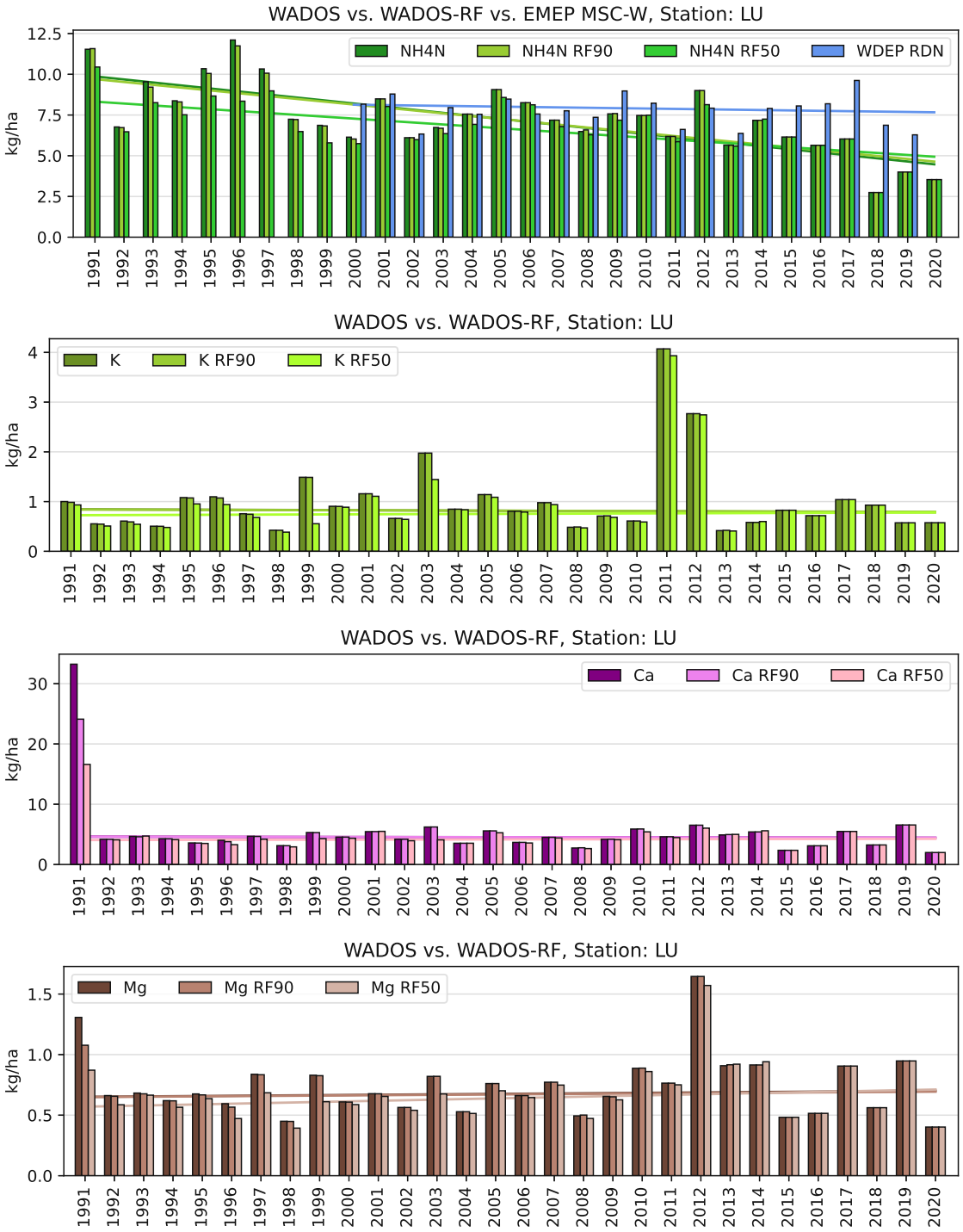


(c) Fig. B.7 (cont.): Sulfur, pH value

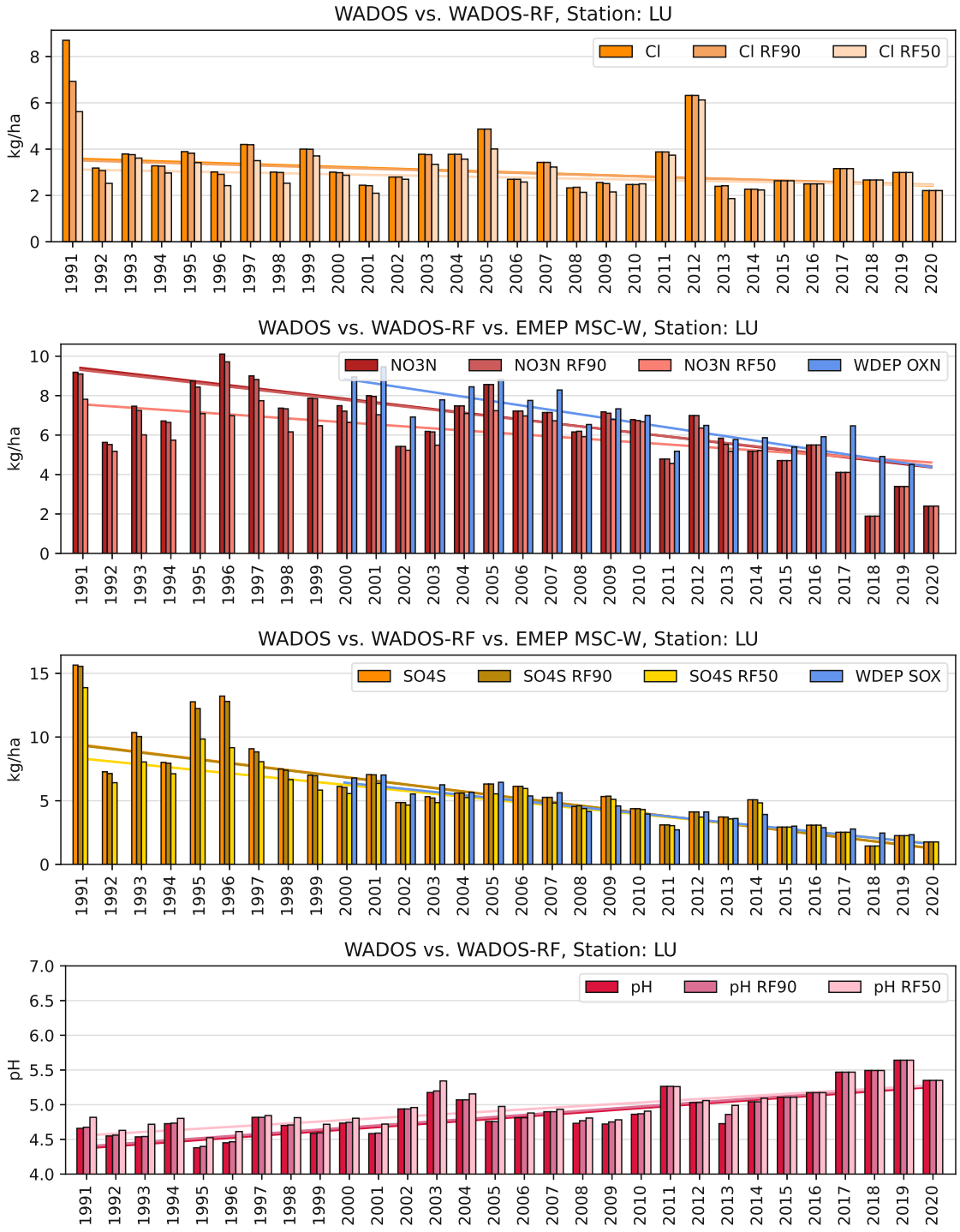


(a) Precipitation amount, sodium

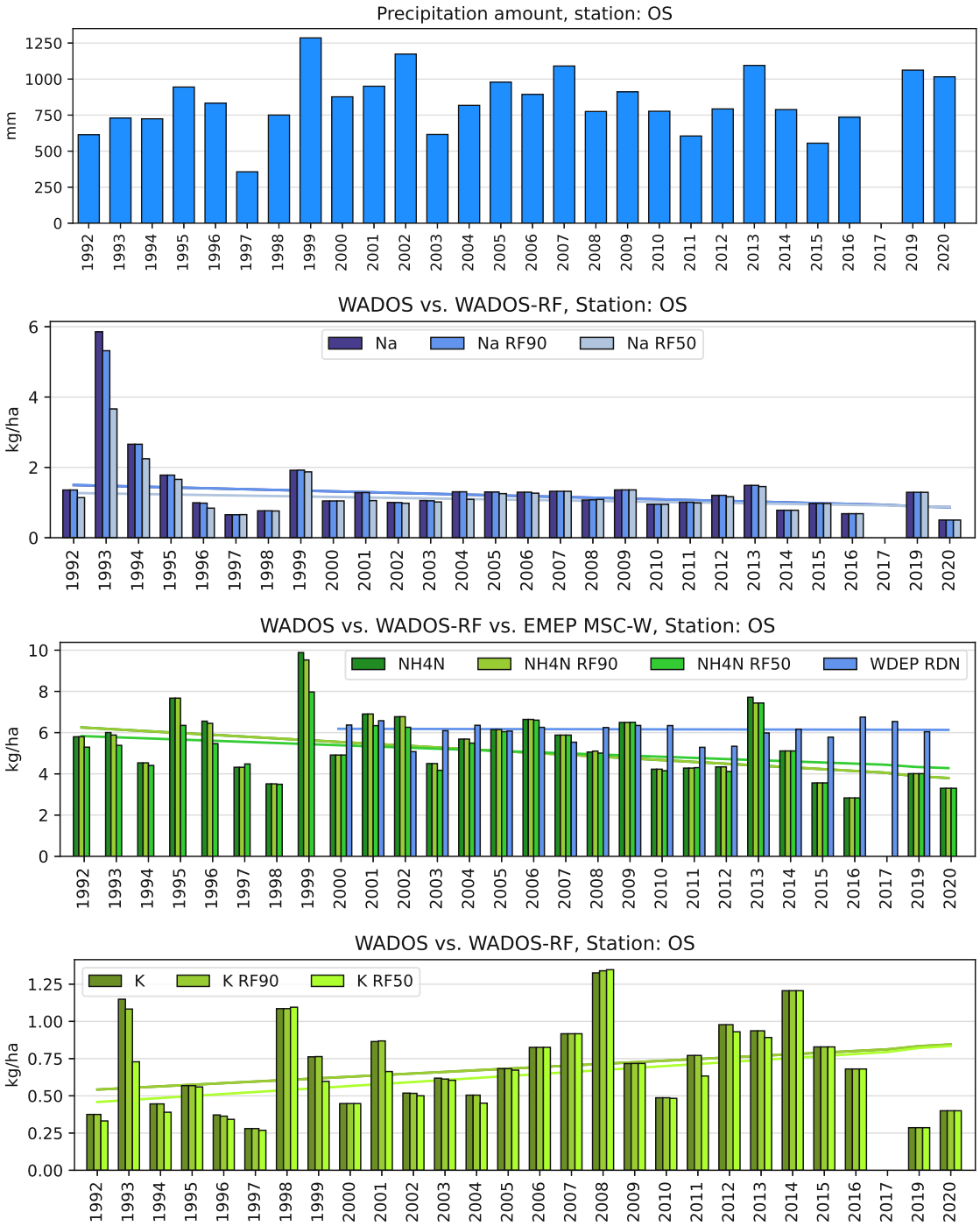
Figure B.8: Influence of RF classification on depositions in Lunz



(b) Fig. B.8 (cont.): Reduced nitrogen, potassium, calcium, magnesium



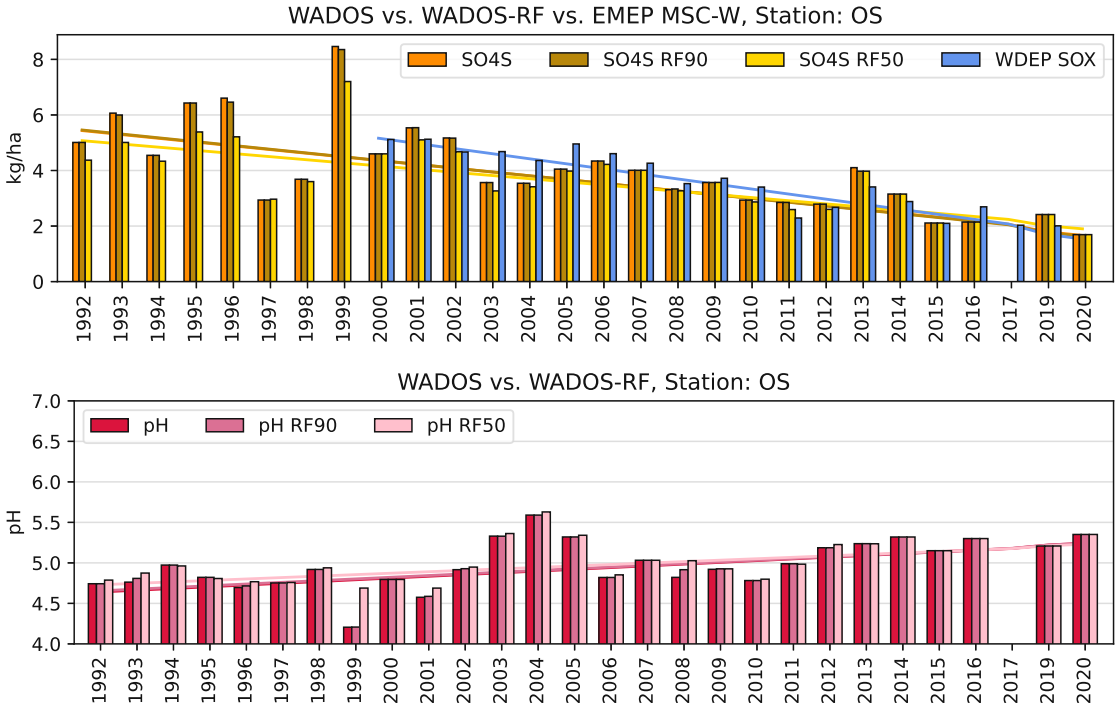
(c) Fig. B.8 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value



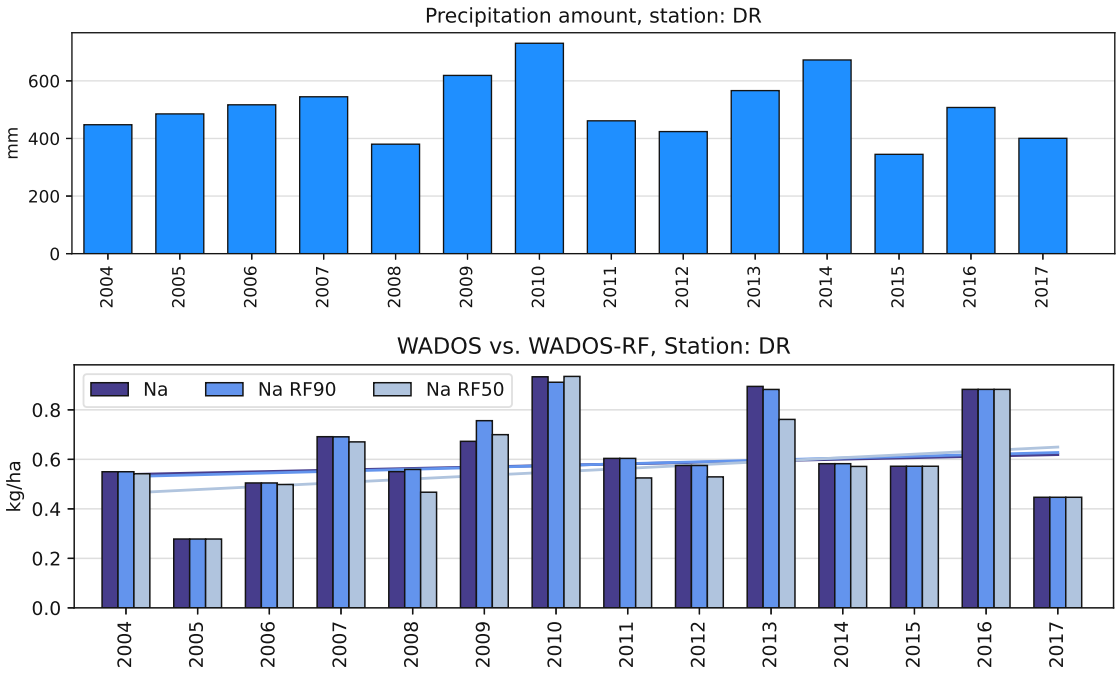
(a) Precipitation amount, sodium, reduced nitrogen, potassium

Figure B.9: Influence of RF classification on depositions at Ostrong

(b) Fig. B.9 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen

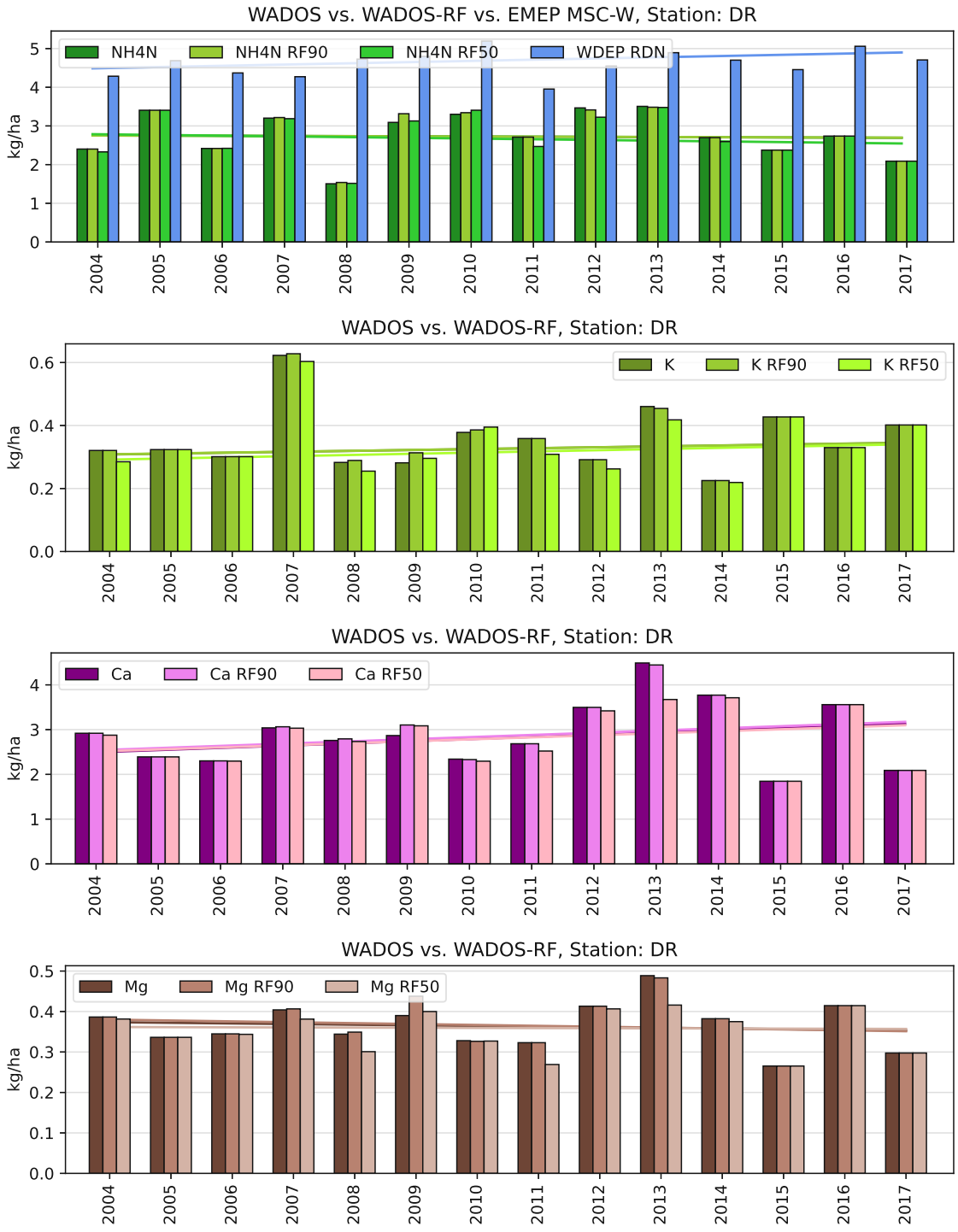


(c) Fig. B.9 (cont.): Sulfur, pH value



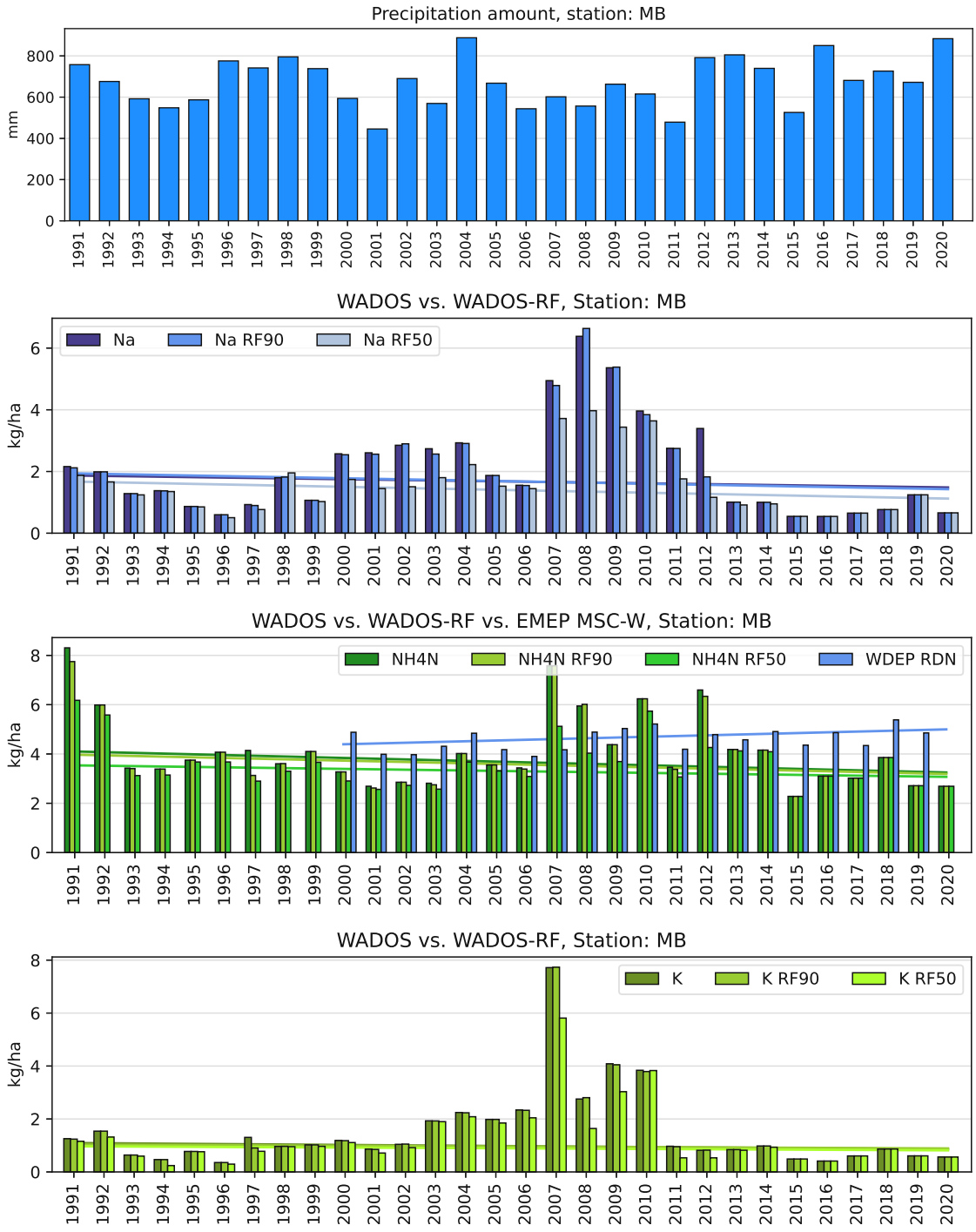
(a) Precipitation amount, sodium

Figure B.10: Influence of RF classification on depositions in Drasenhofen



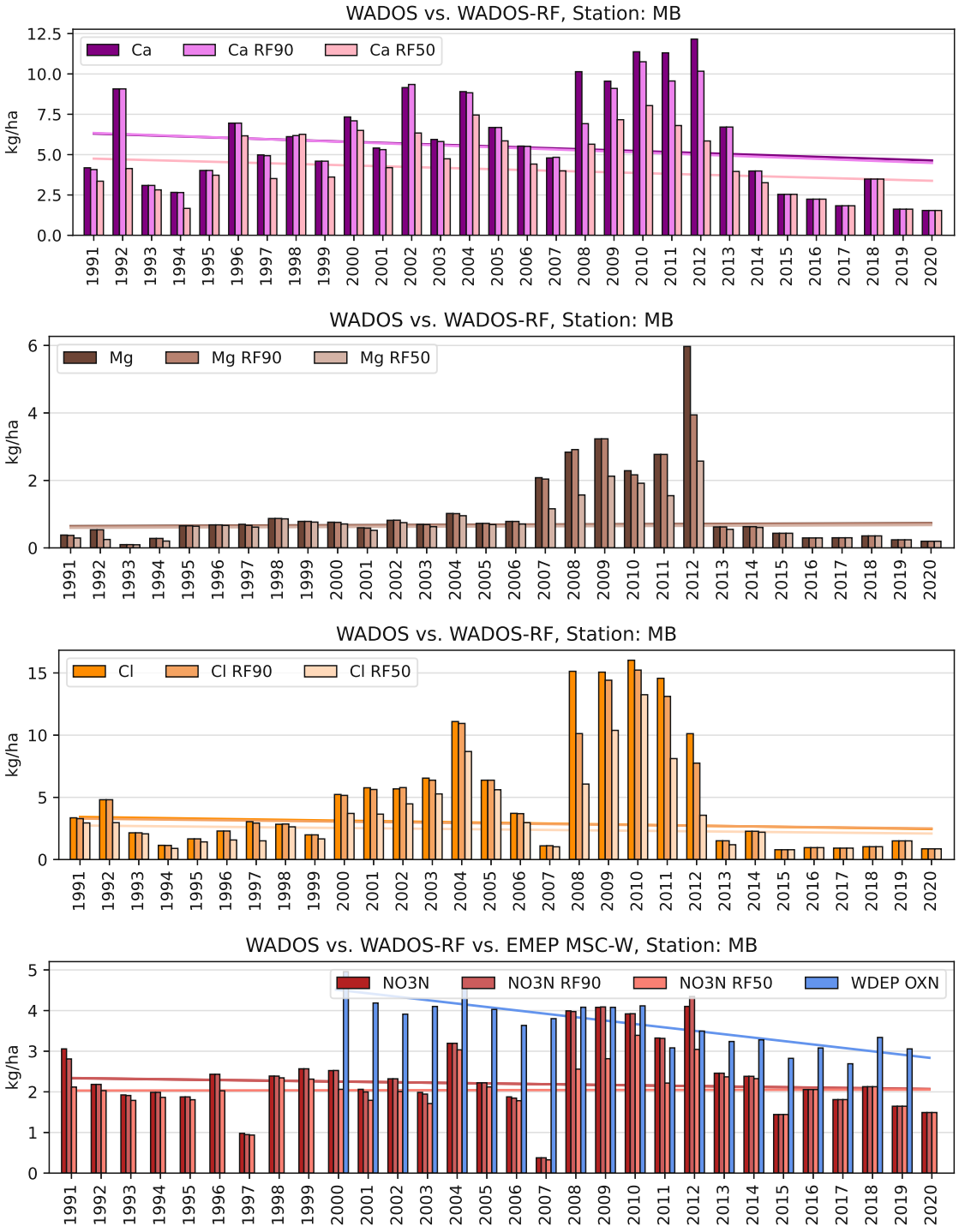
(b) Fig. B.10 (cont.): Reduced nitrogen, potassium, calcium, magnesium

(c) Fig. B.10 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

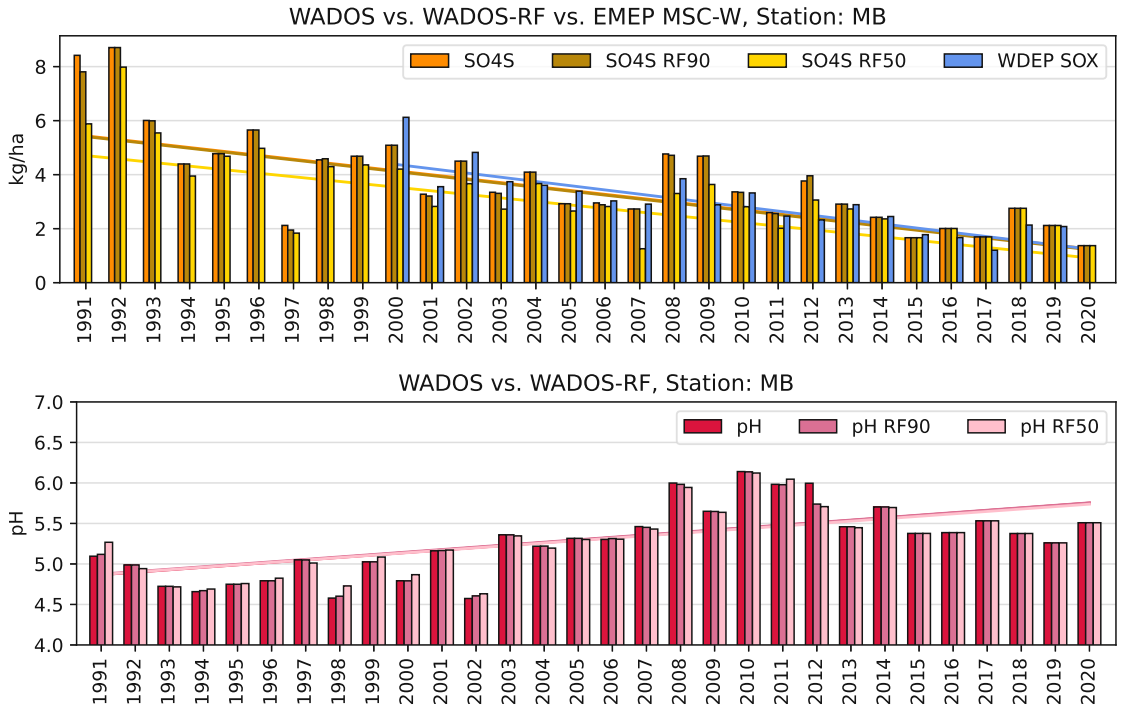


(a) Precipitation amount, sodium, reduced nitrogen, potassium

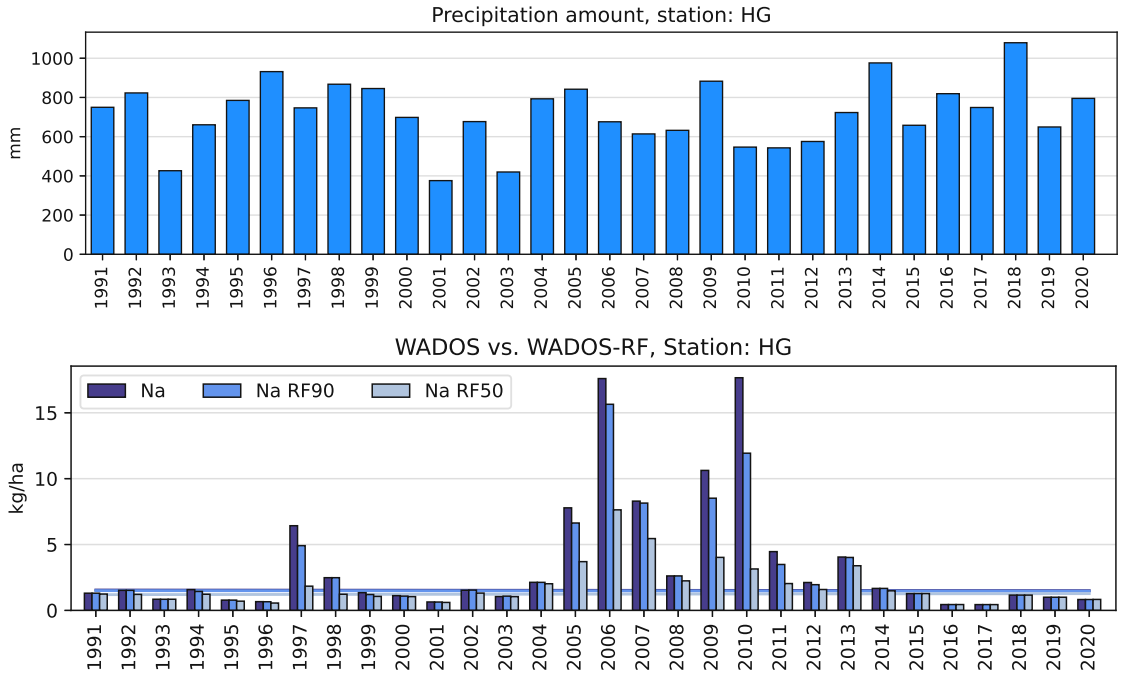
Figure B.11: Influence of RF classification on depositions at Masenberg



(b) Fig. B.11 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen

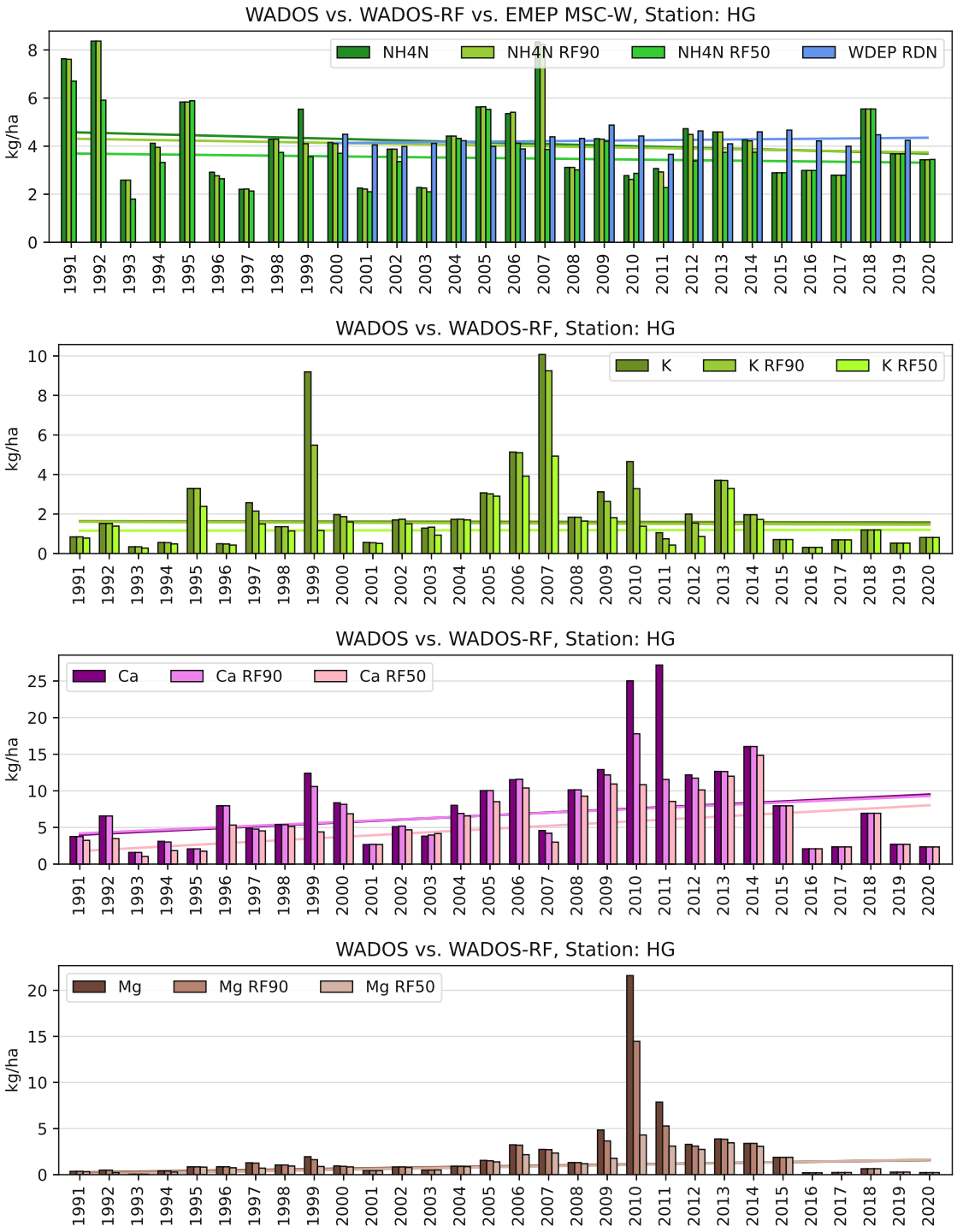


(c) Fig. B.11 (cont.): Sulfur, pH value

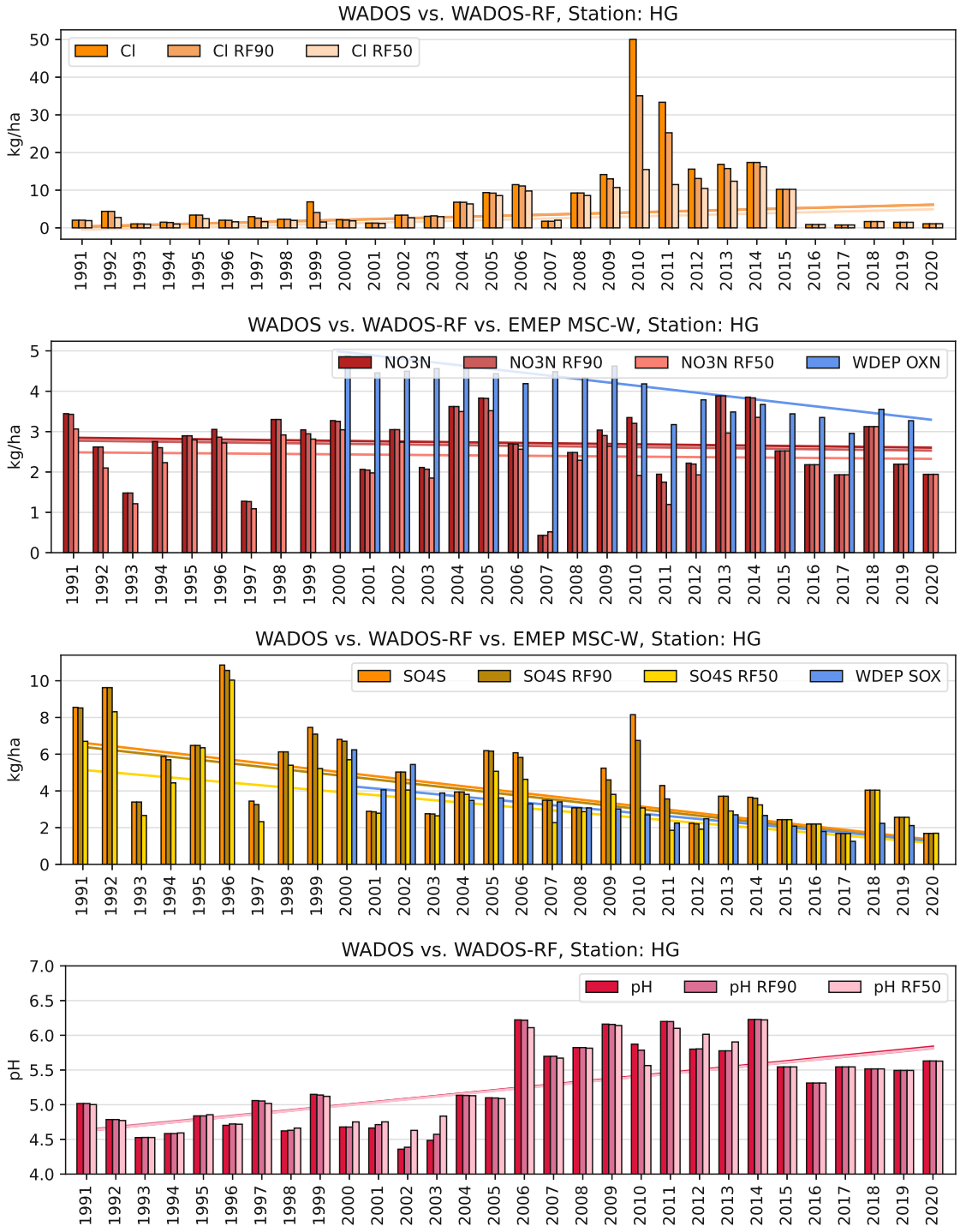


(a) Precipitation amount, sodium

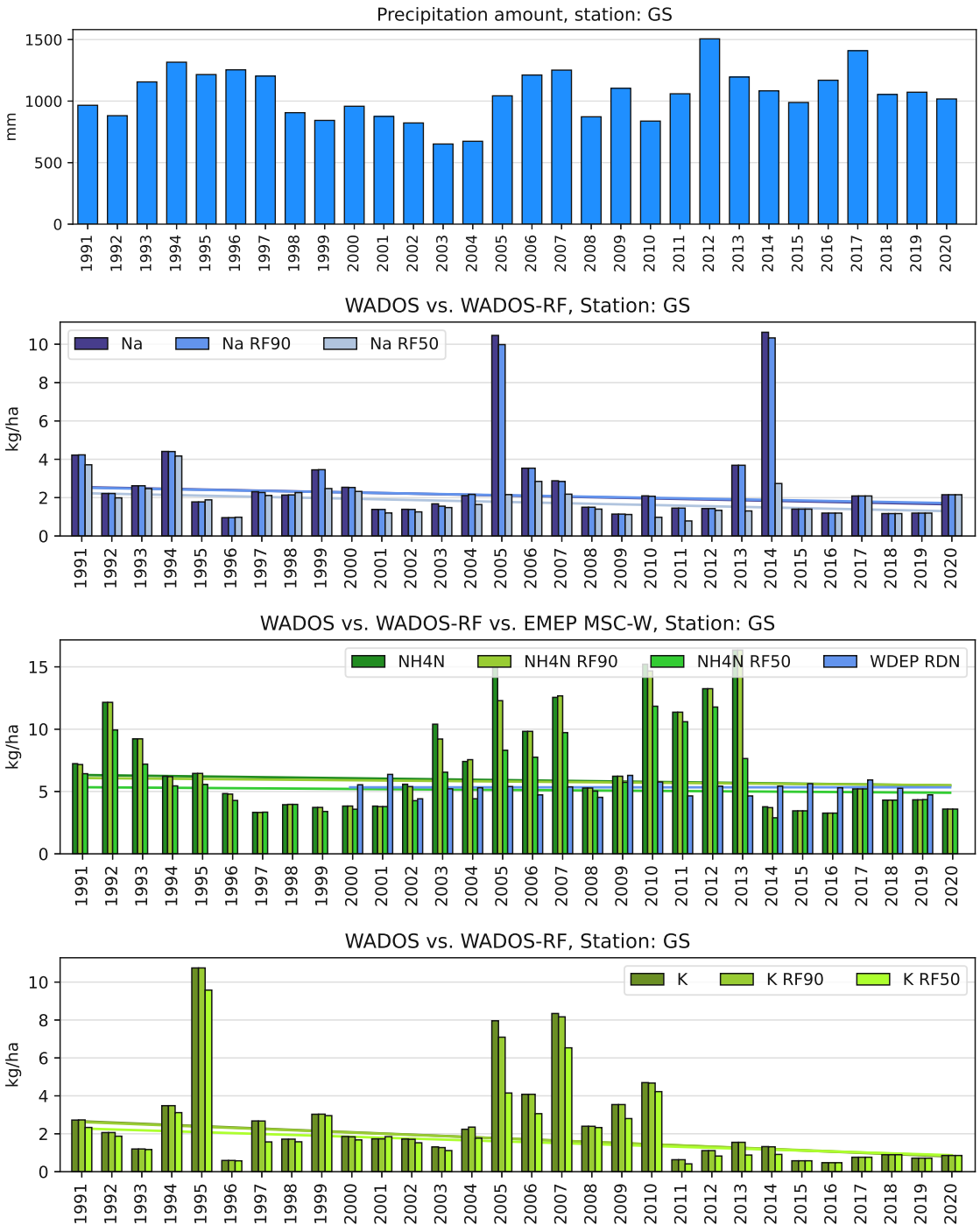
Figure B.12: Influence of RF classification on depositions in Hochgößnitz



(b) Fig. B.12 (cont.): Reduced nitrogen, potassium, calcium, magnesium

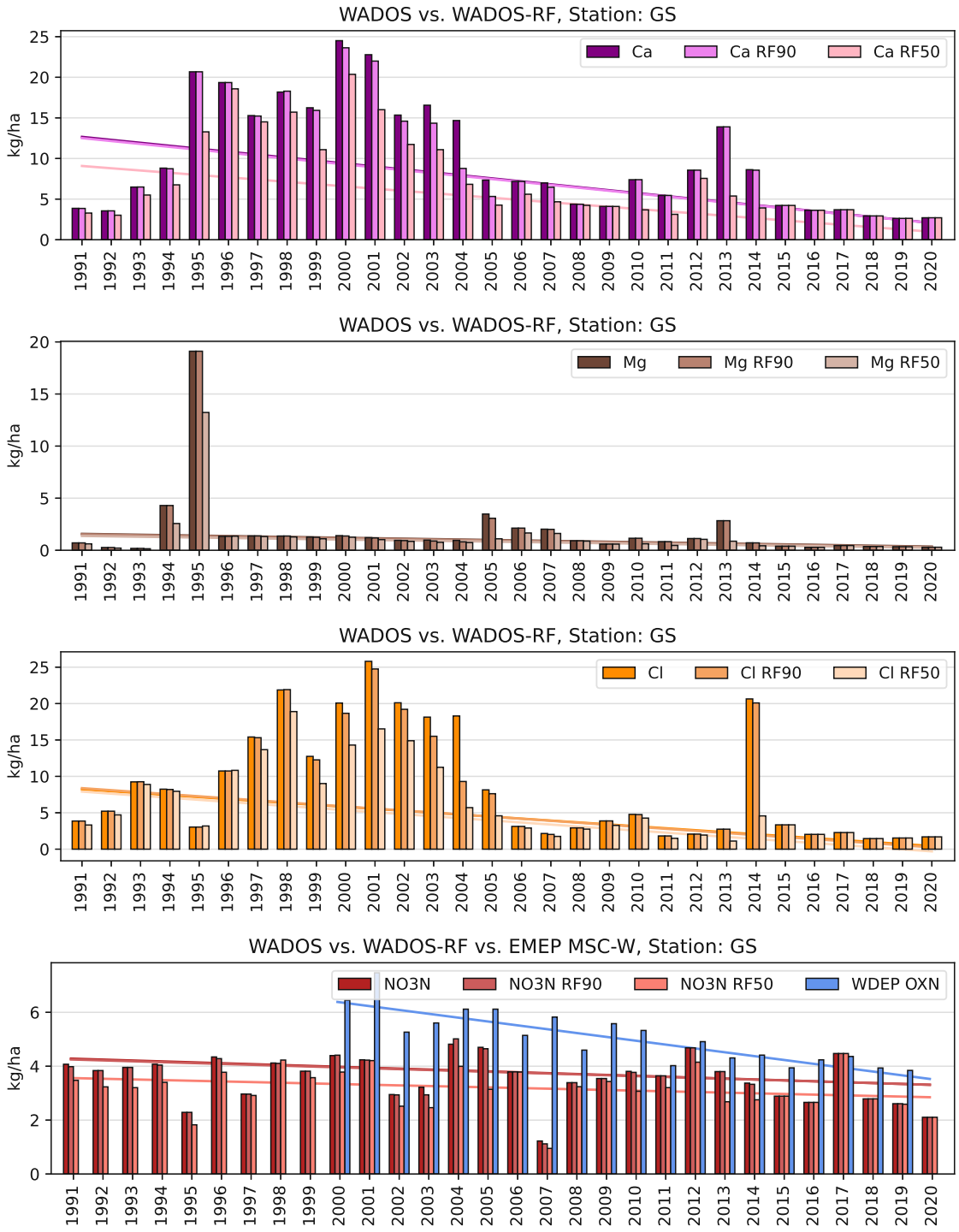


(c) Fig. B.12 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

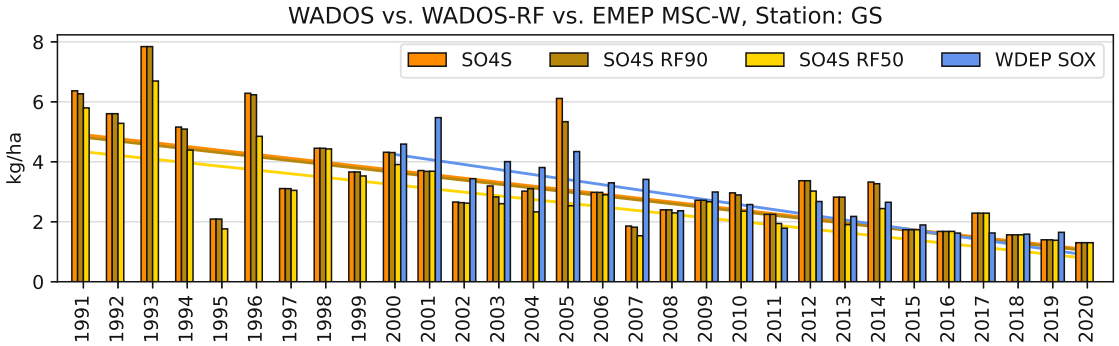


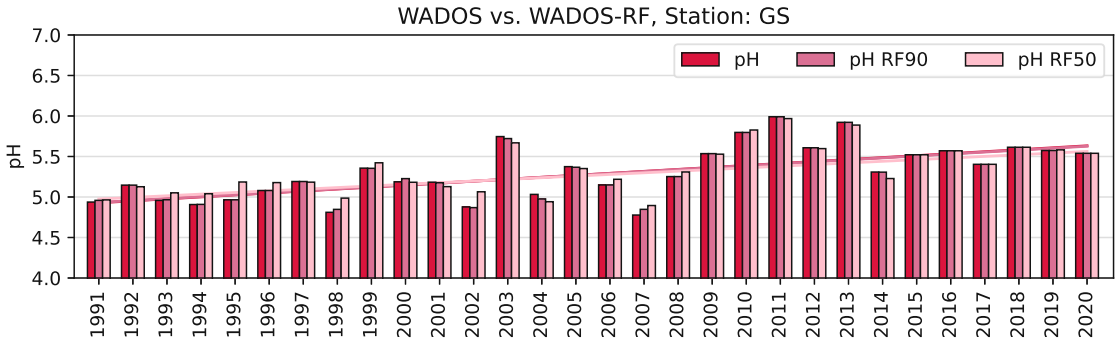
(a) Precipitation amount, sodium, reduced nitrogen, potassium

Figure B.13: Influence of RF classification on depositions at Grundlsee

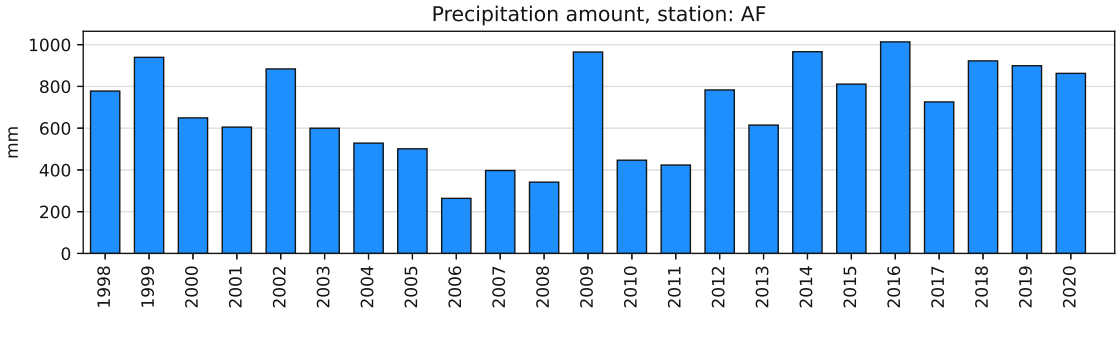


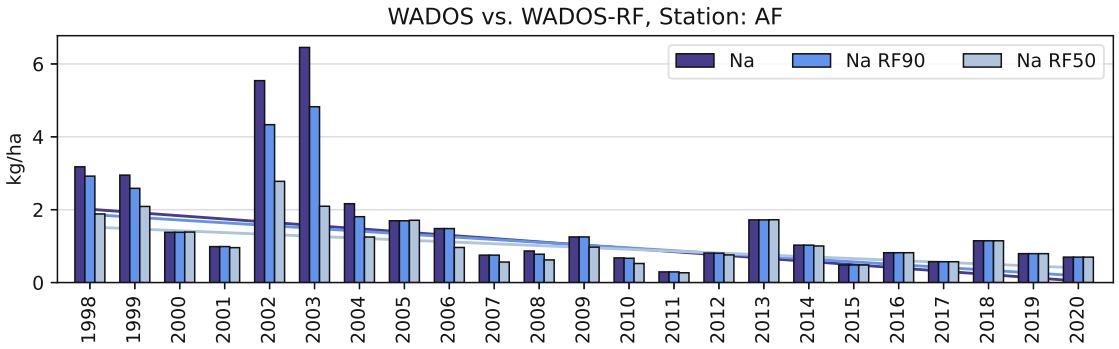
(b) Fig. B.13 (cont.): Calcium, magnesium, chlorid, oxidized nitrogen





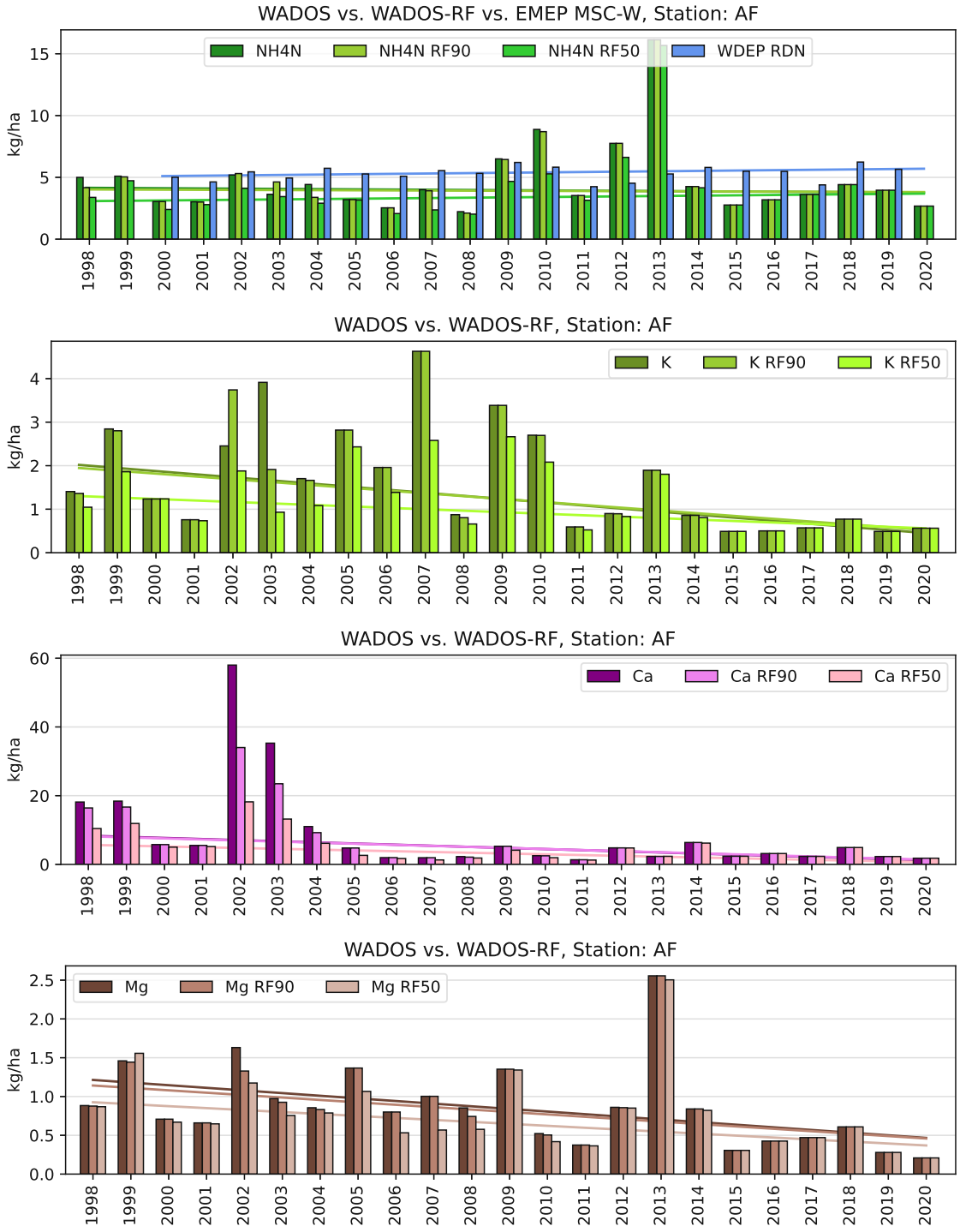
(c) Fig. B.13 (cont.): Sulfur, pH value



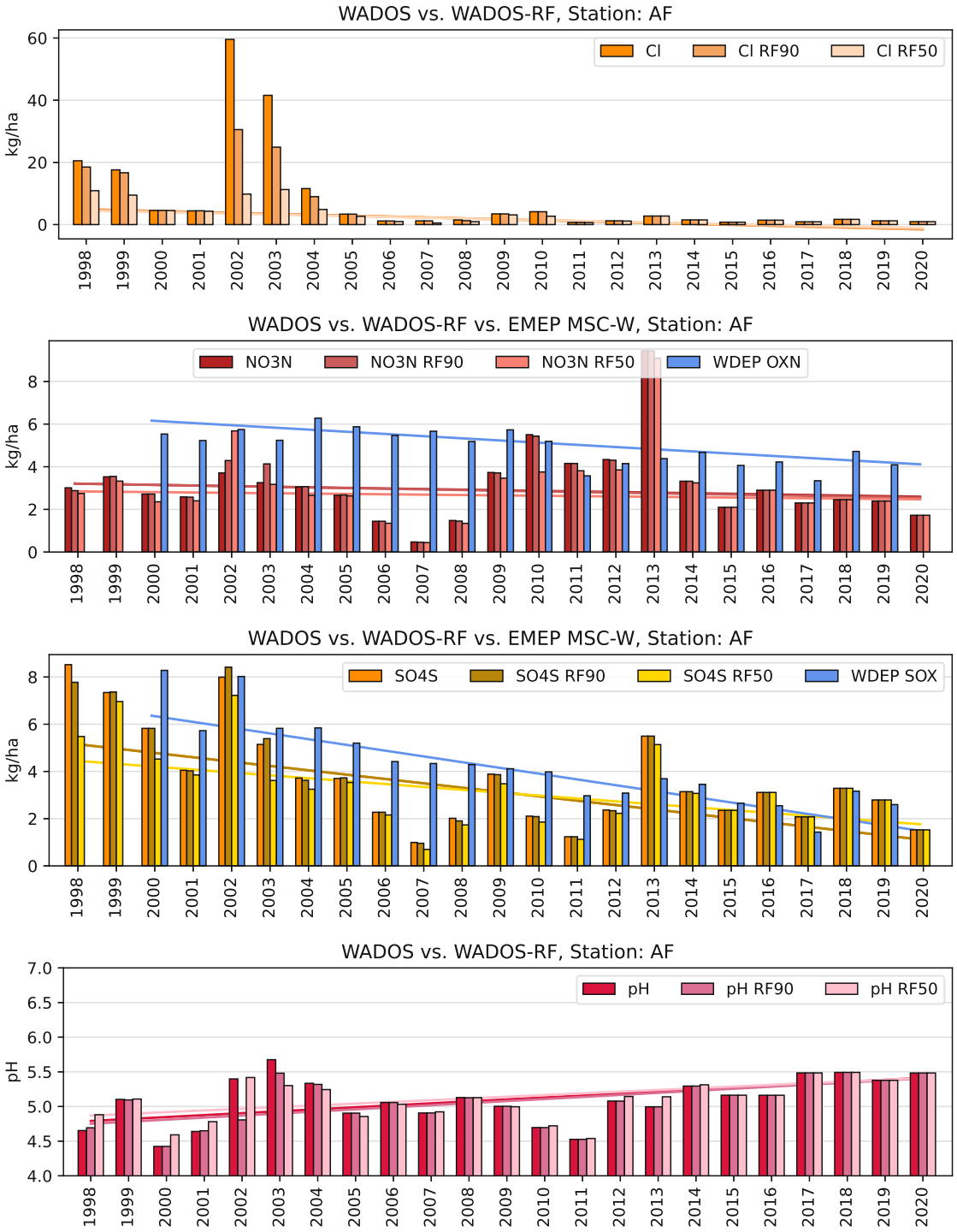


(a) Precipitation amount, sodium

Figure B.14: Influence of RF classification on depositions in Arnfels

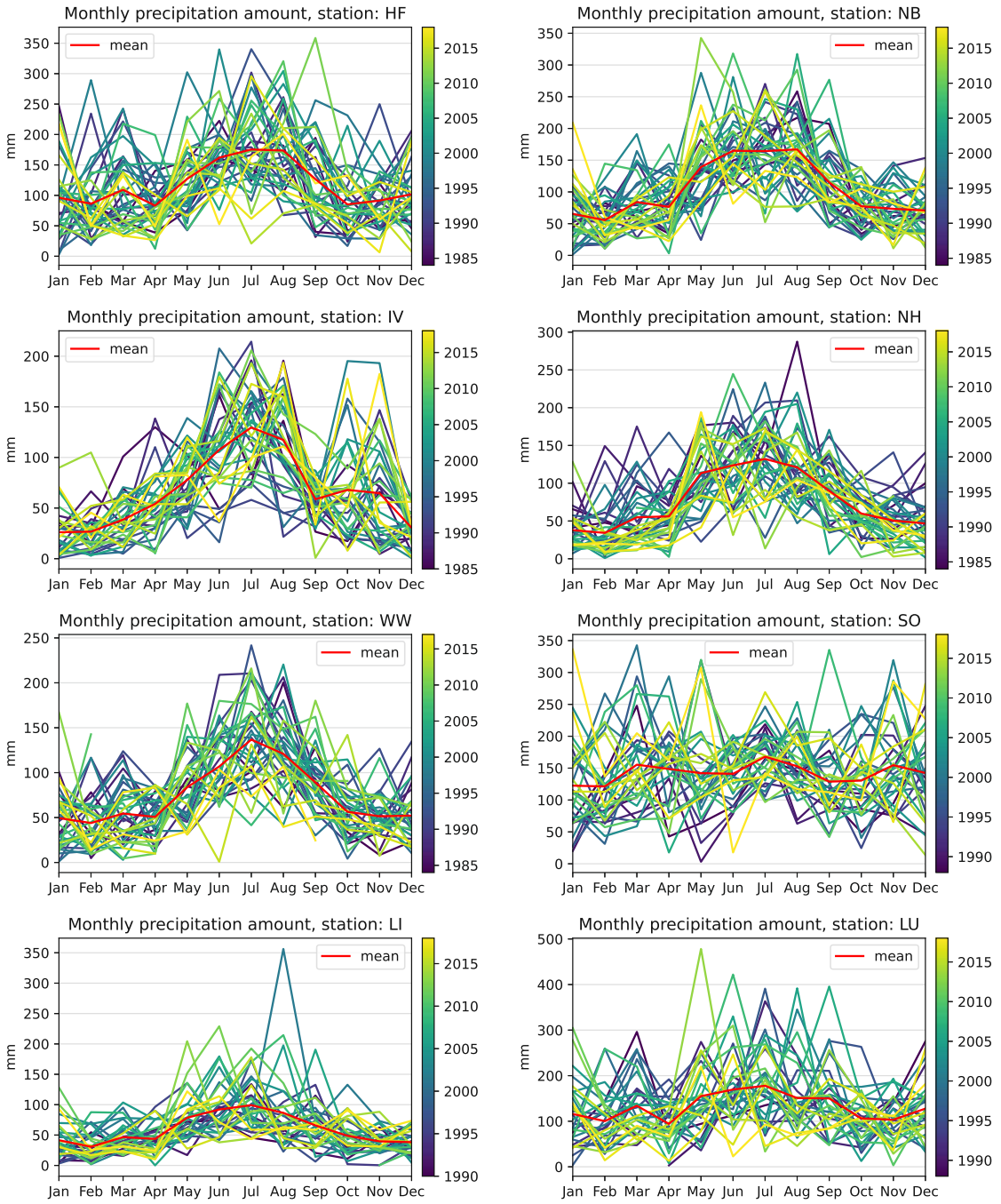


(b) Fig. B.14 (cont.): Reduced nitrogen, potassium, calcium, magnesium



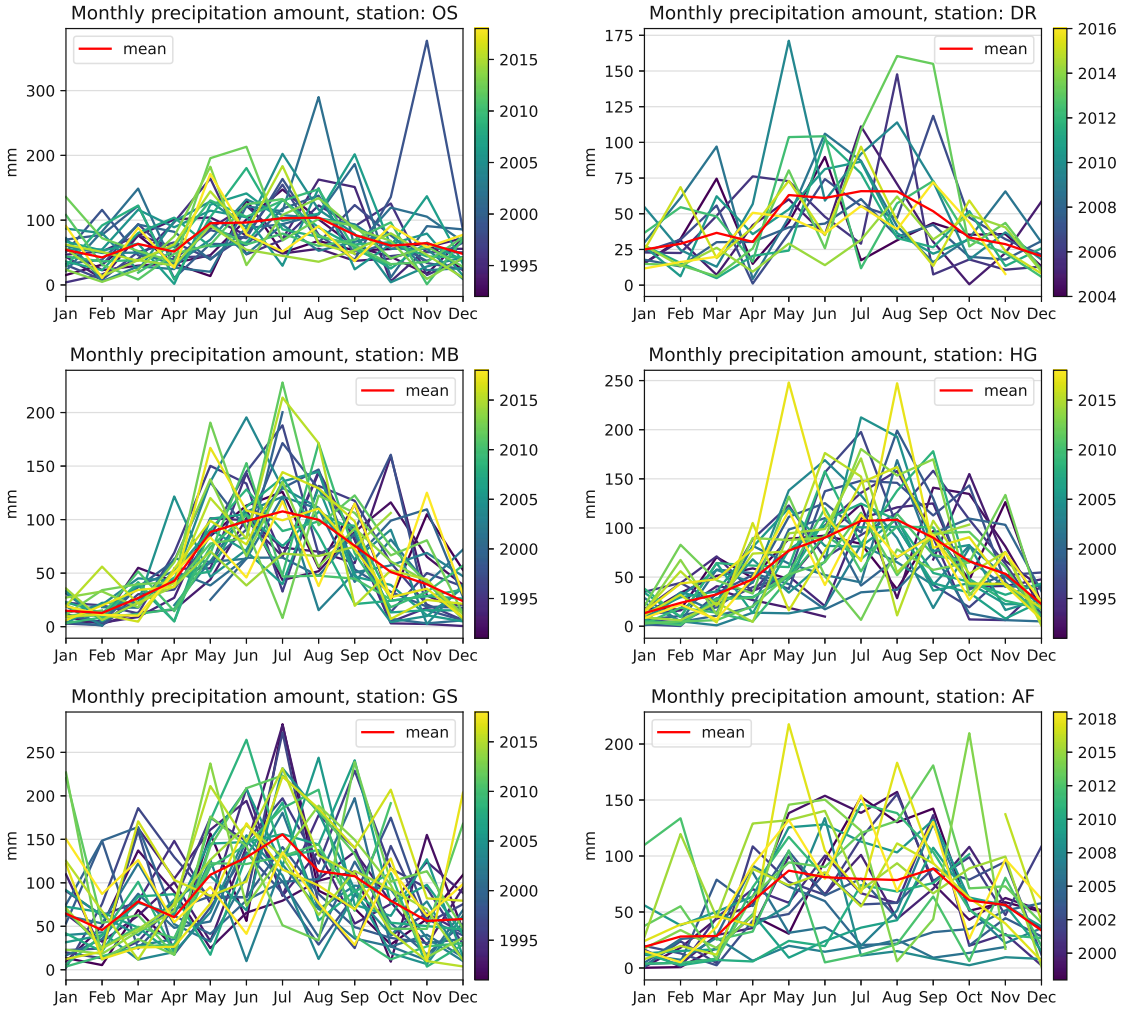
(c) Fig. B.14 (cont.): Chlorid, oxidized nitrogen, sulfur, pH value

Appendix C Seasonality plots

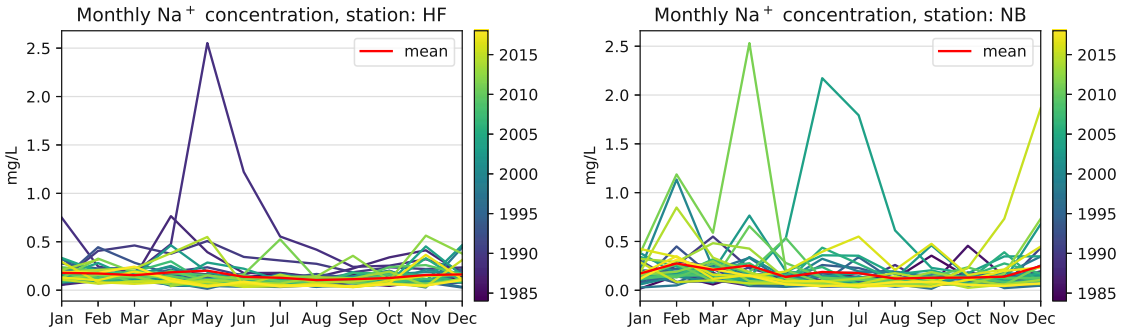


(a)

Figure C.1: Monthly precipitation amounts

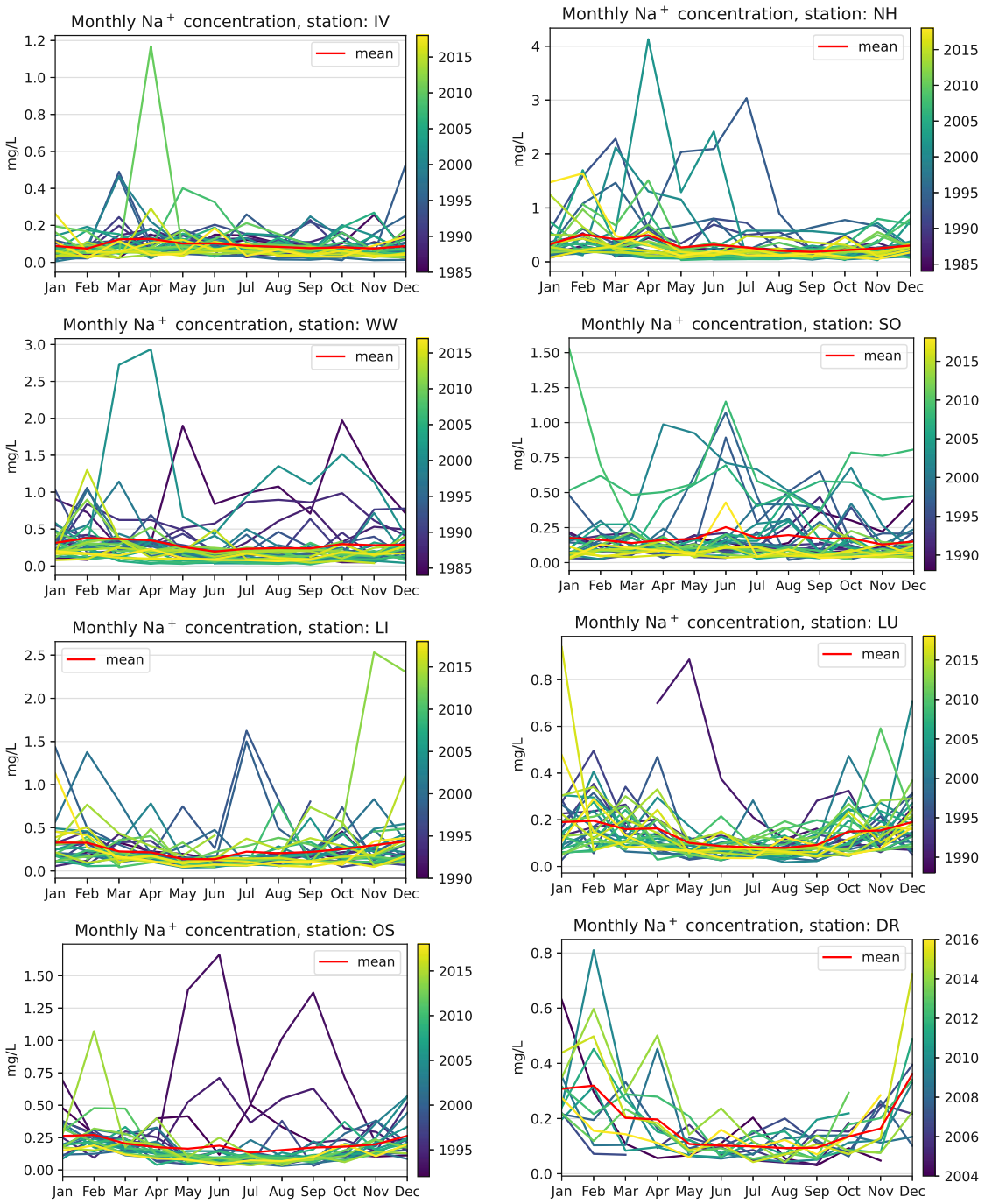


(b) Fig. C.1 (cont.)

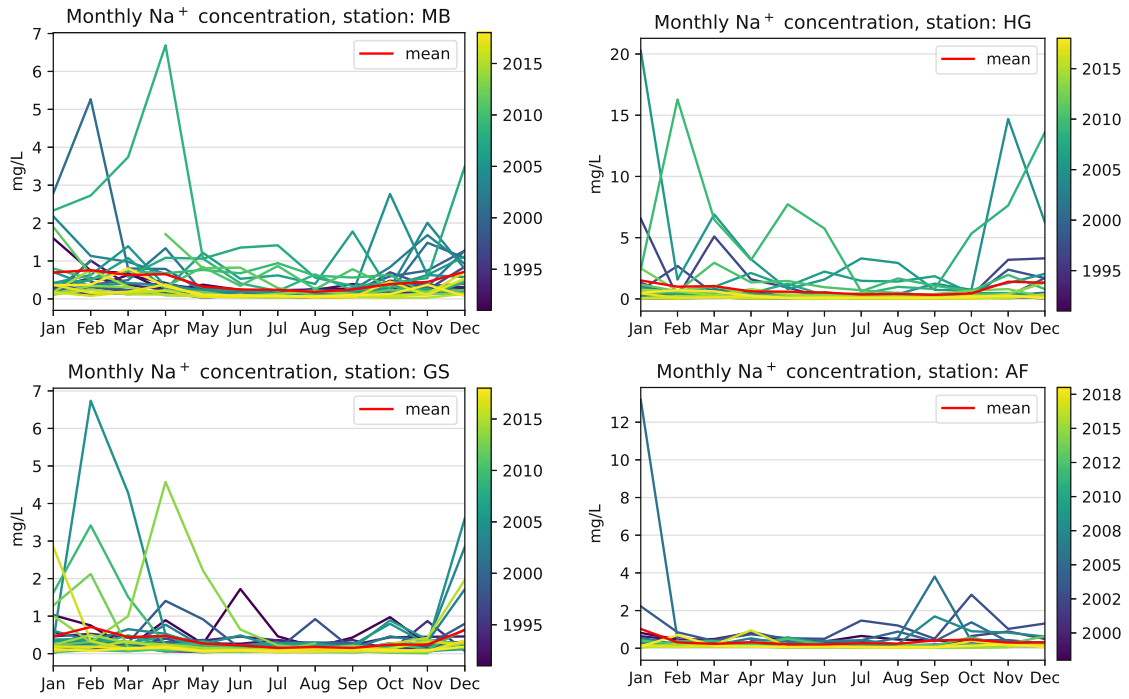


(a)

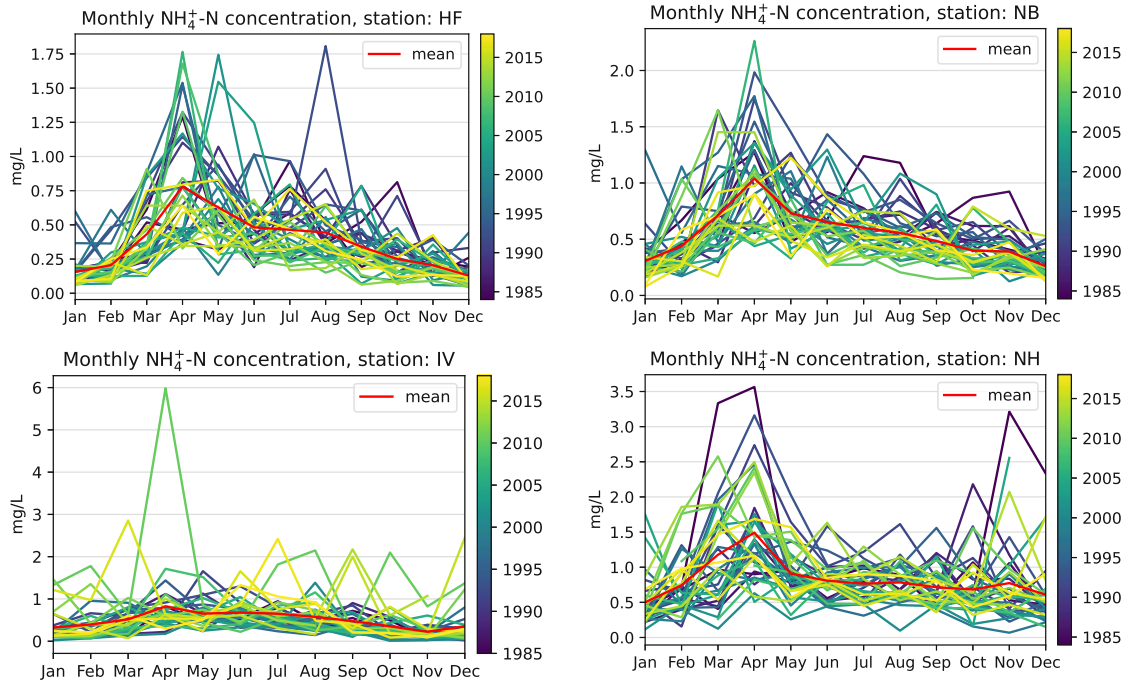
Figure C.2: Monthly Na^+ concentrations



(b) Fig. C.2 (cont.)

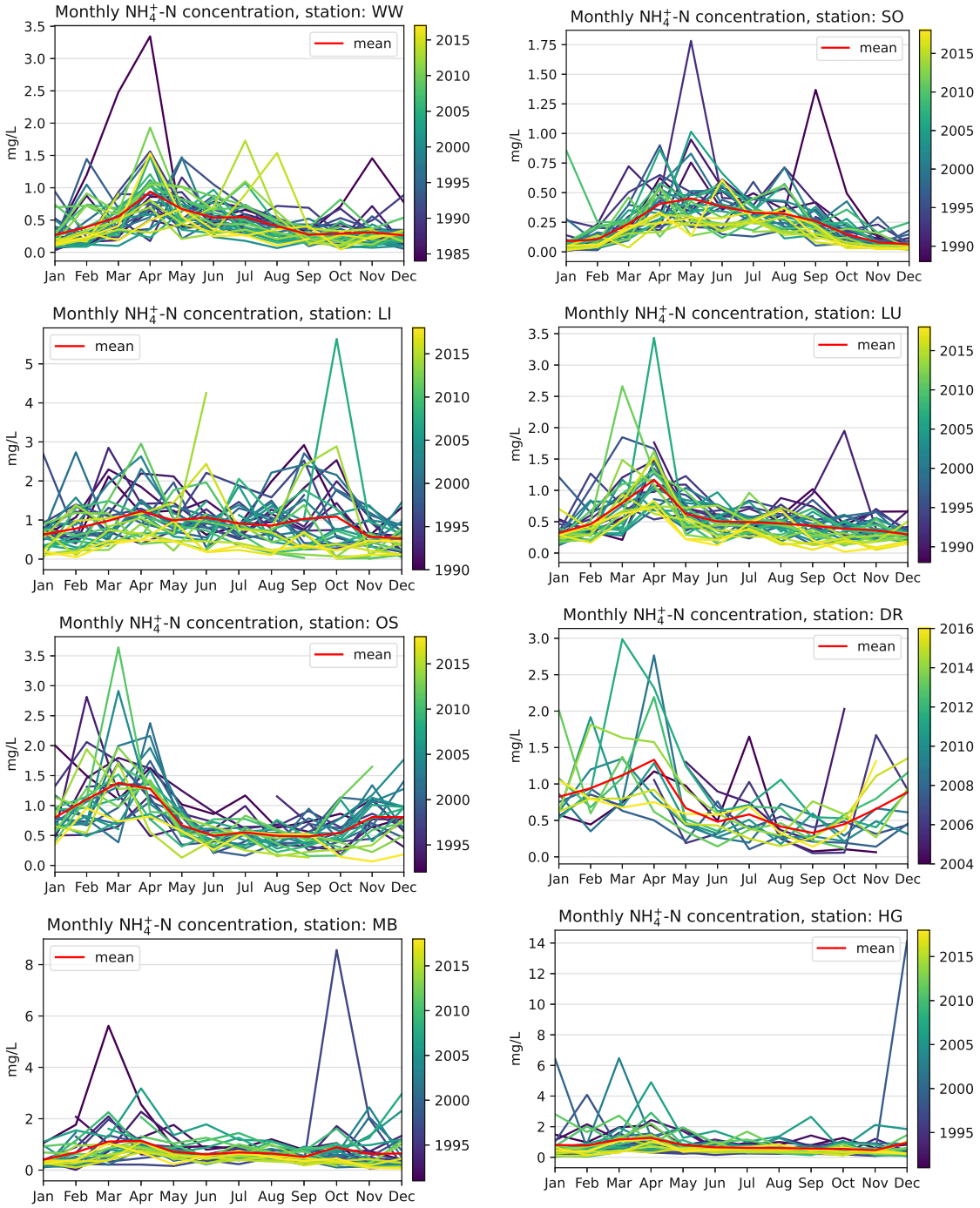


(c) Fig. C.2 (cont.)

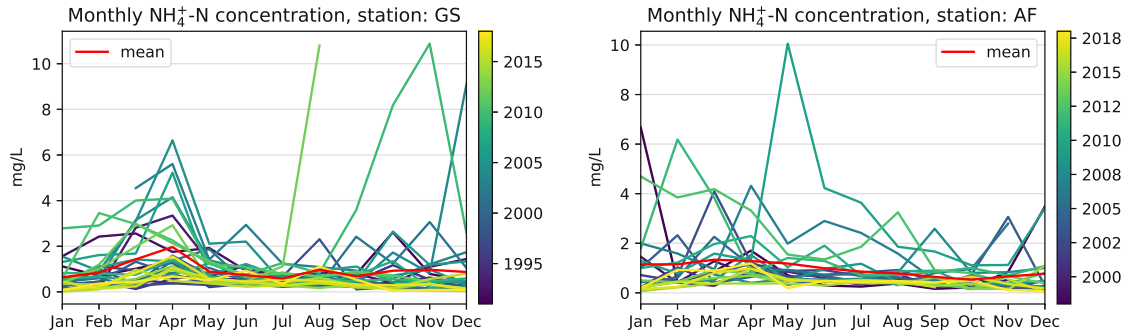


(a)

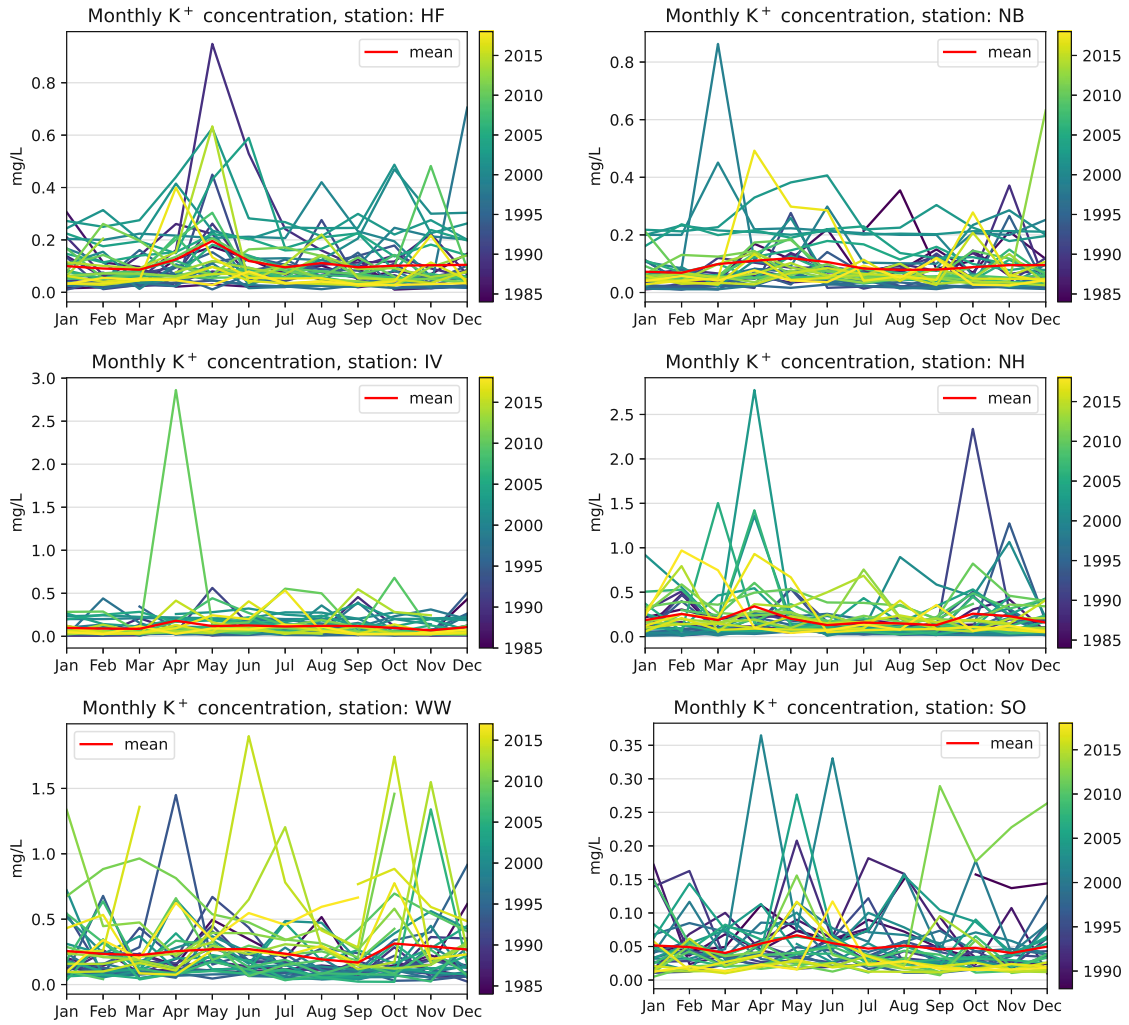
Figure C.3: Monthly $NH_4{}^+$-N concentrations



(b) Fig. C.3 (cont.)

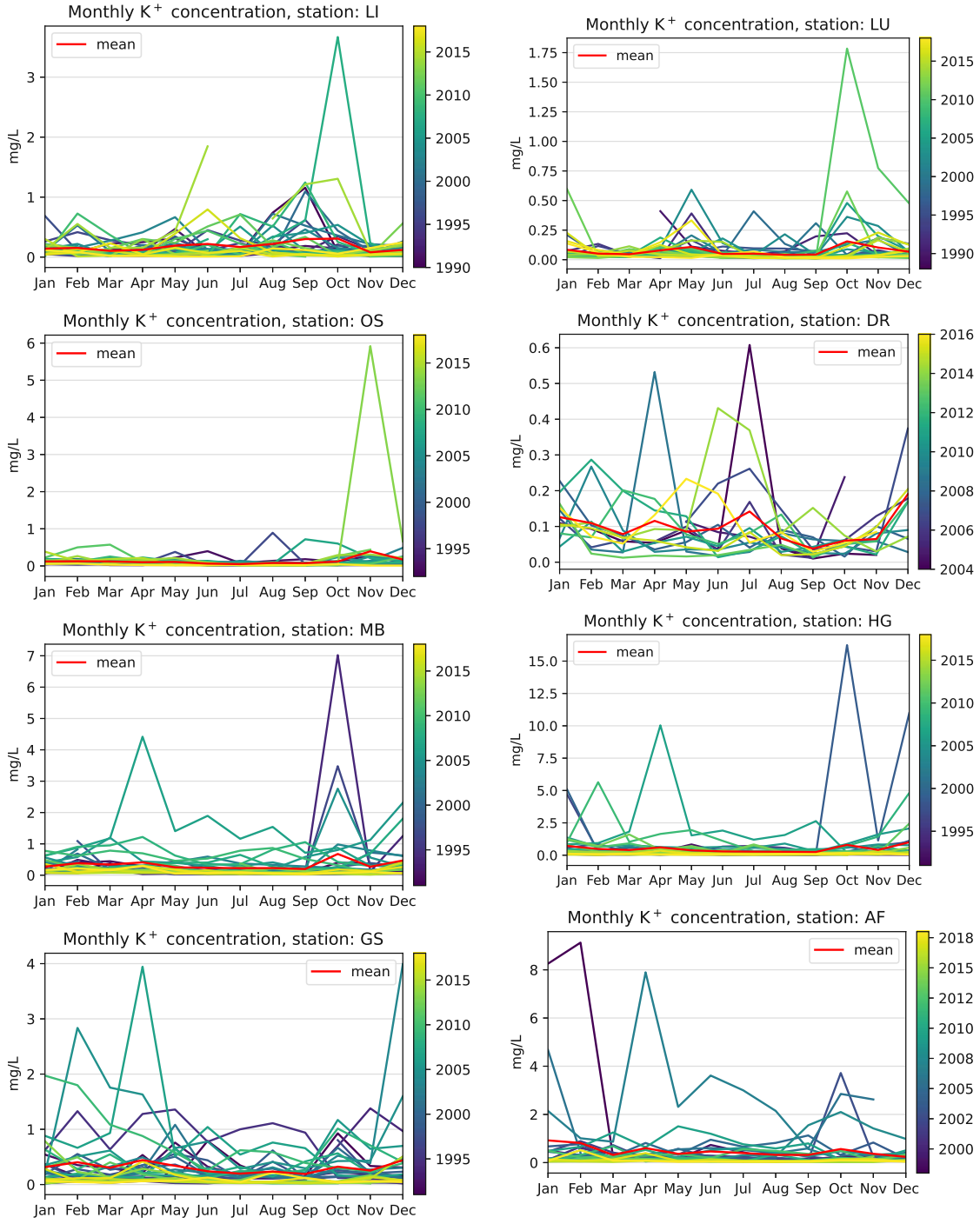


(c) Fig. C.3 (cont.)

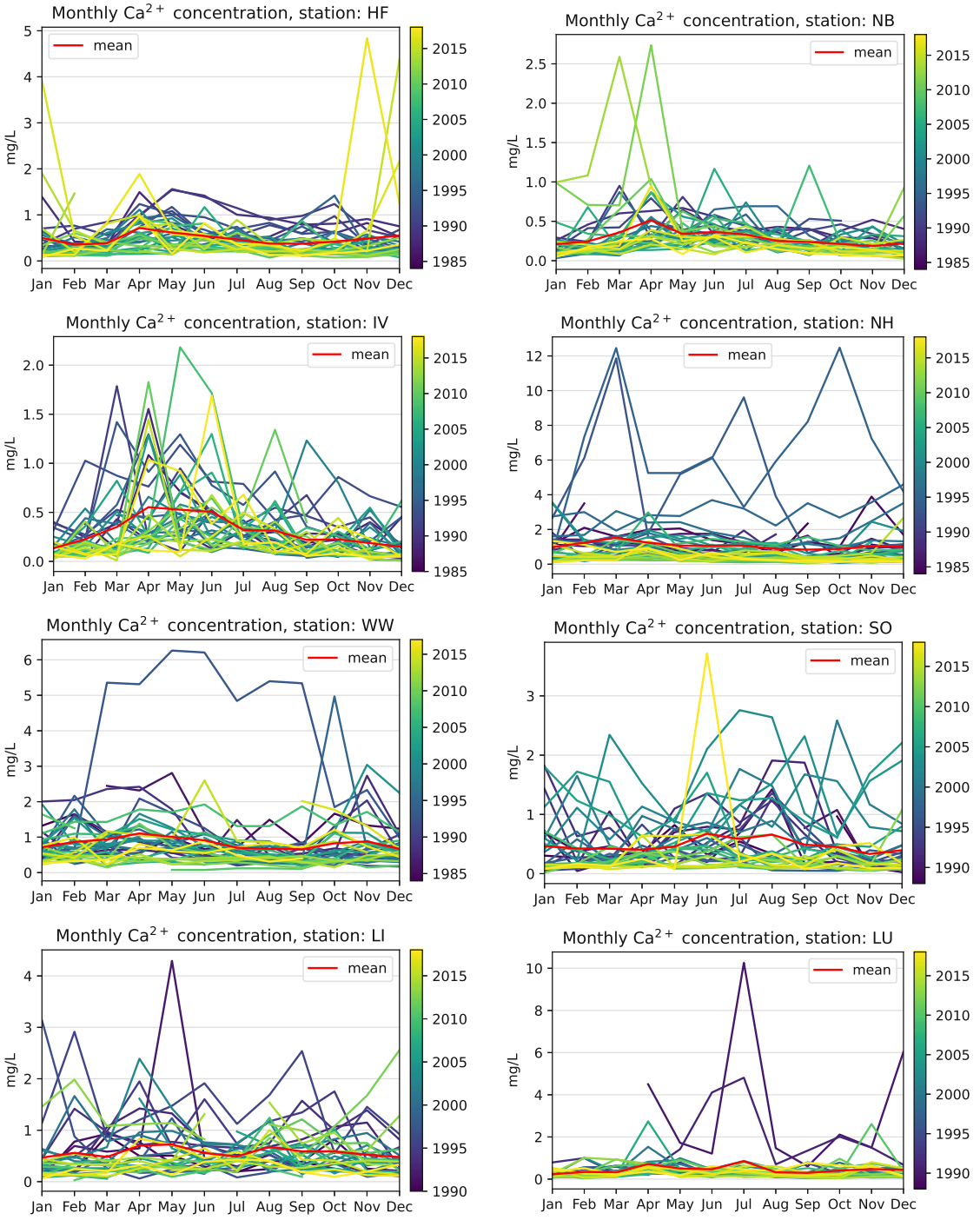


(a)

Figure C.4: Monthly K^+ concentrations

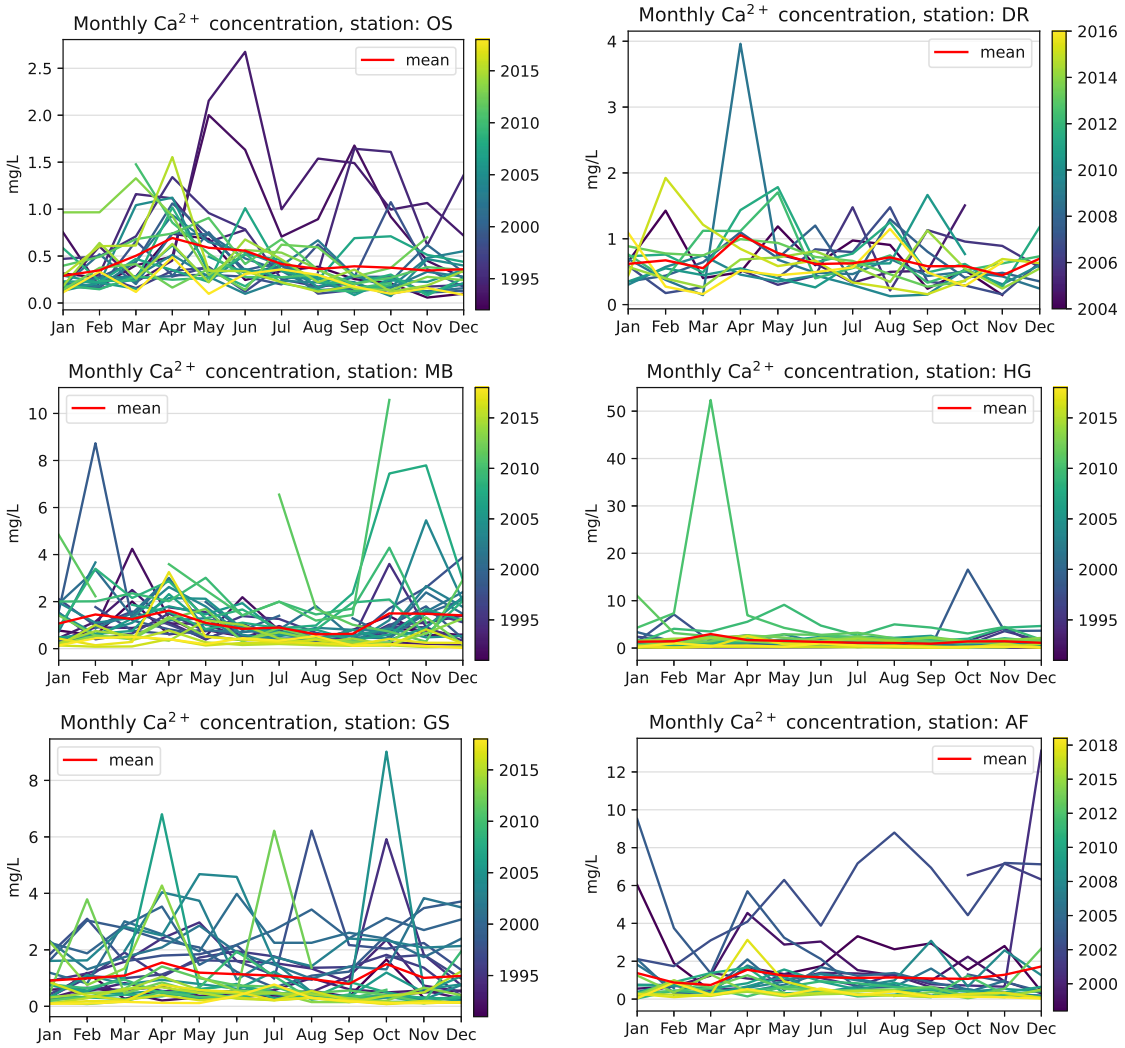


(b) Fig. C.4 (cont.)

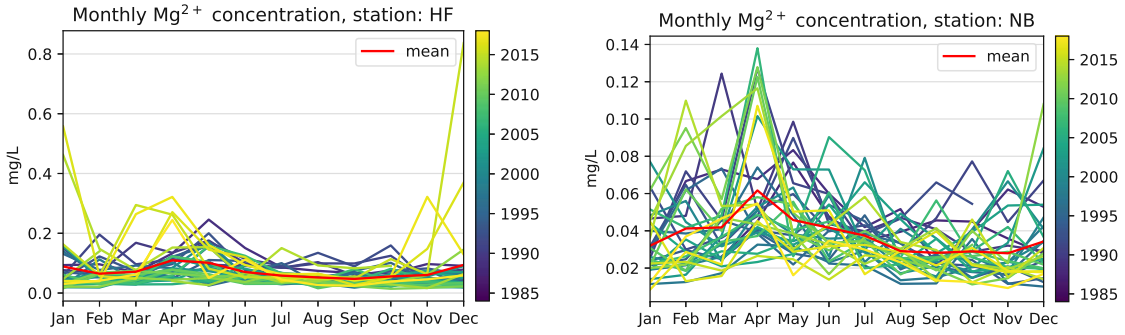


(a)

Figure C.5: Monthly Ca^{2+} concentrations

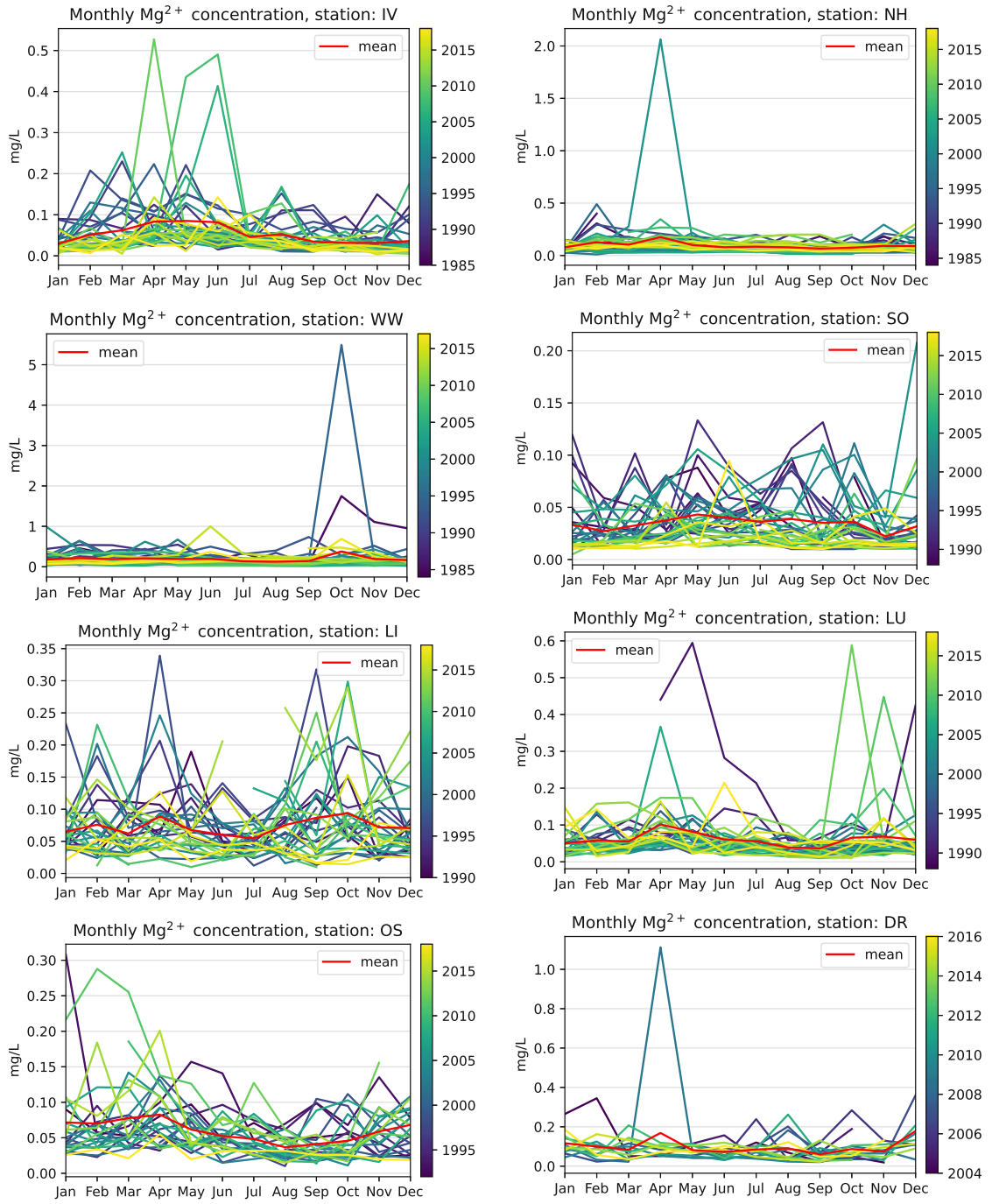


(b) Fig. C.5 (cont.)

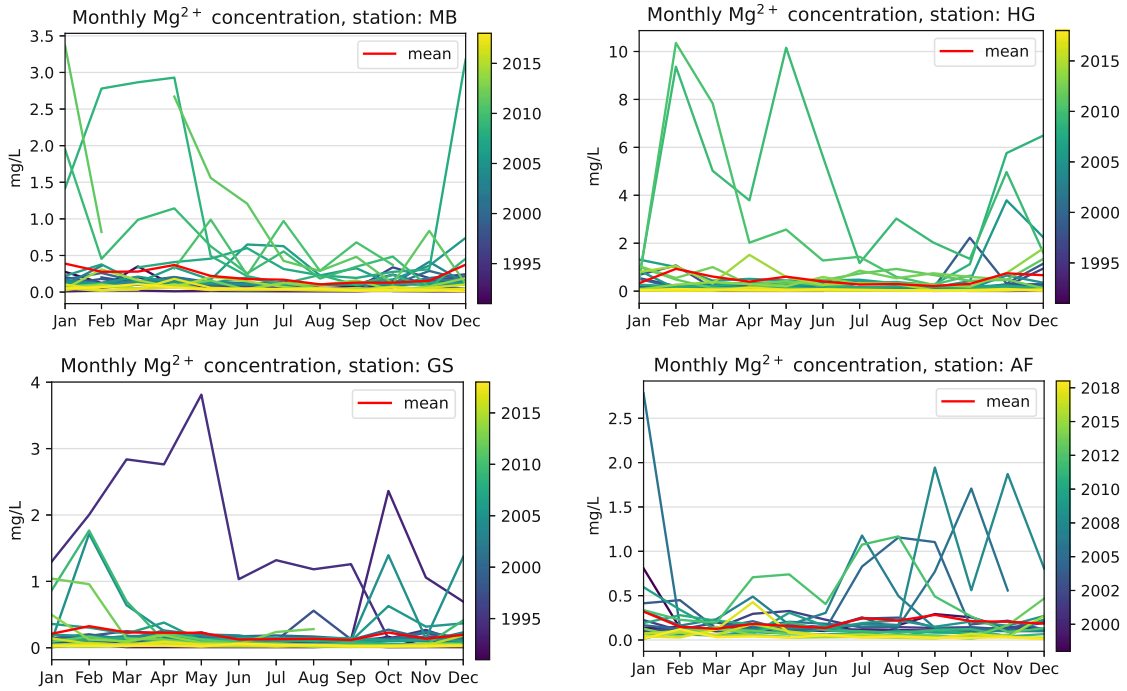


(a)

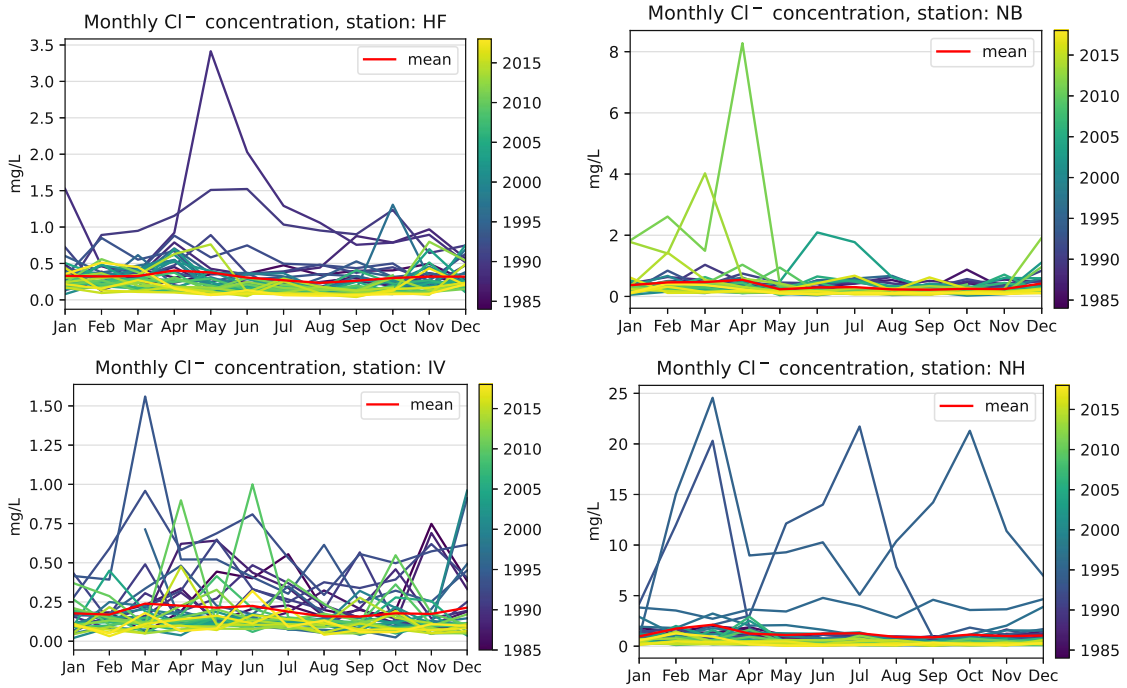
Figure C.6: Monthly Mg^{2+} concentrations



(b) Fig. C.6 (cont.)

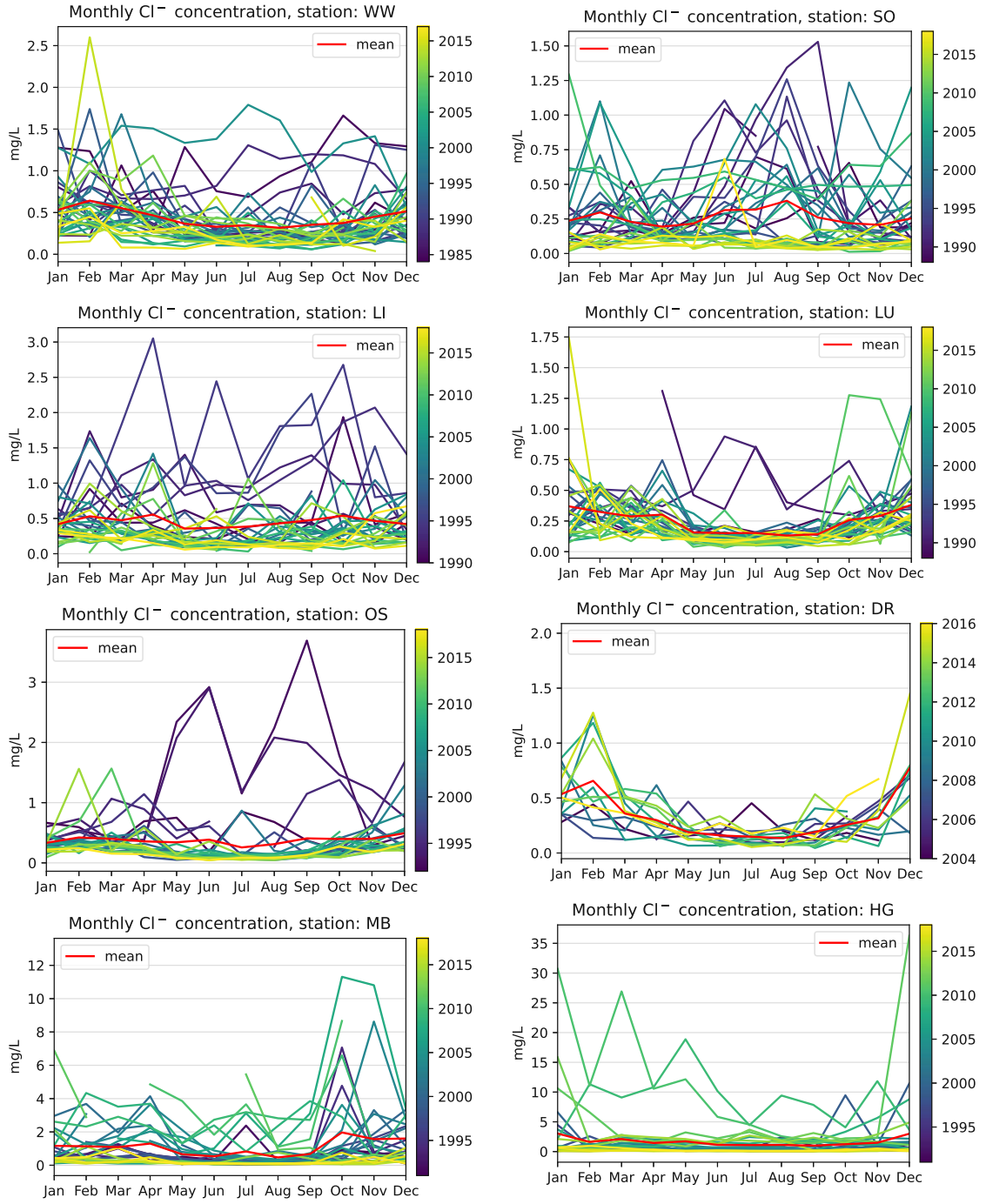


(c) Fig. C.6 (cont.)

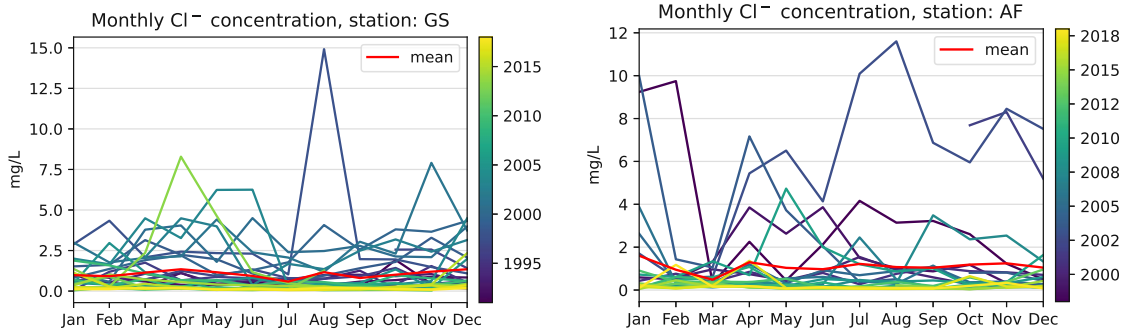


(a)

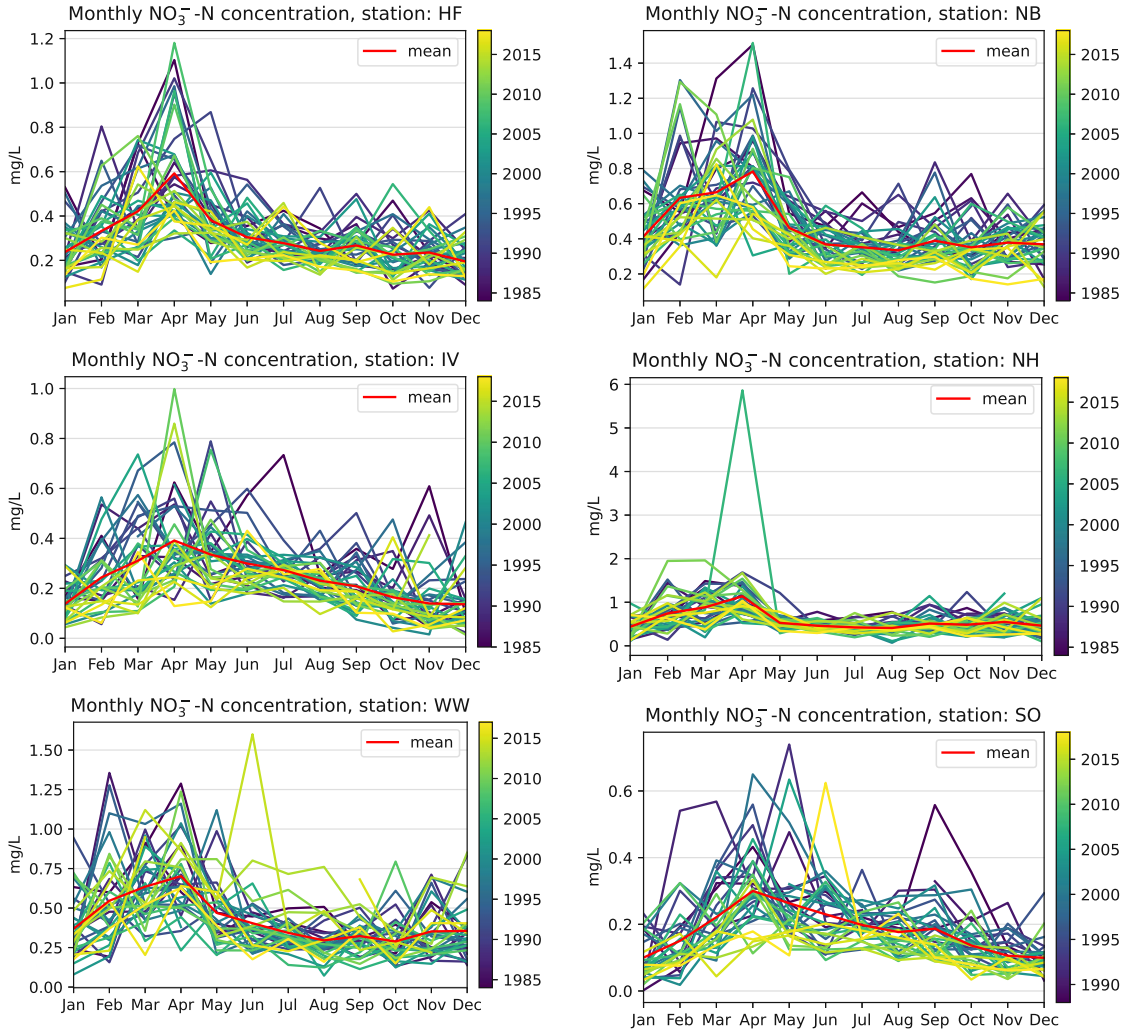
Figure C.7: Monthly Cl^- concentrations



(b) Fig. C.7 (cont.)

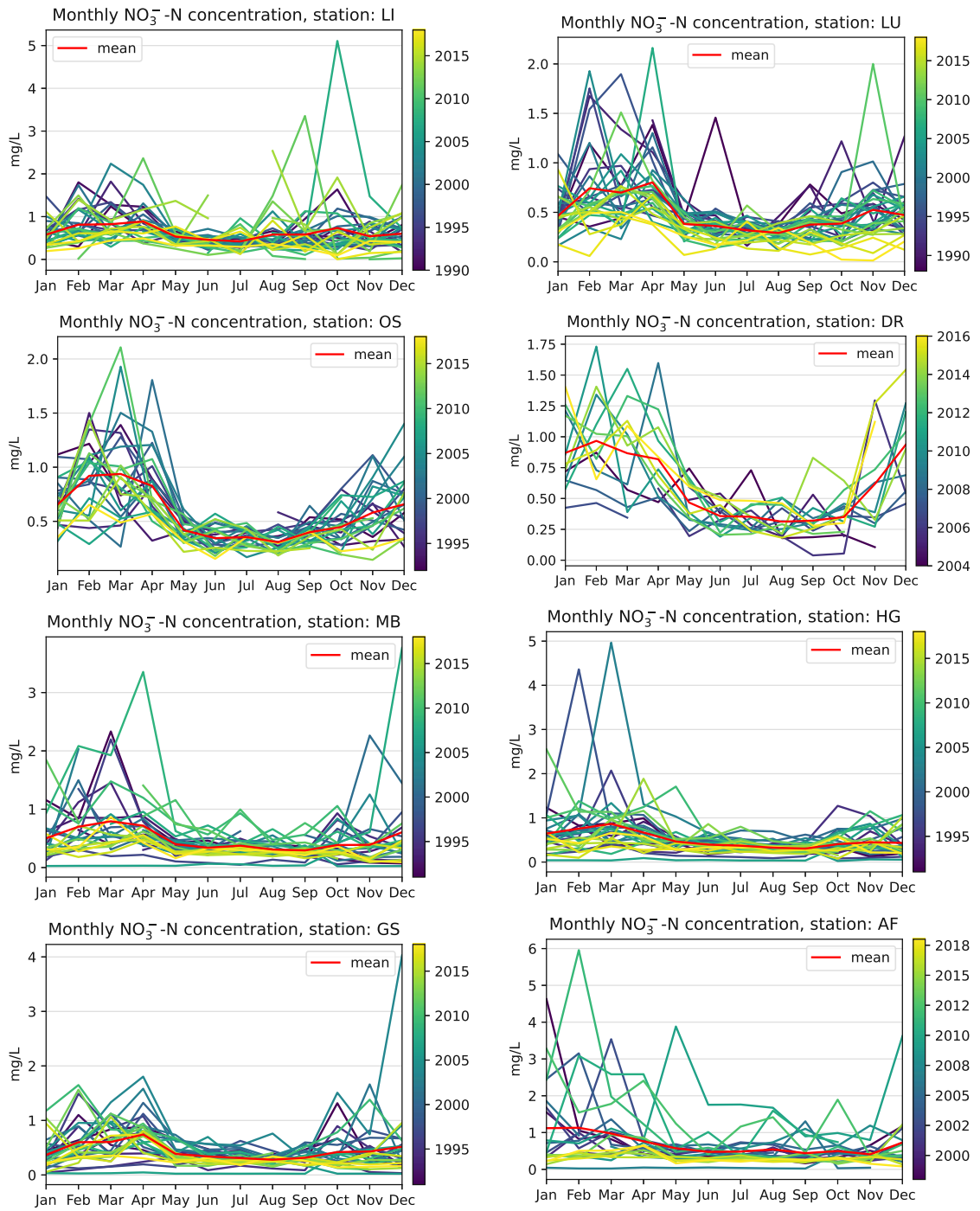


(c) Fig. C.7 (cont.)

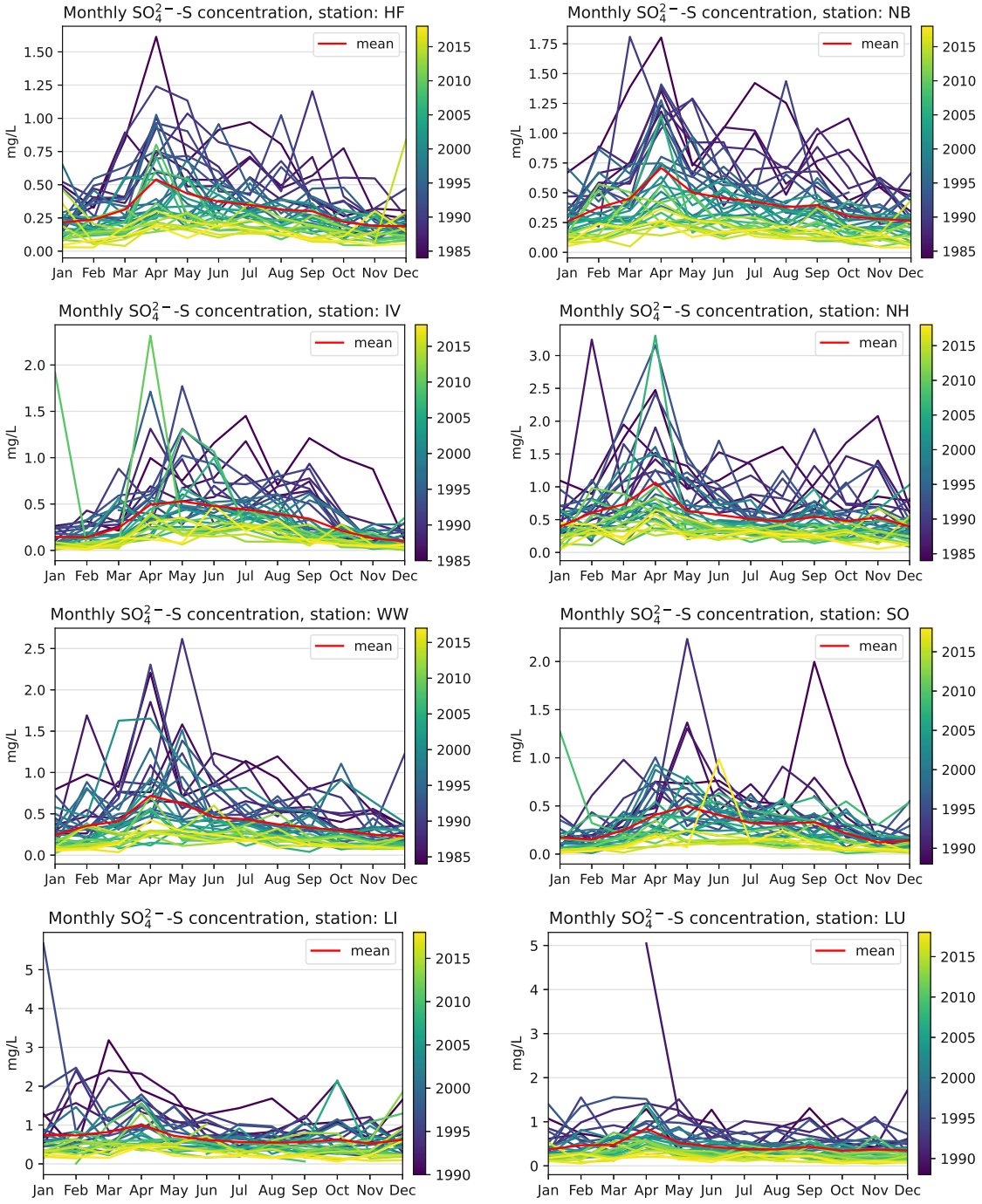


(a)

Figure C.8: Monthly NO_3^--N concentrations

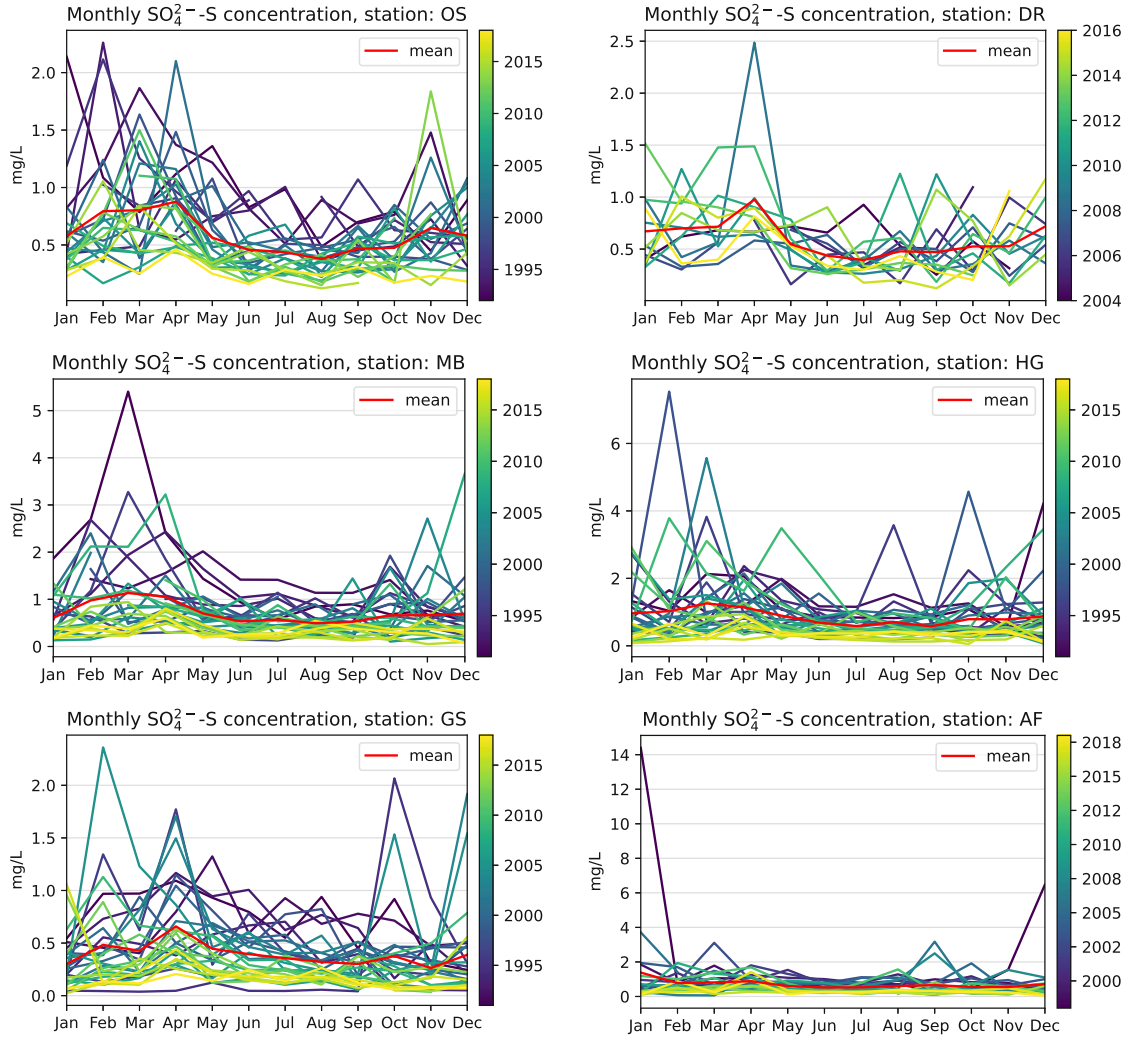


(b) Fig. C.8 (cont.)

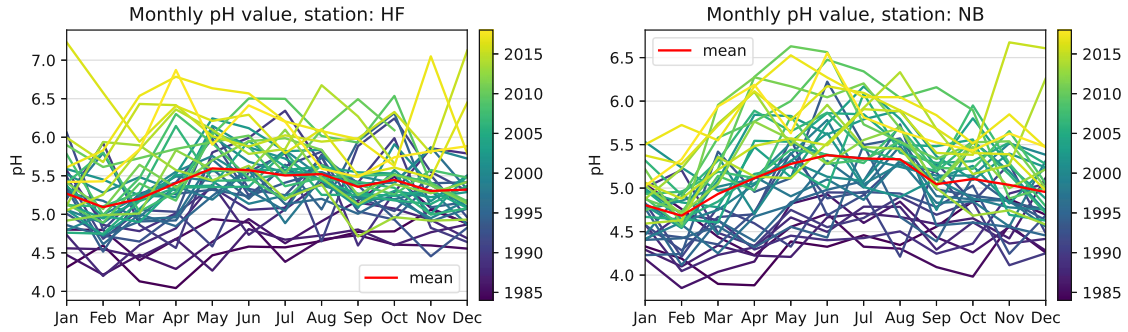


(a)

Figure C.9: Monthly SO_4^{2-}-S concentrations

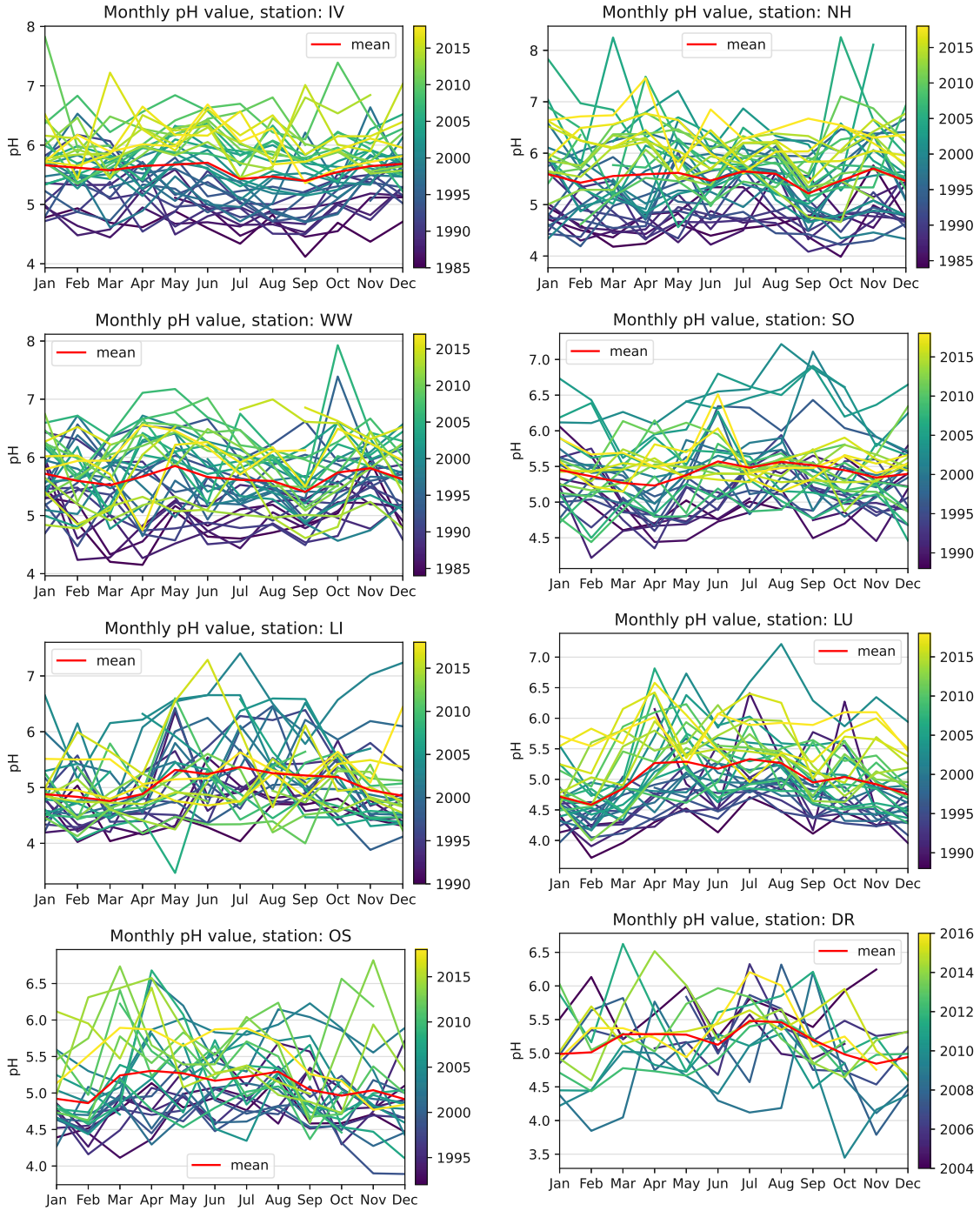


(b) Fig. C.9 (cont.)

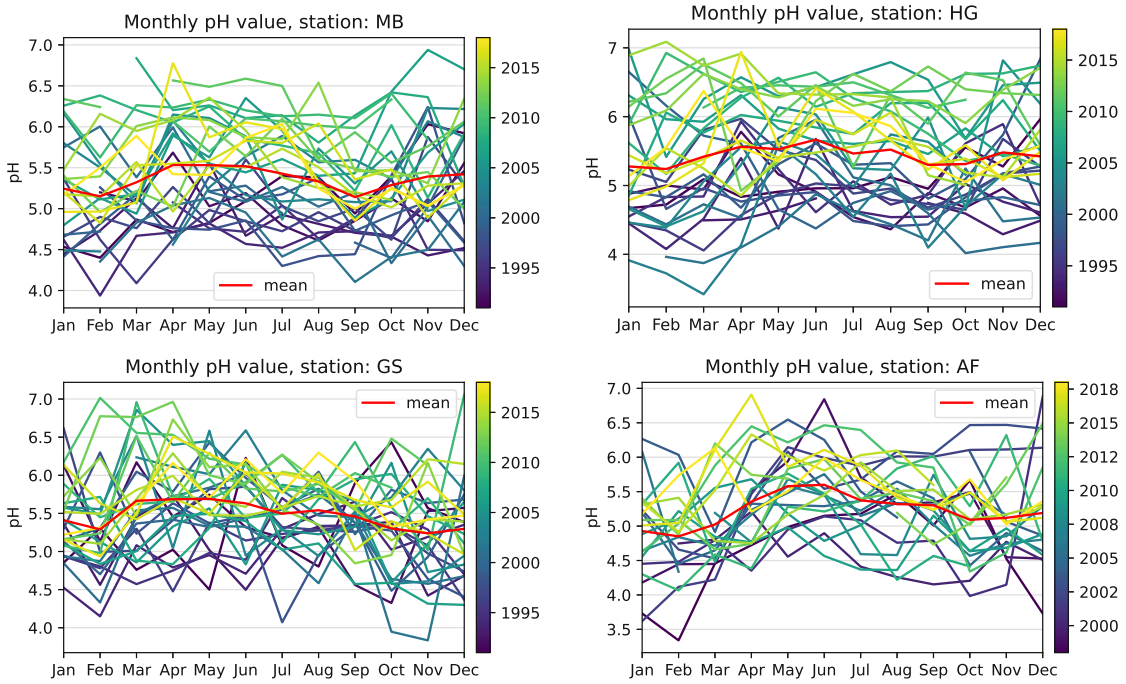


(a)

Figure C.10: Monthly pH values

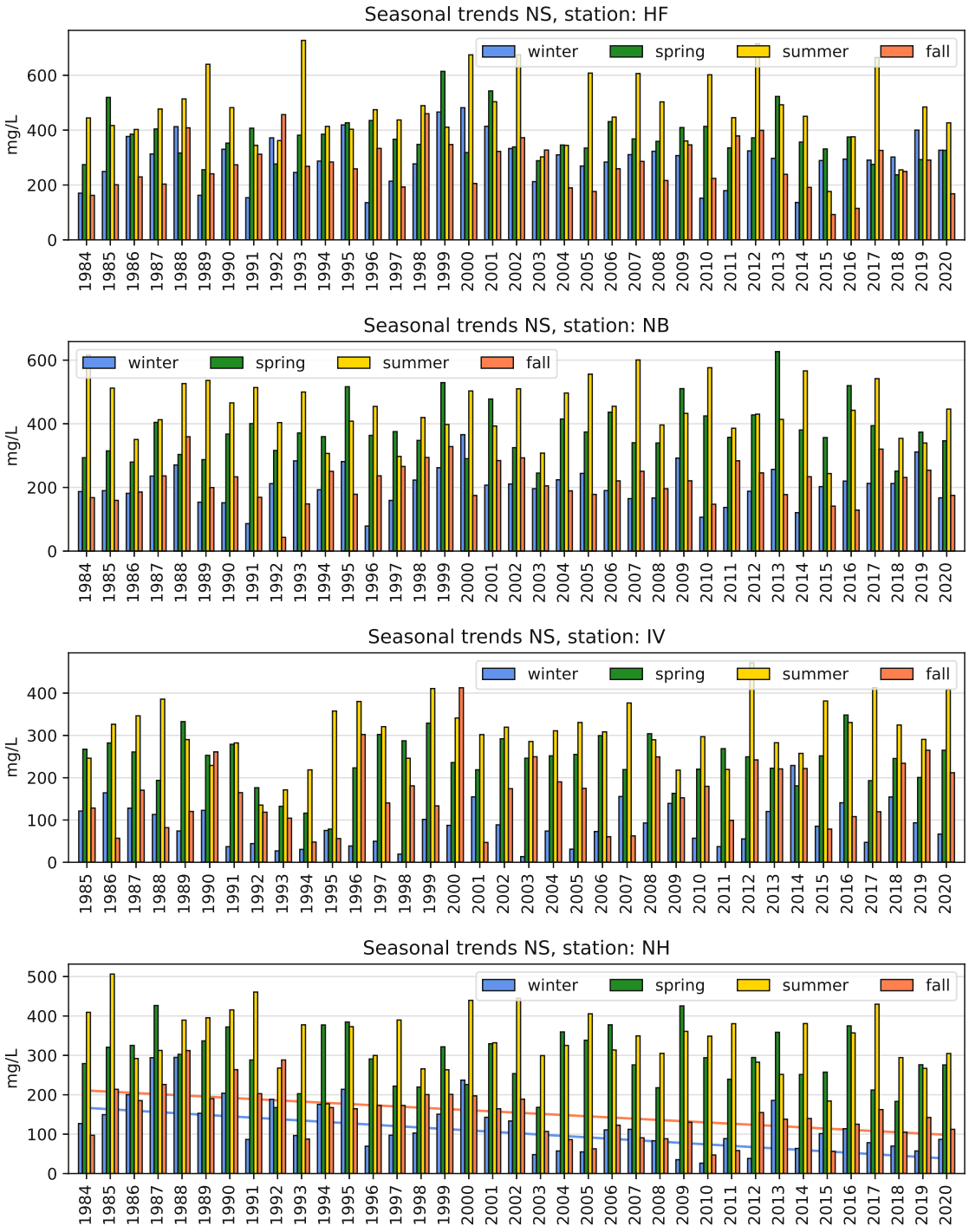


(b) Fig. C.10 (cont.)



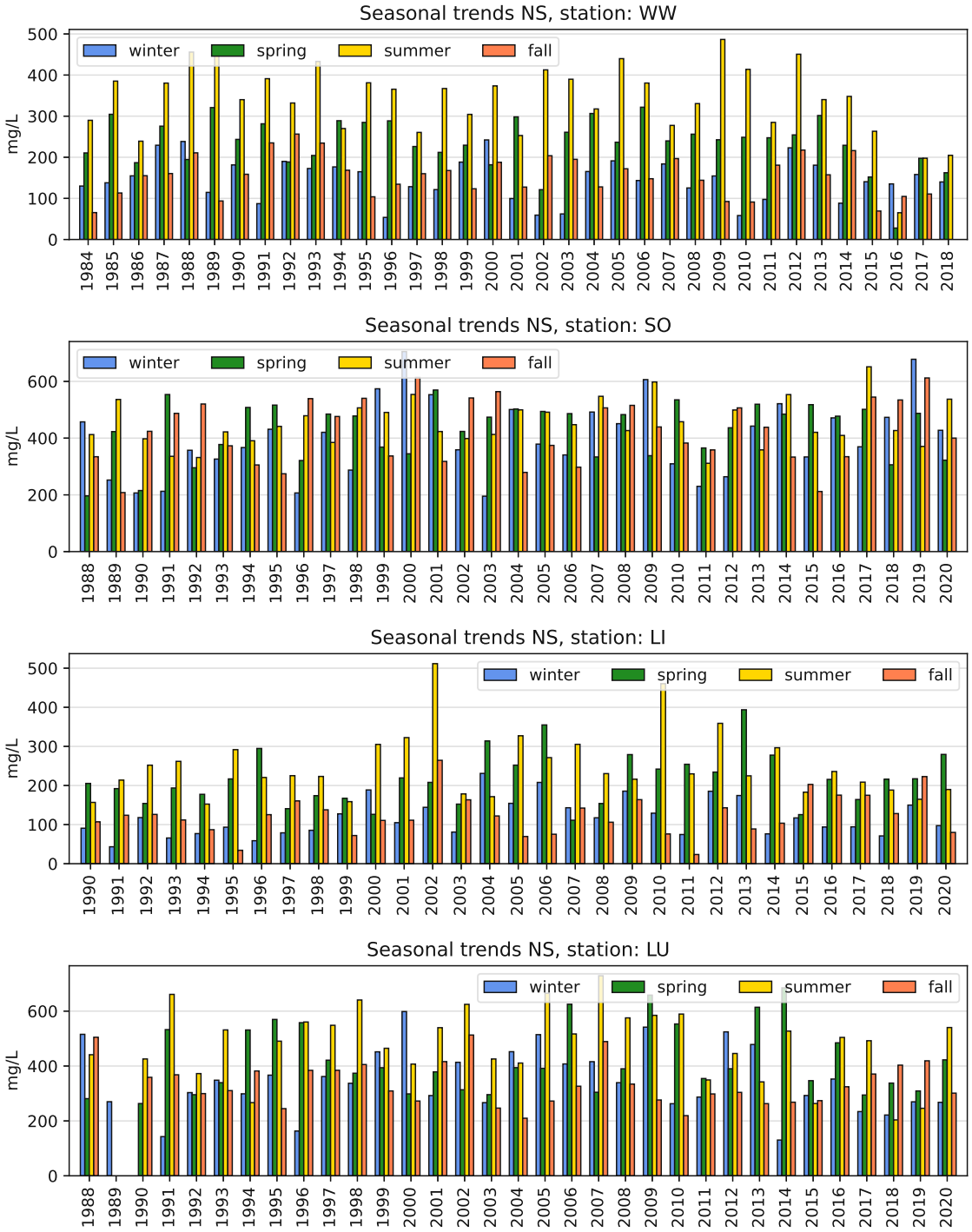
(c) Fig. C.10 (cont.)

Appendix D Seasonal time series

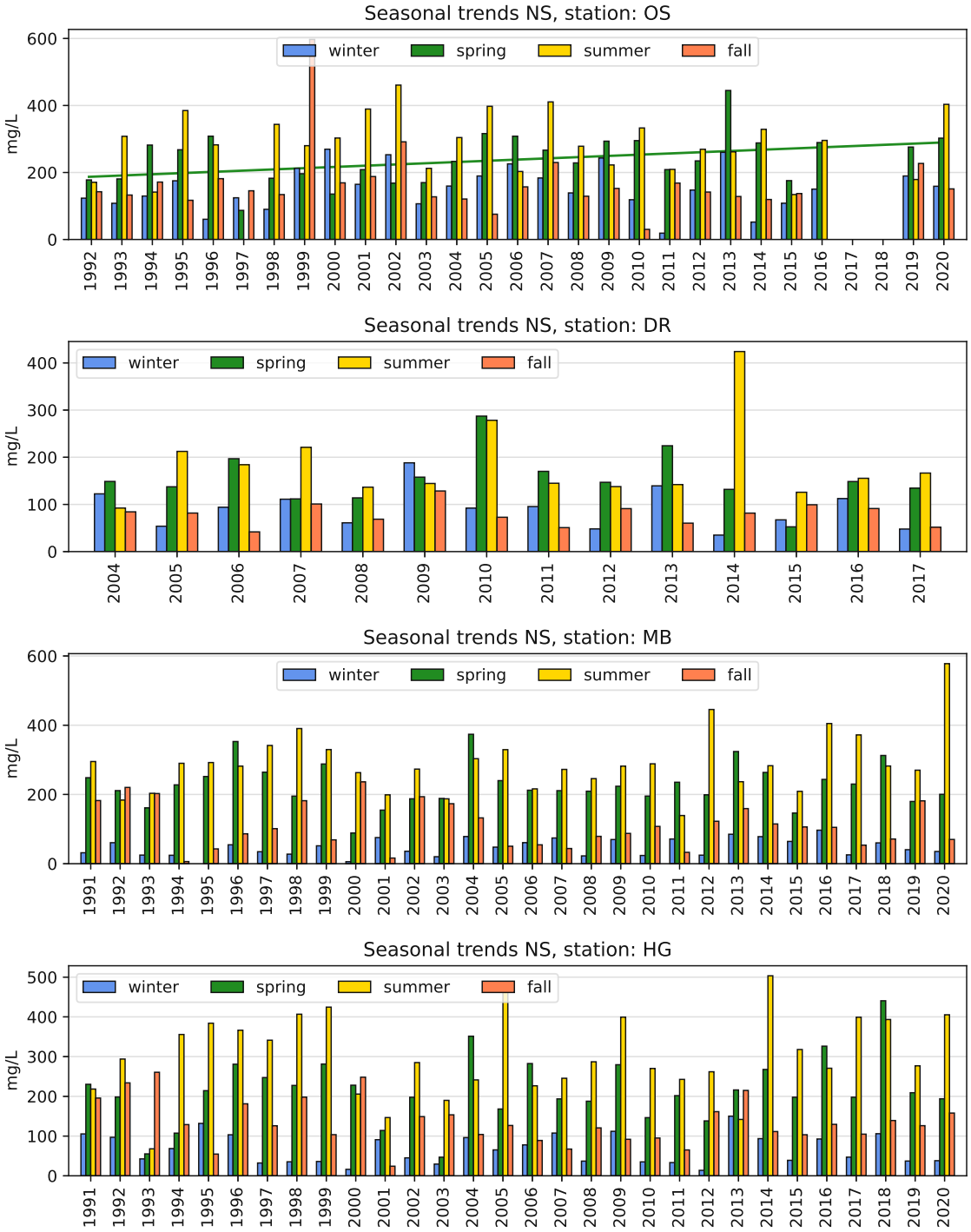


(a) Höfen, Niederndorferberg, Innervillgraten, Haunsberg

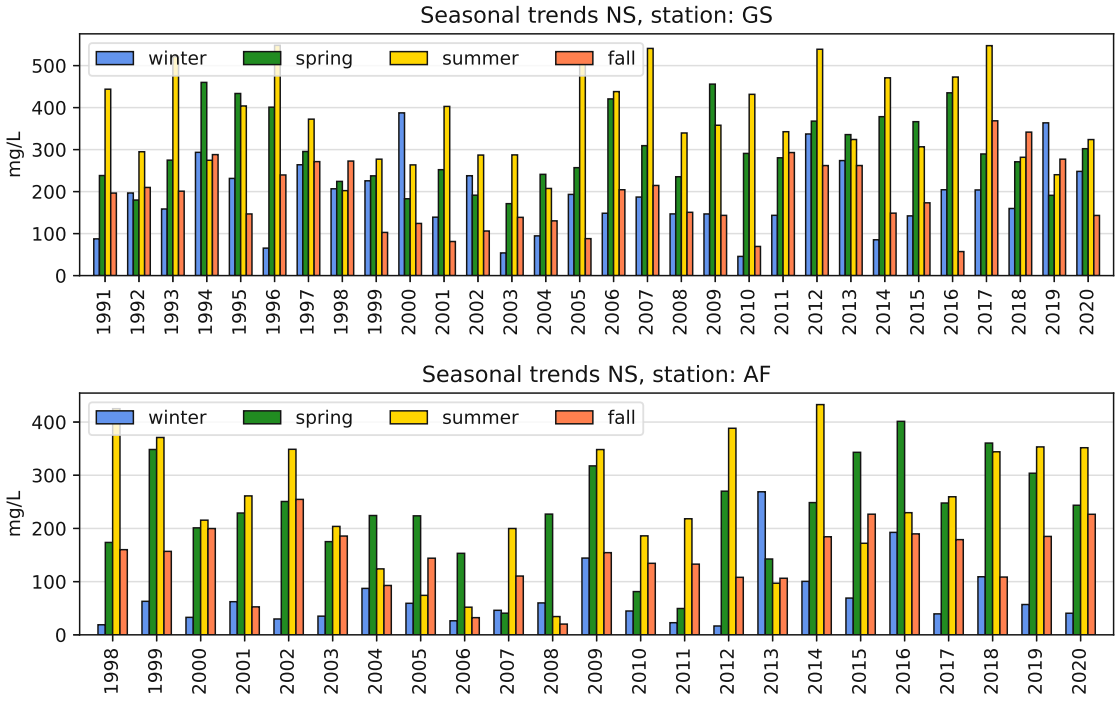
Figure D.1: Separated seasonal precipitation amount trends



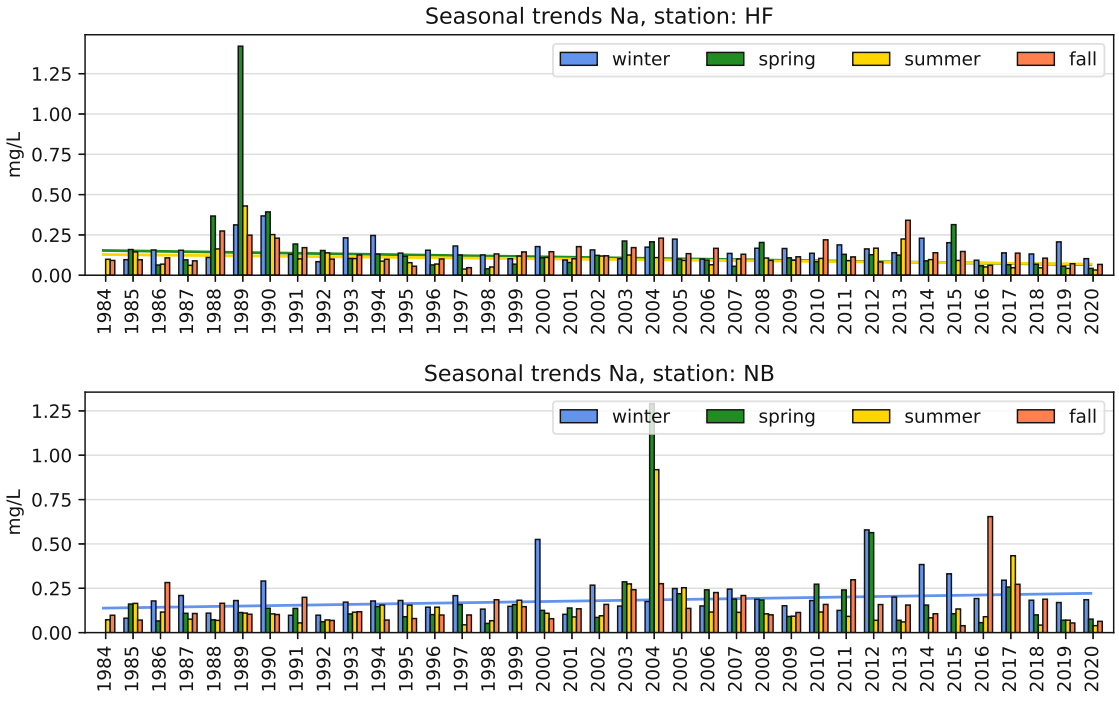
(b) Fig. D.1 (cont.): Werfenweng, Sonnblick, Litschau, Lunz



(c) Fig. D.1 (cont.): Ostrong, Drasenhofen, Masenberg, Hochgößnitz

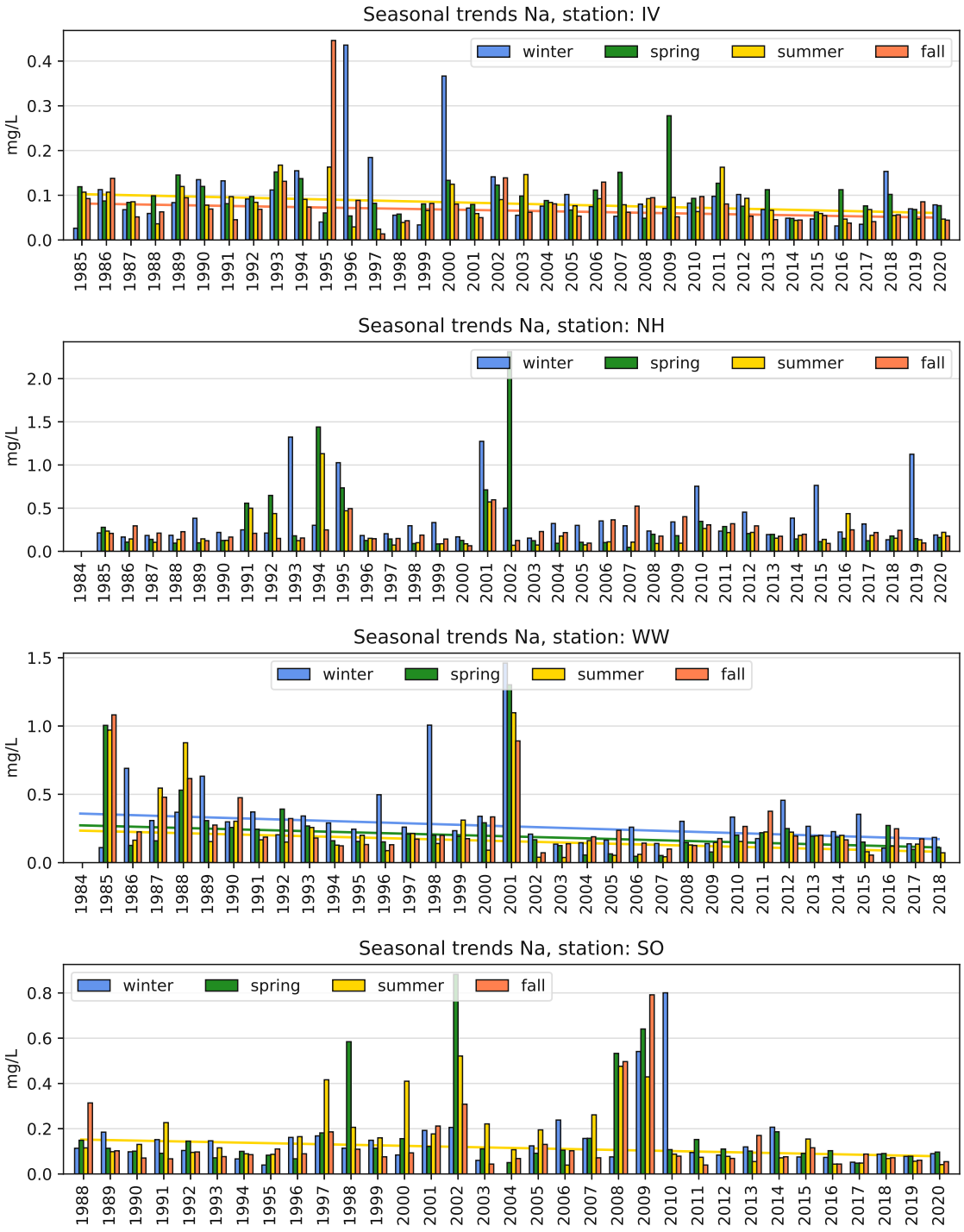


(d) Fig. D.1 (cont.): Grundlsee, Arnfels

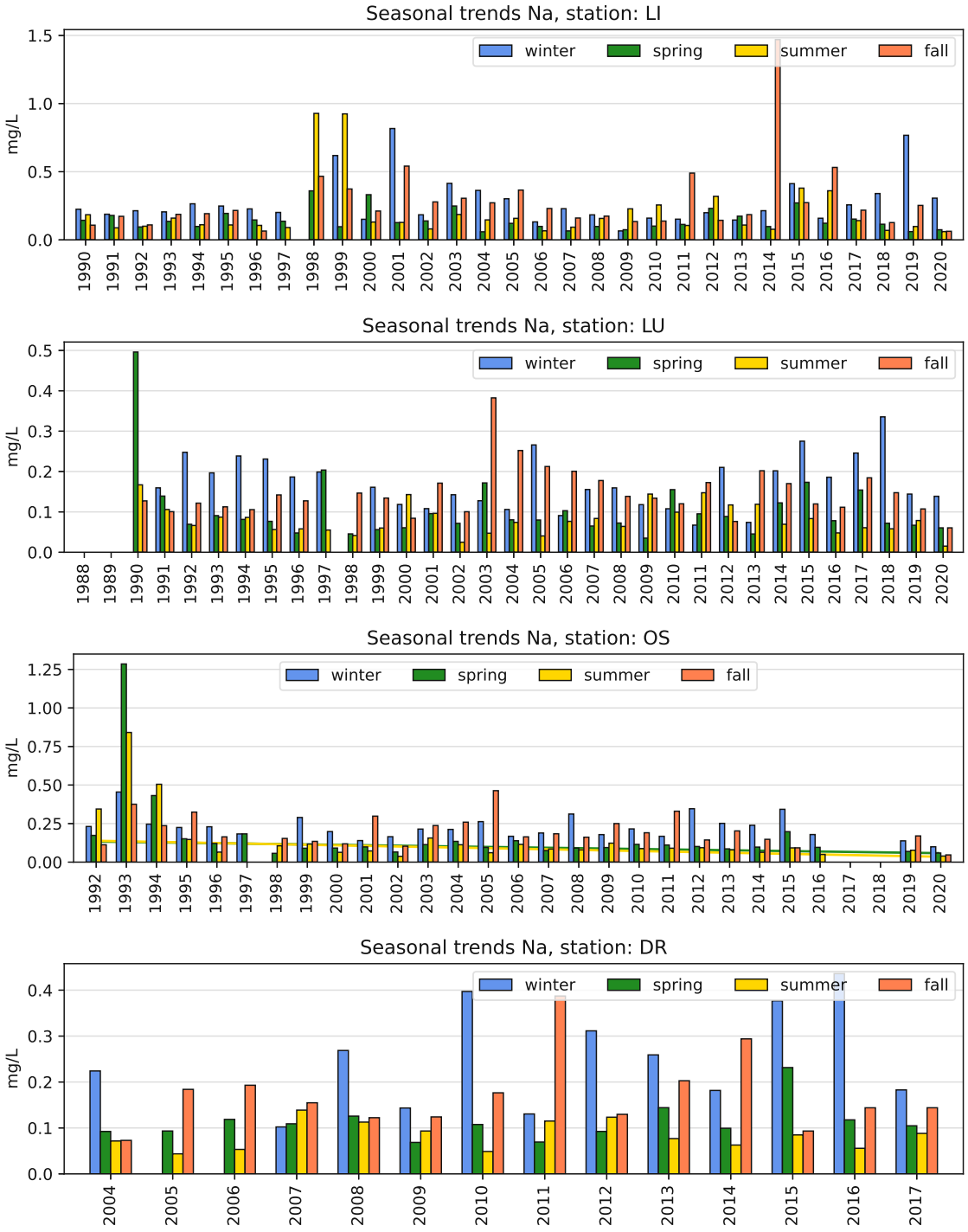


(a) Höfen, Niederndorferberg

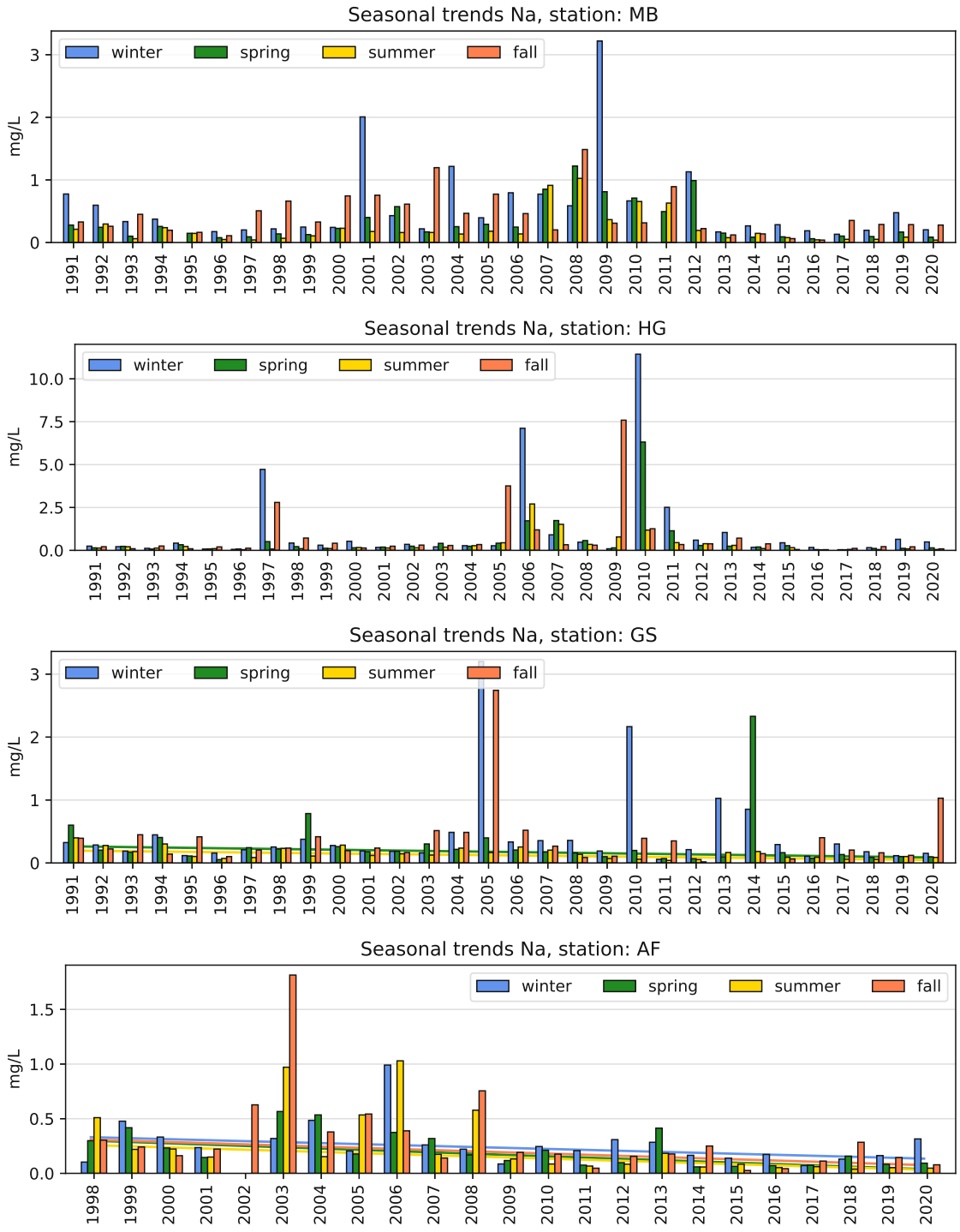
Figure D.2: Separated seasonal sodium concentration trends



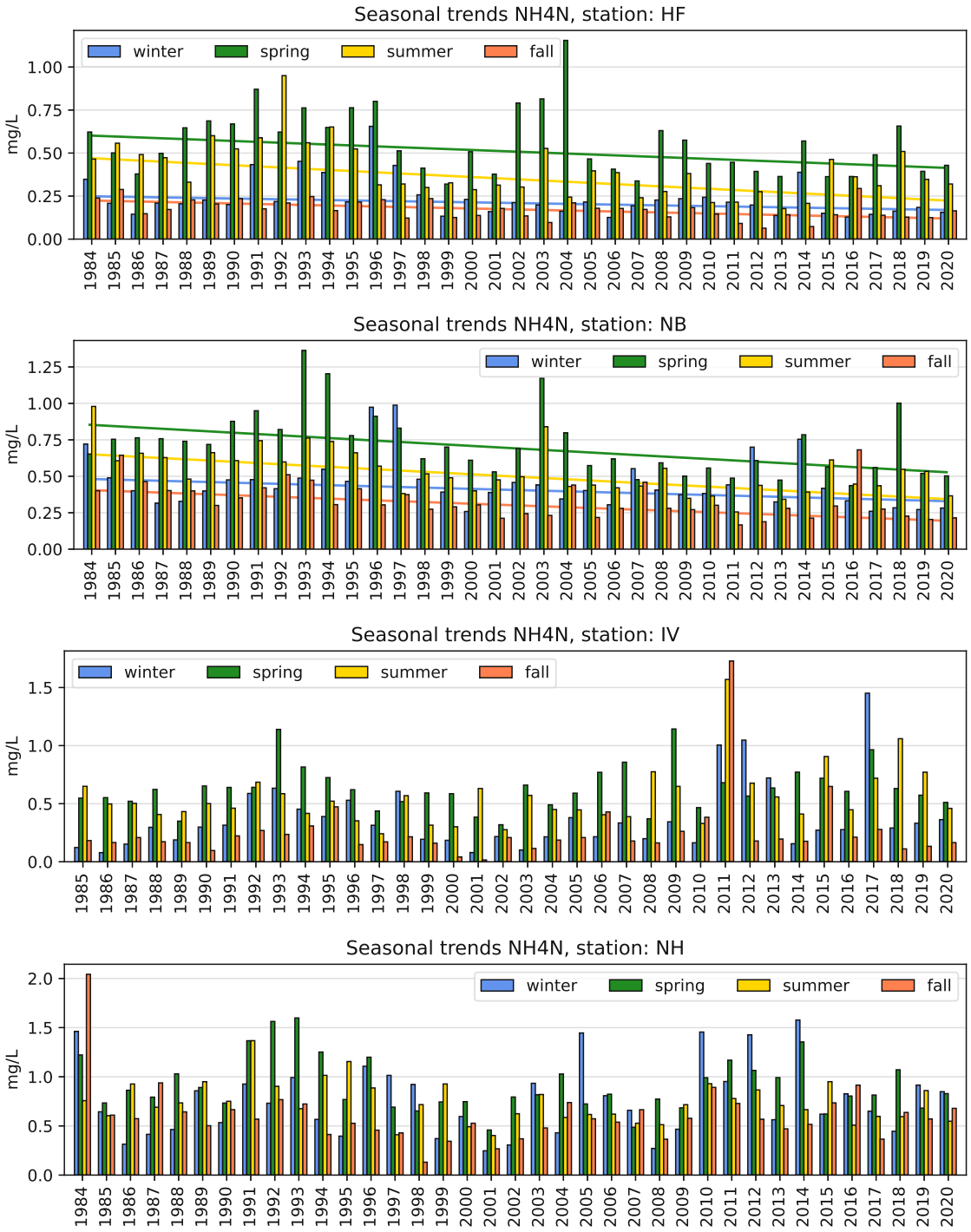
(b) Fig. D.2 (cont.): Innervillgraten, Haunsberg, Werfenweng, Sonnblick



(c) Fig. D.2 (cont.): Litschau, Lunz, Ostrong, Drasenhofen

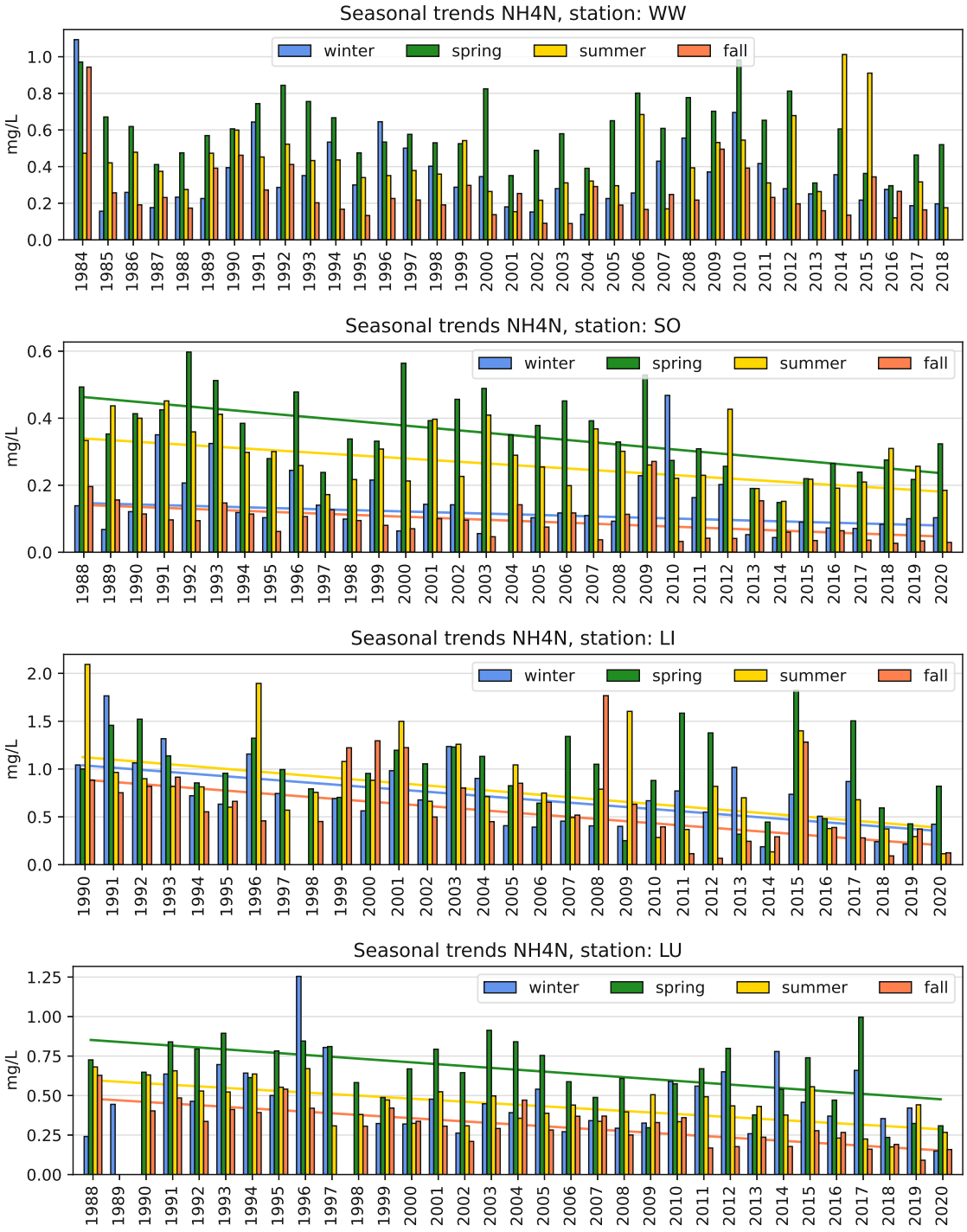


(d) Fig. D.2 (cont.): Masenberg, Hochgößnitz, Grundlsee, Arnfels

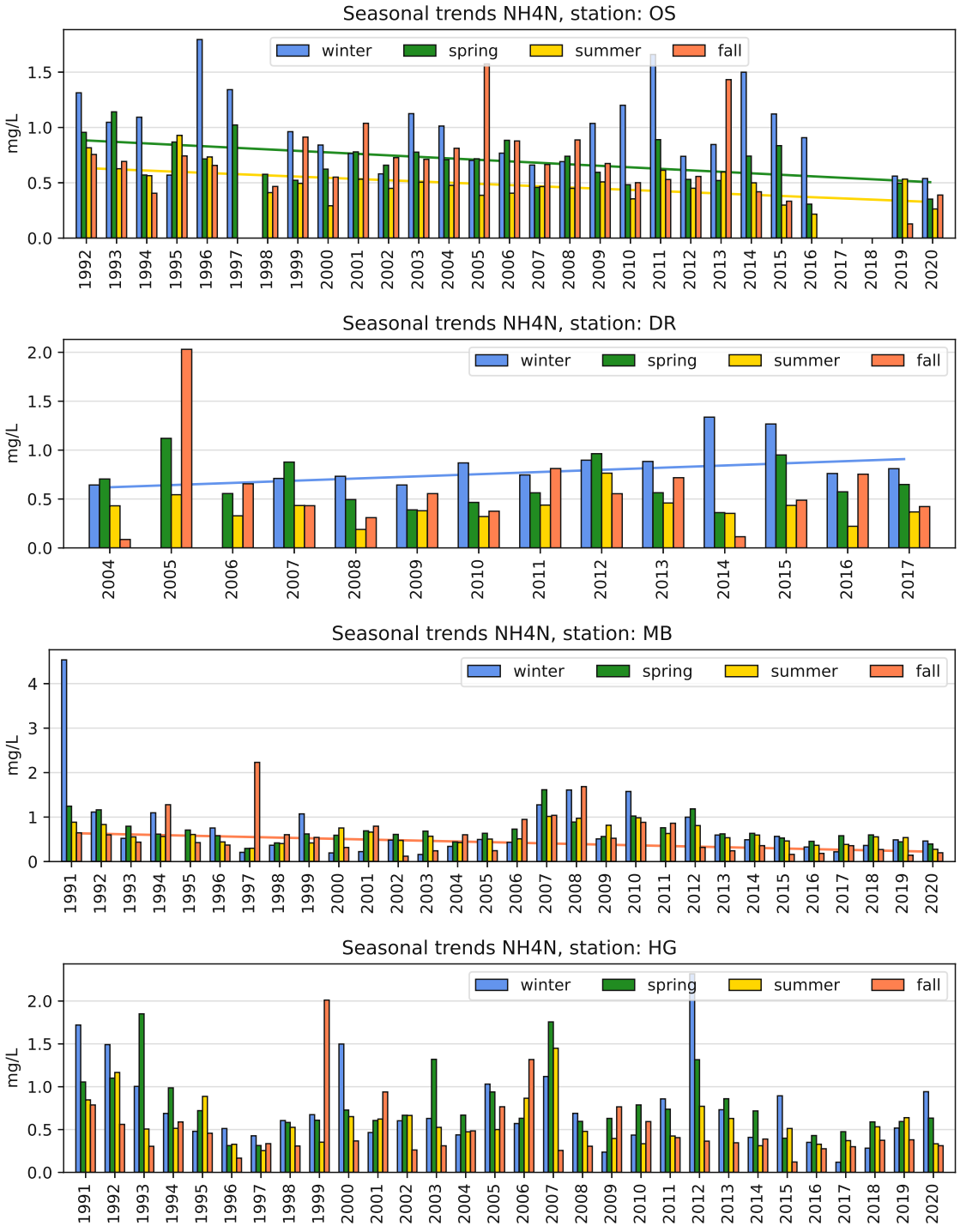


(a) Höfen, Niederndorferberg, Innervillgraten, Haunsberg

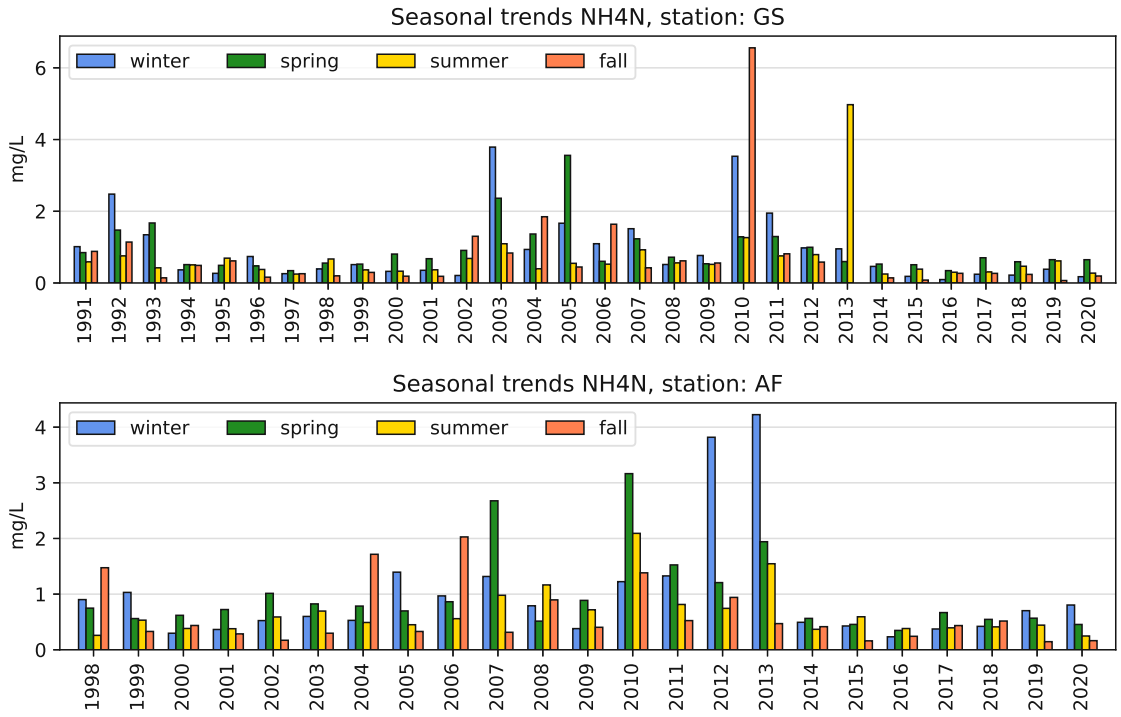
Figure D.3: Separated seasonal reduced nitrogen concentration trends



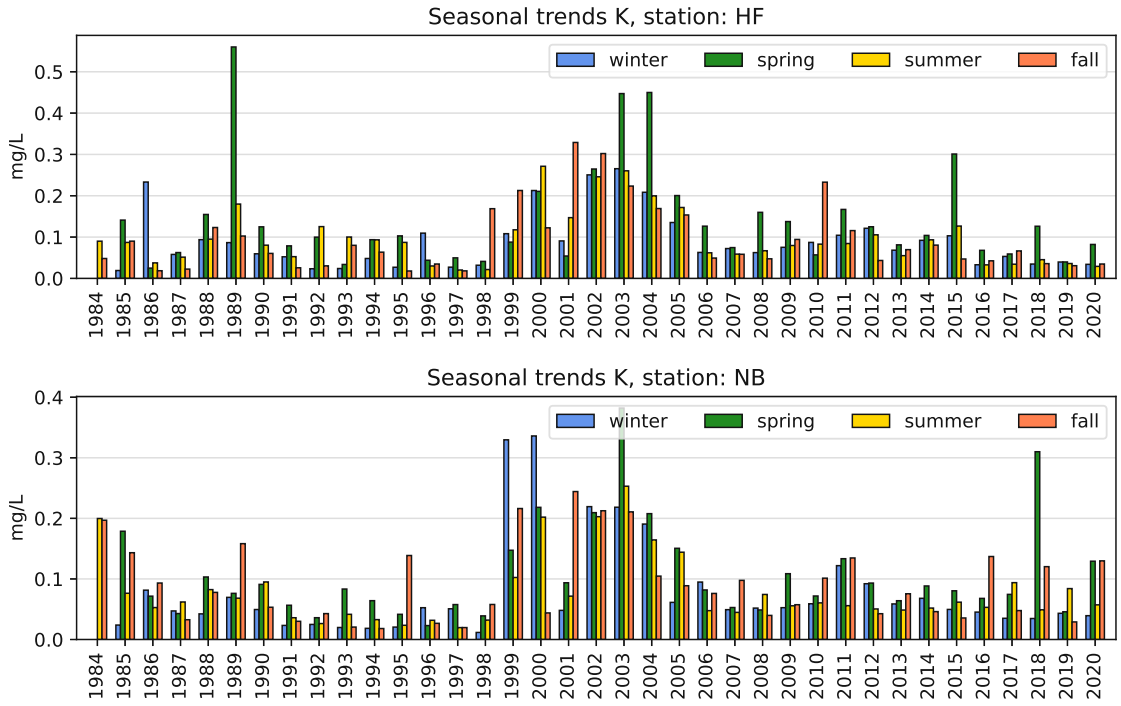
(b) Fig. D.3 (cont.): Werfenweng, Sonnblick, Litschau, Lunz



(c) Fig. D.3 (cont.): Ostrong, Drasenhofen, Masenberg, Hochgößnitz

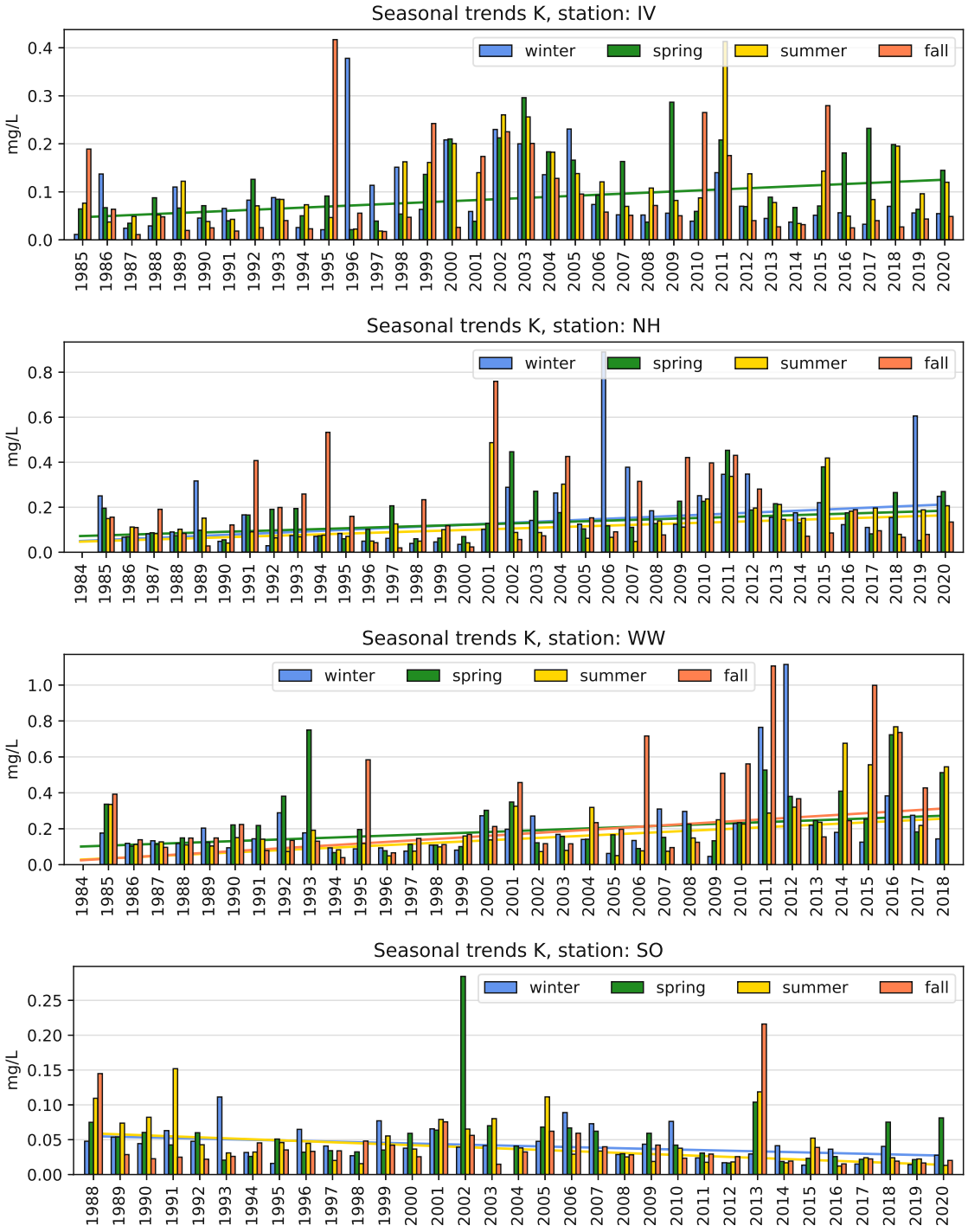


(d) Fig. D.3 (cont.): Grundlsee, Arnfels



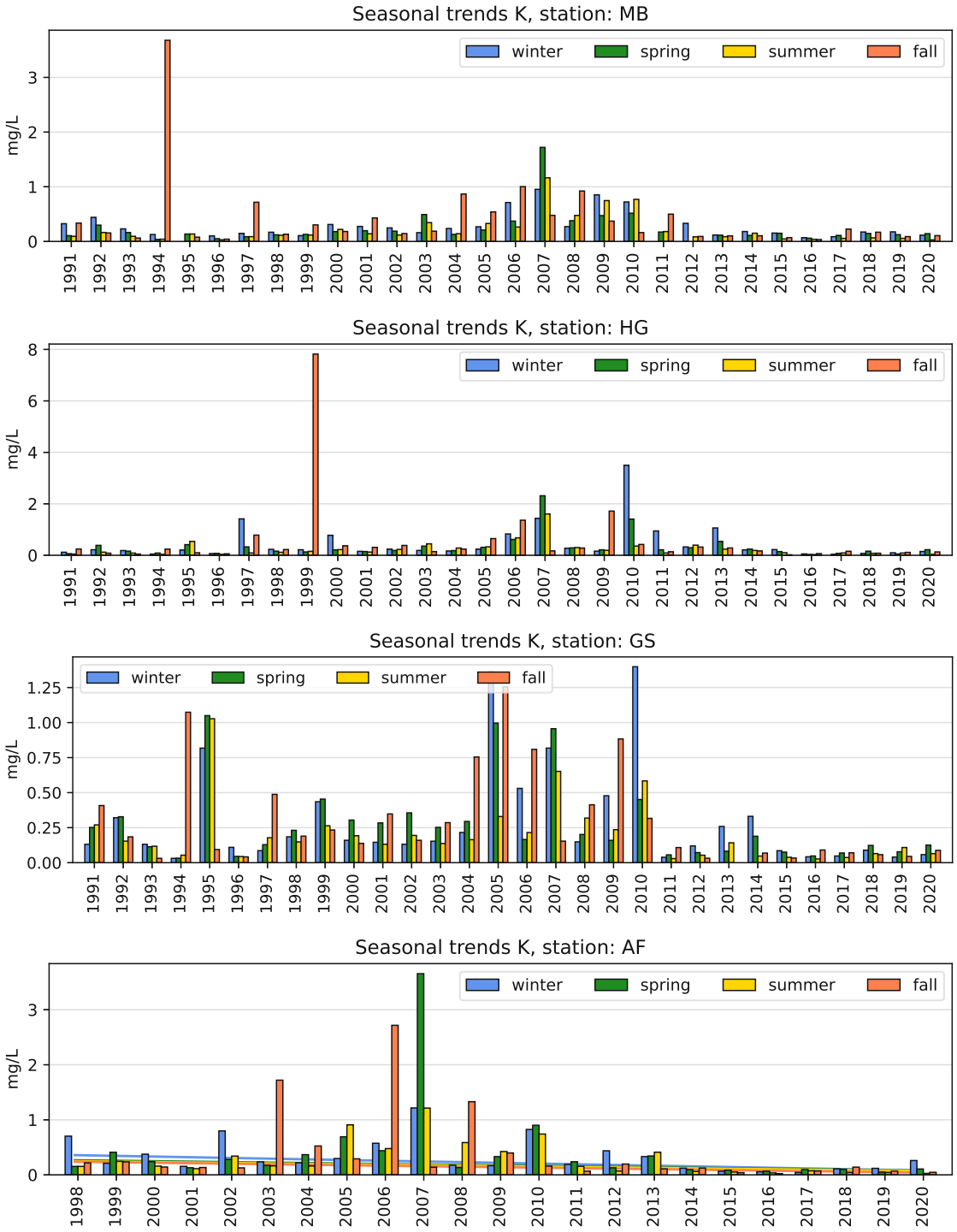
(a) Höfen, Niederndorferberg

Figure D.4: Separated seasonal potassium concentration trends

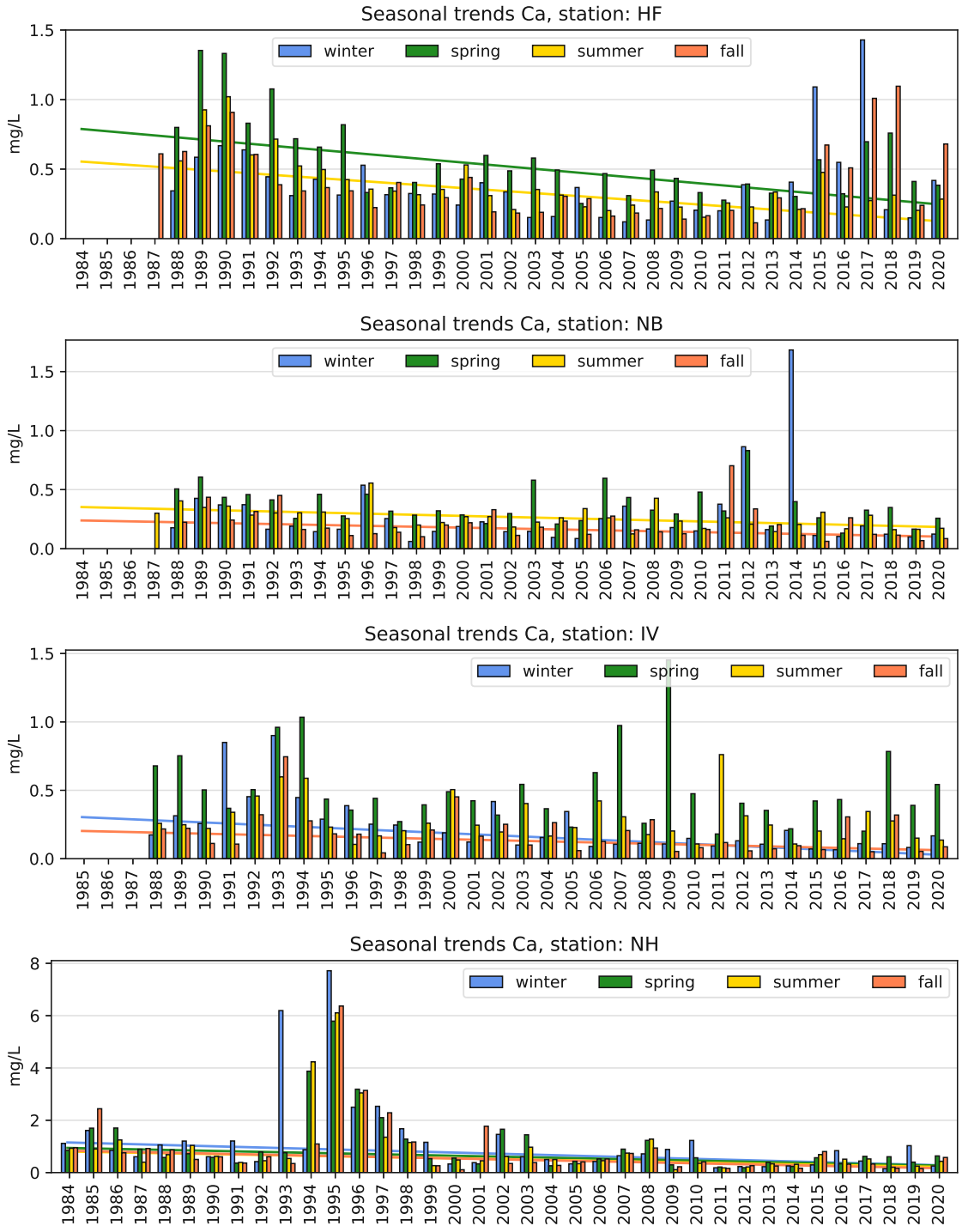


(b) Fig. D.4 (cont.): Innervillgraten, Haunsberg, Werfenweng, Sonnblick

(c) Fig. D.4 (cont.): Litschau, Lunz, Ostrong, Drasenhofen

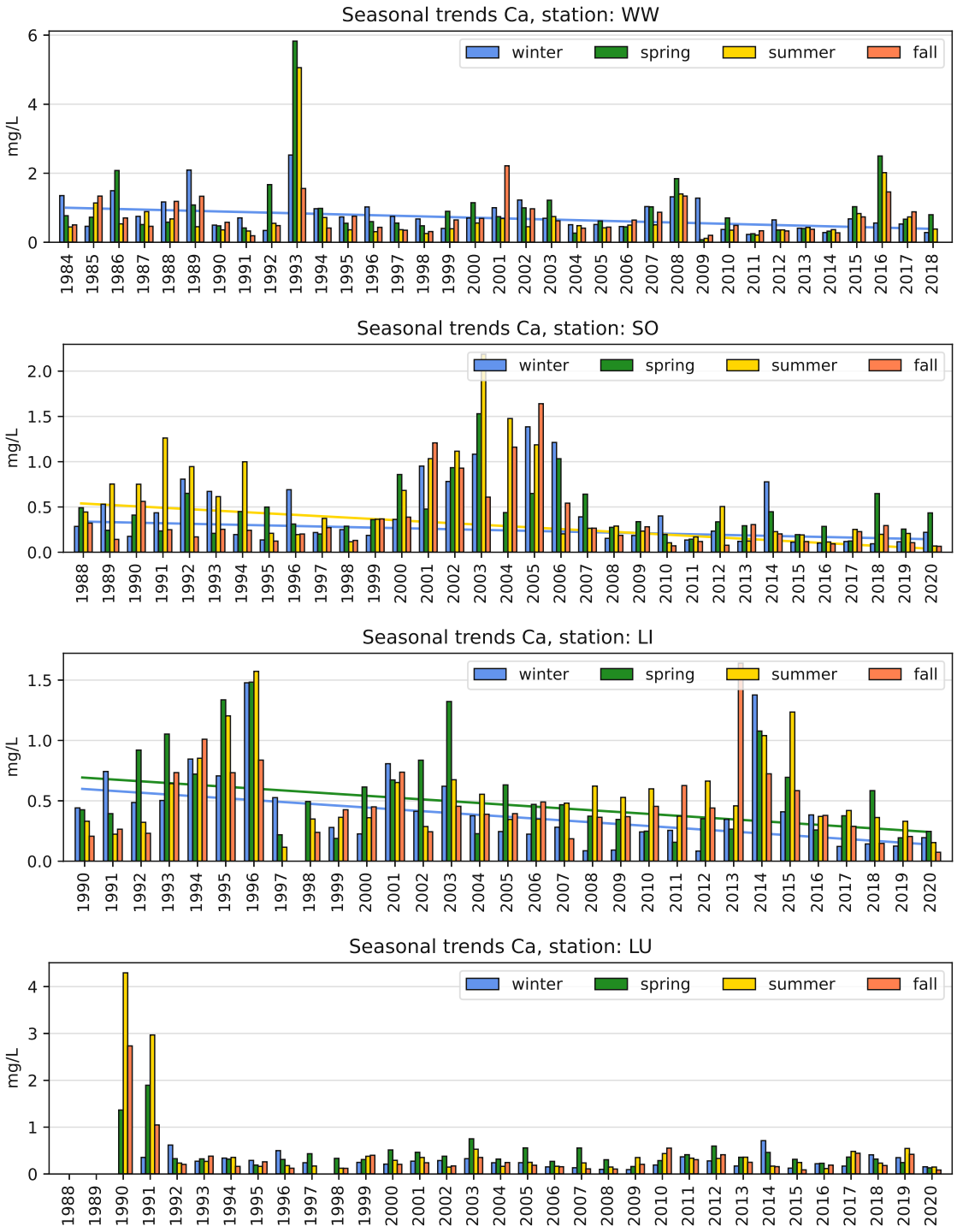


(d) Fig. D.4 (cont.): Masenberg, Hochgößnitz, Grundlsee, Arnfels

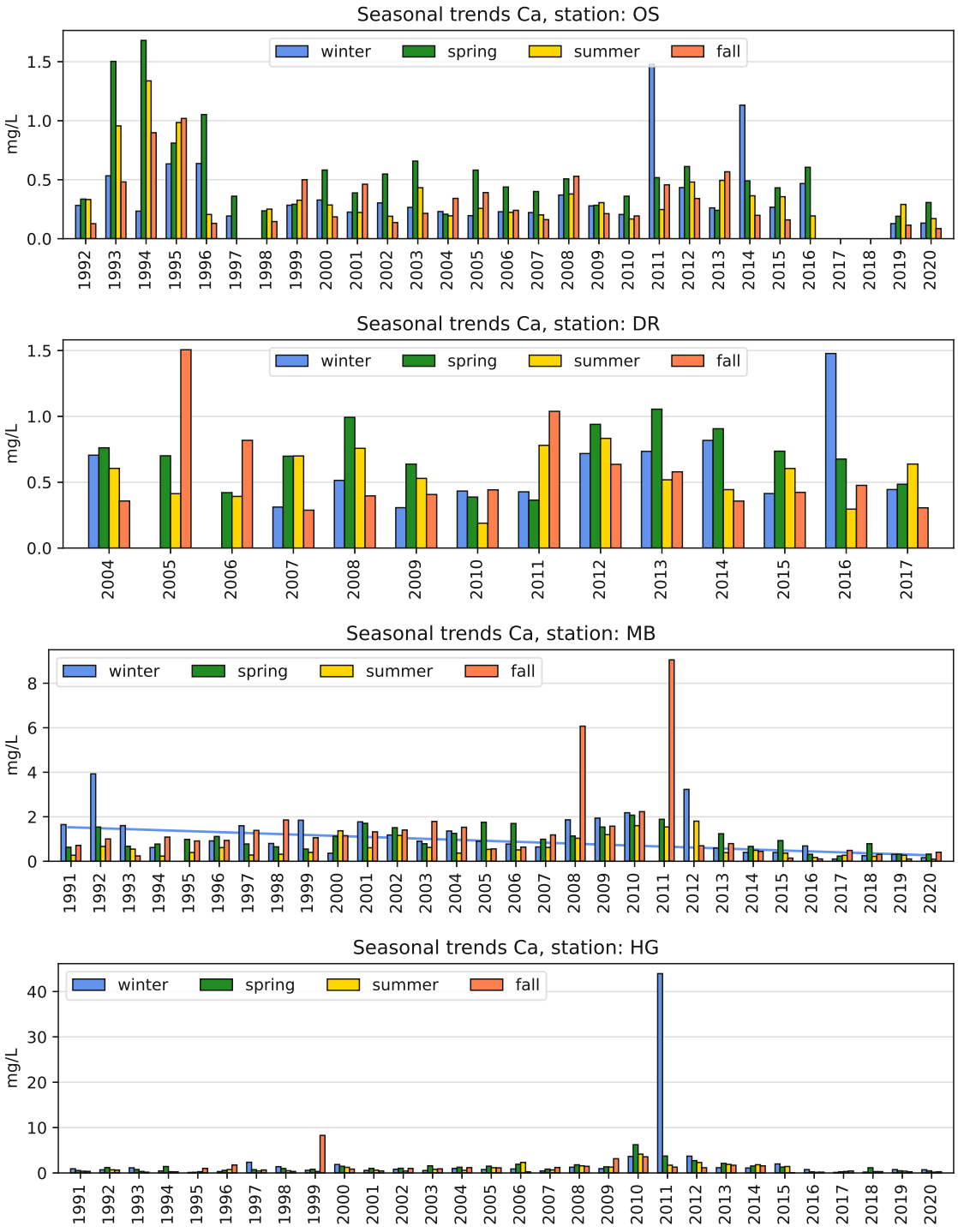


(a) Höfen, Niederndorferberg, Innervillgraten, Haunsberg

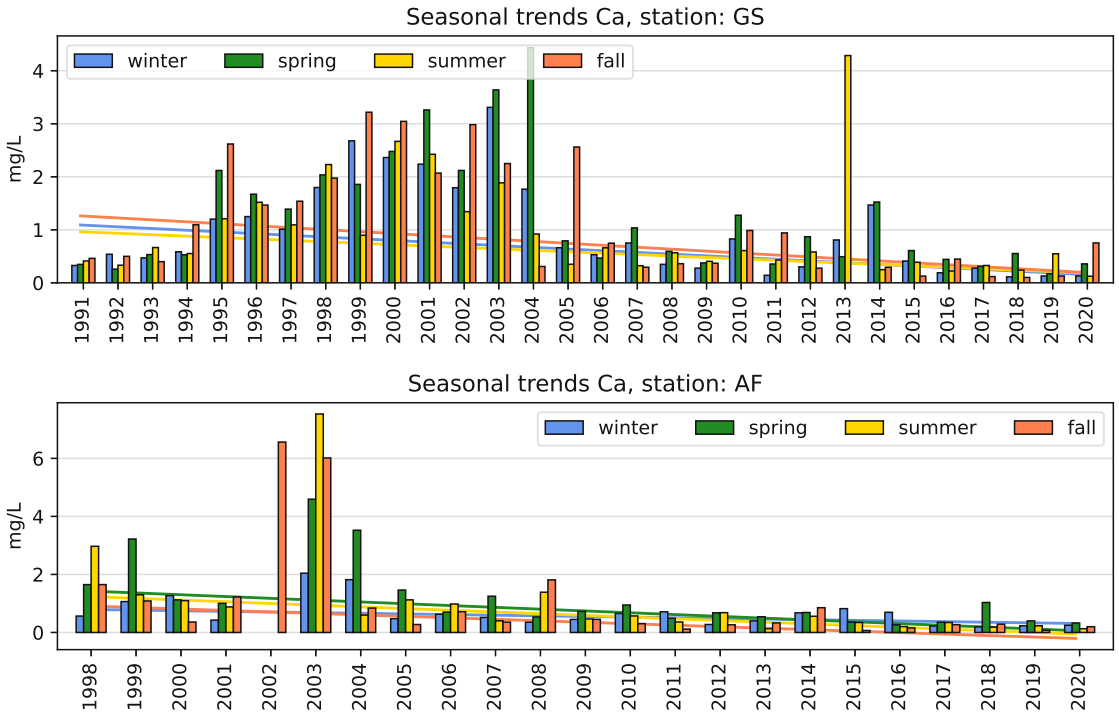
Figure D.5: Separated seasonal calcium concentration trends



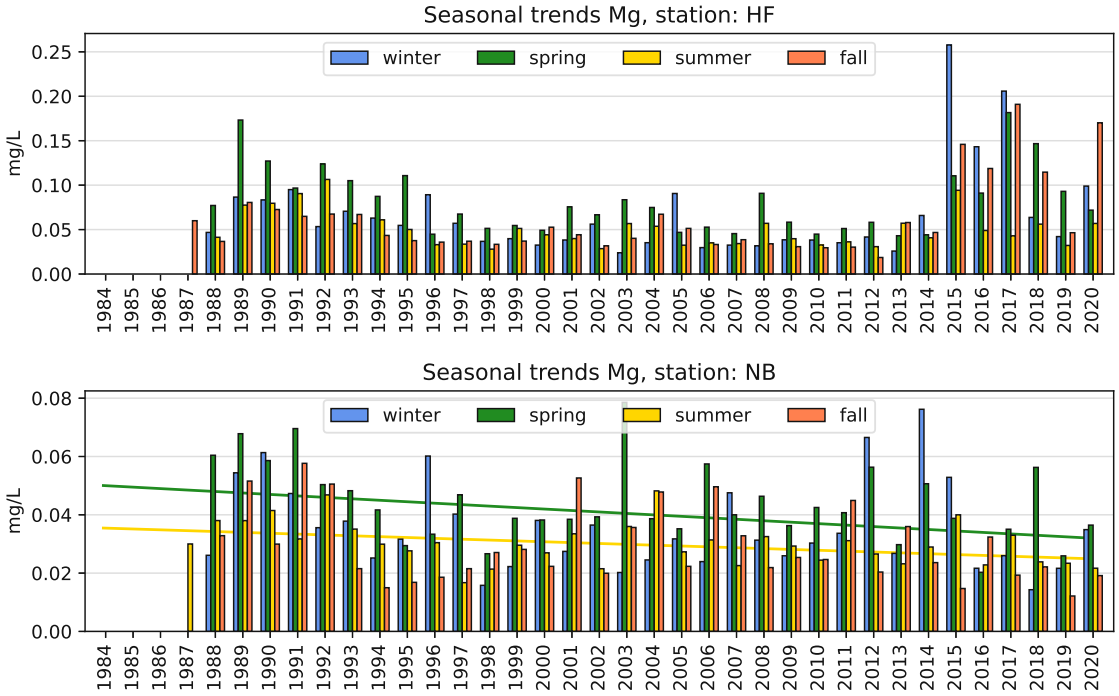
(b) Fig. D.5 (cont.): Werfenweng, Sonnblick, Litschau, Lunz



(c) Fig. D.5 (cont.): Ostrong, Drasenhofen, Masenberg, Hochgößnitz

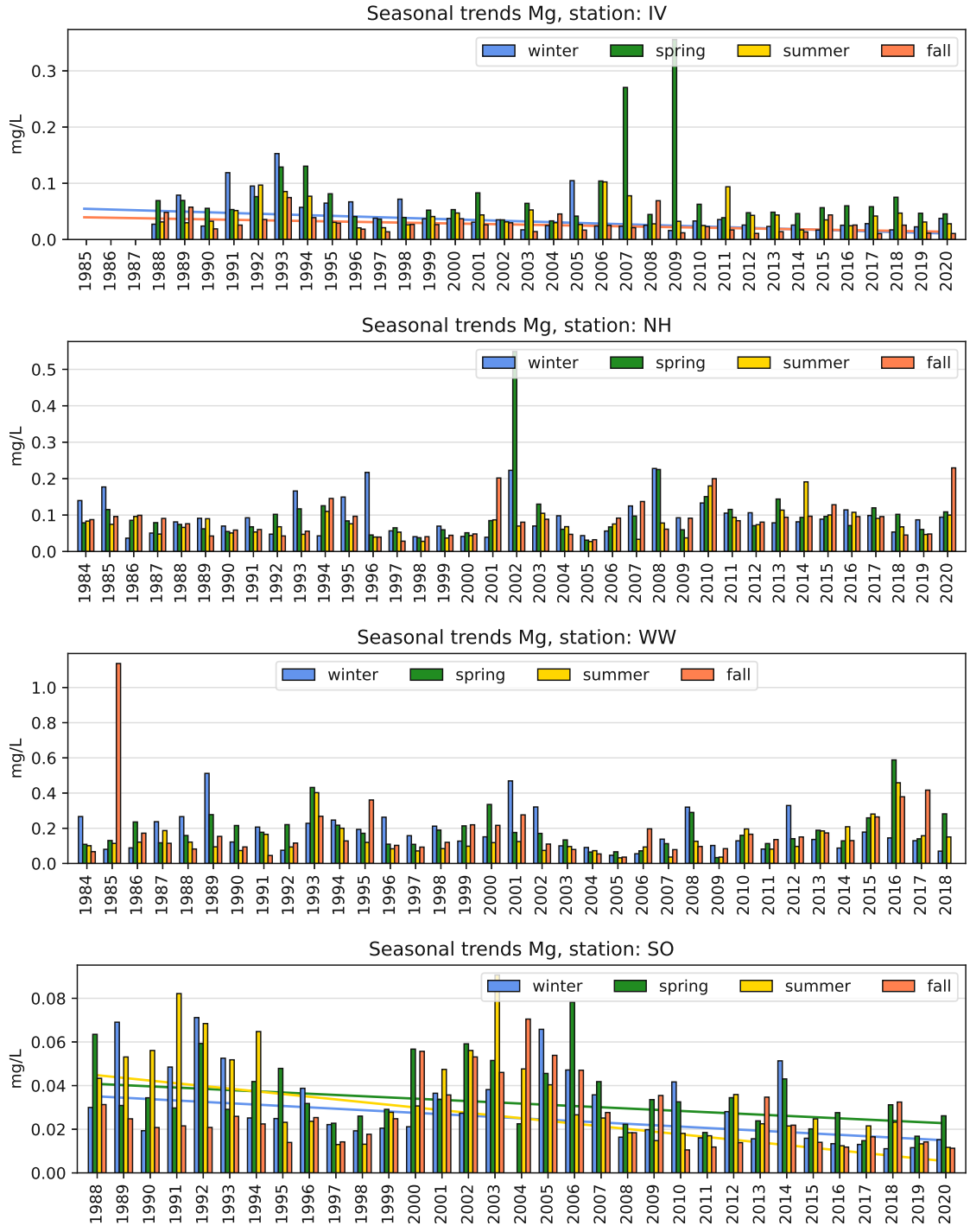


(d) Fig. D.5 (cont.): Grundlsee, Arnfels



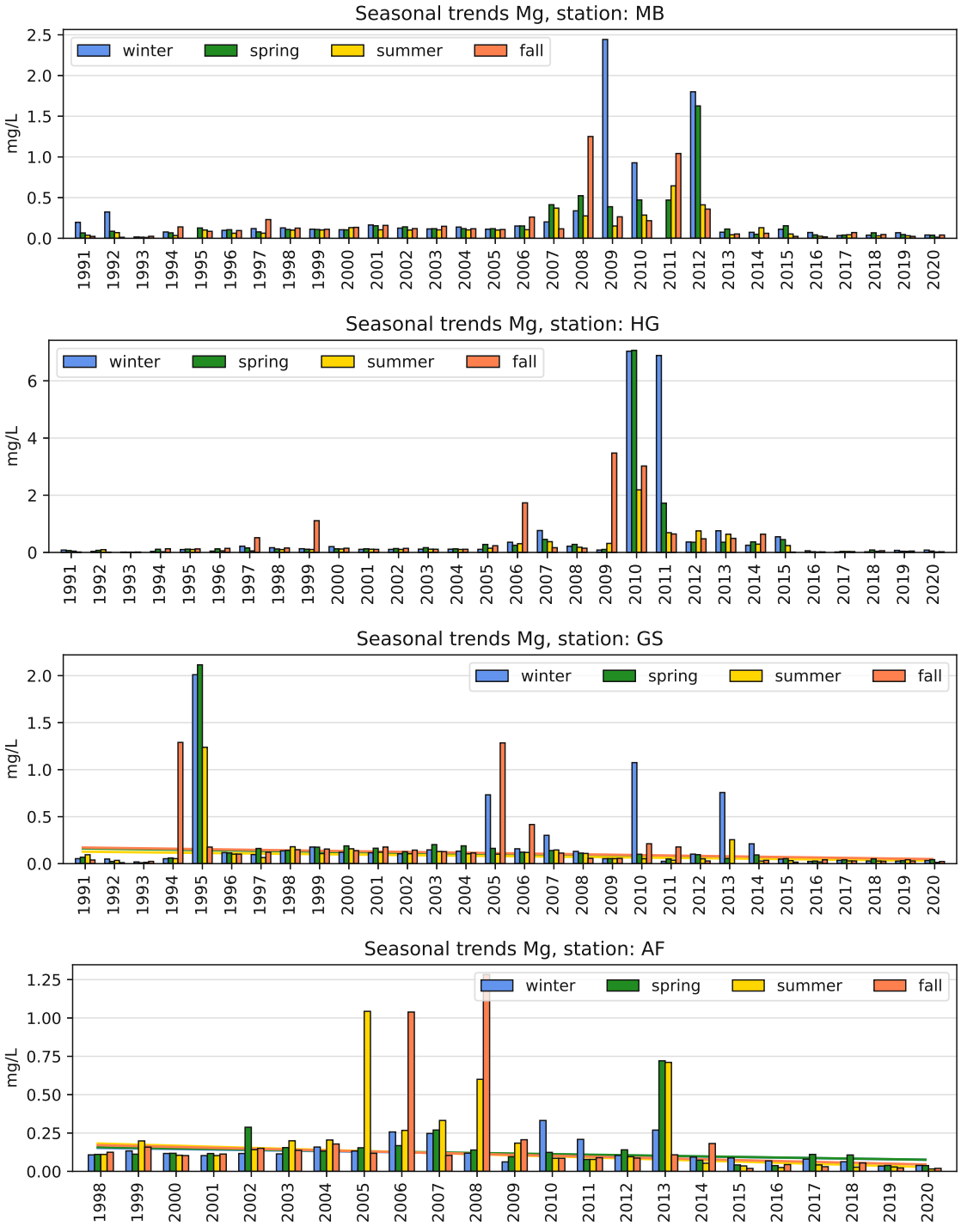
(a) Höfen, Niederndorferberg

Figure D.6: Separated seasonal magnesium concentration trends

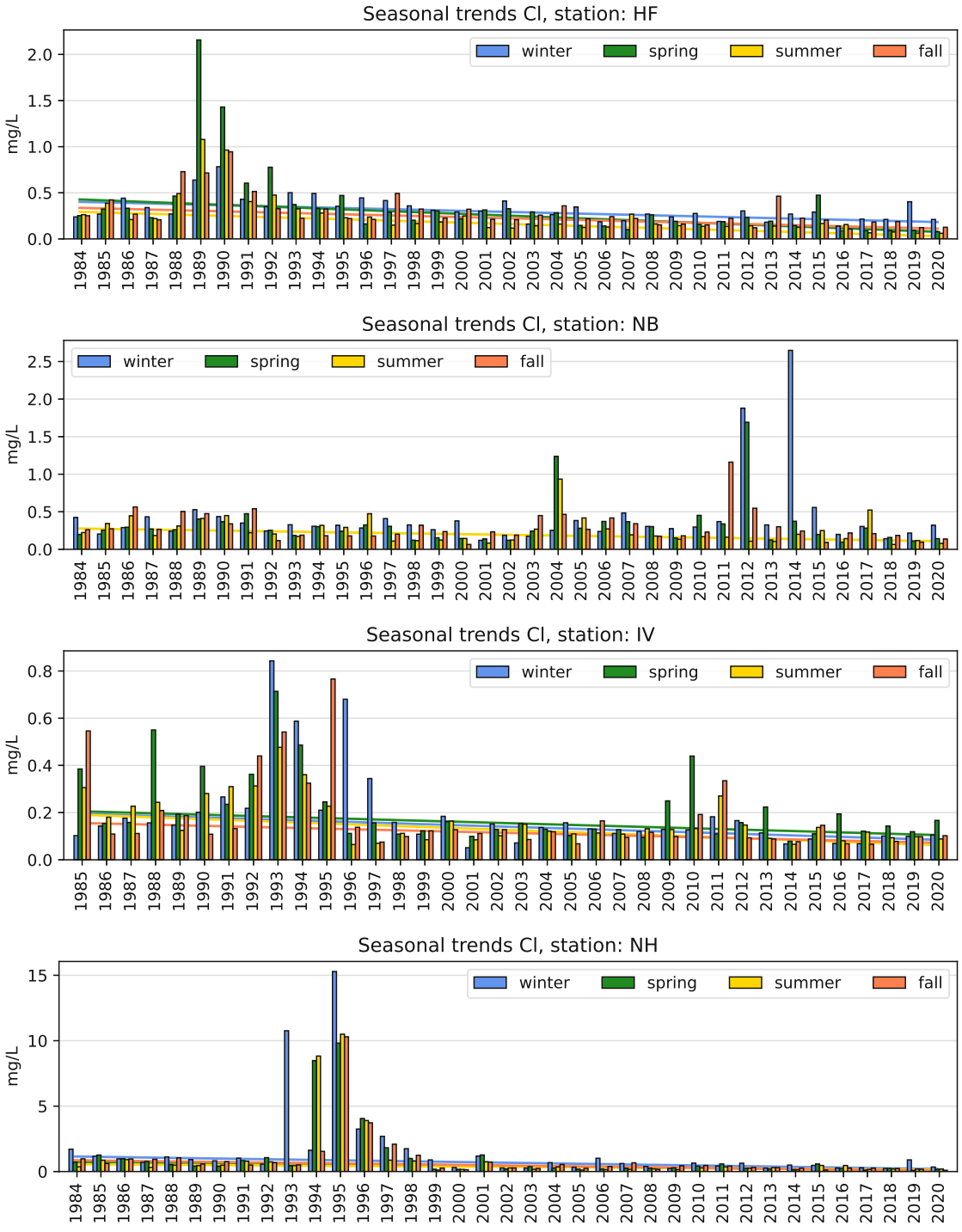


(b) Fig. D.6 (cont.): Innervillgraten, Haunsberg, Werfenweng, Sonnblick

(c) Fig. D.6 (cont.): Litschau, Lunz, Ostrong, Drasenhofen

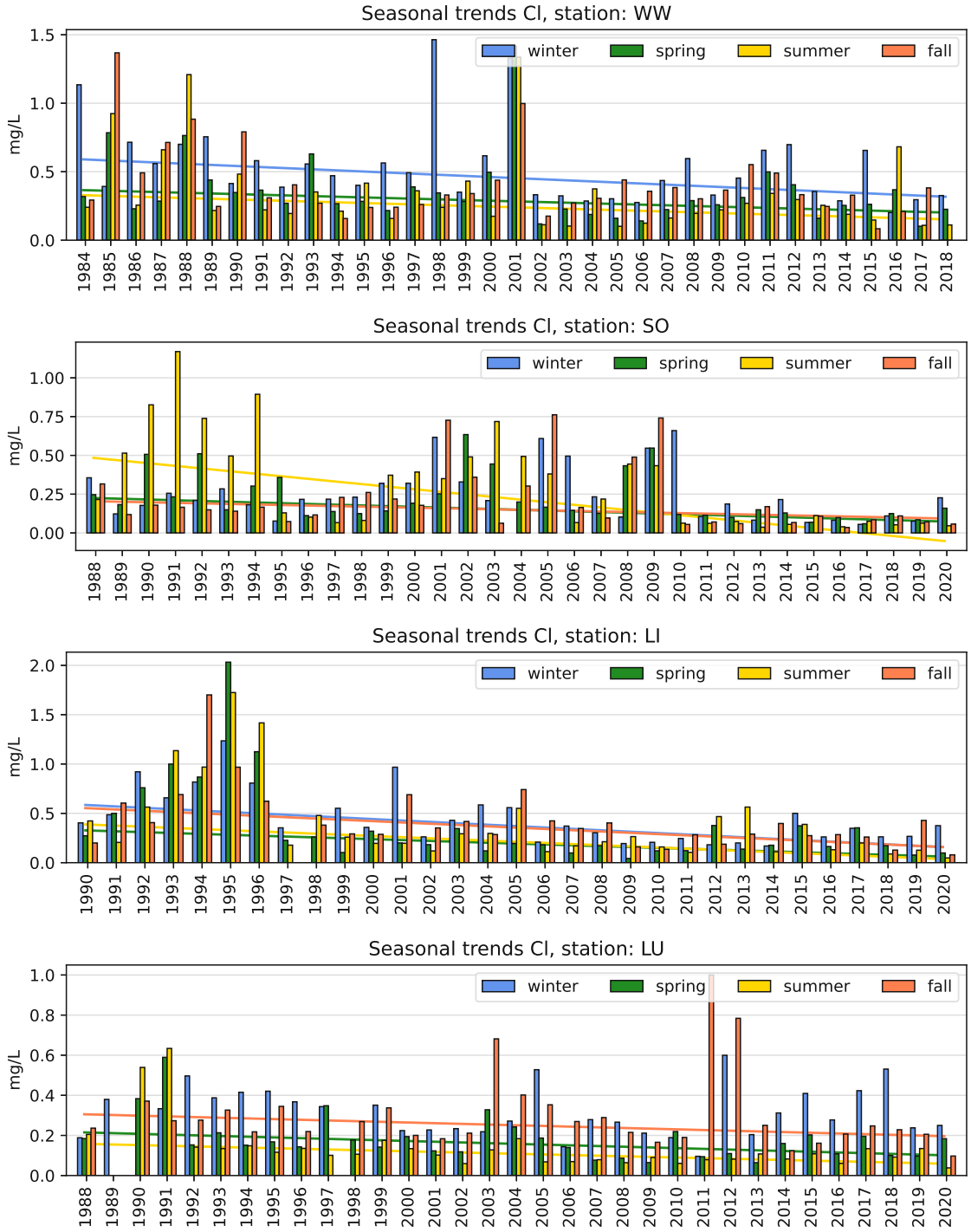


(d) Fig. D.6 (cont.): Masenberg, Hochgößnitz, Grundlsee, Arnfels

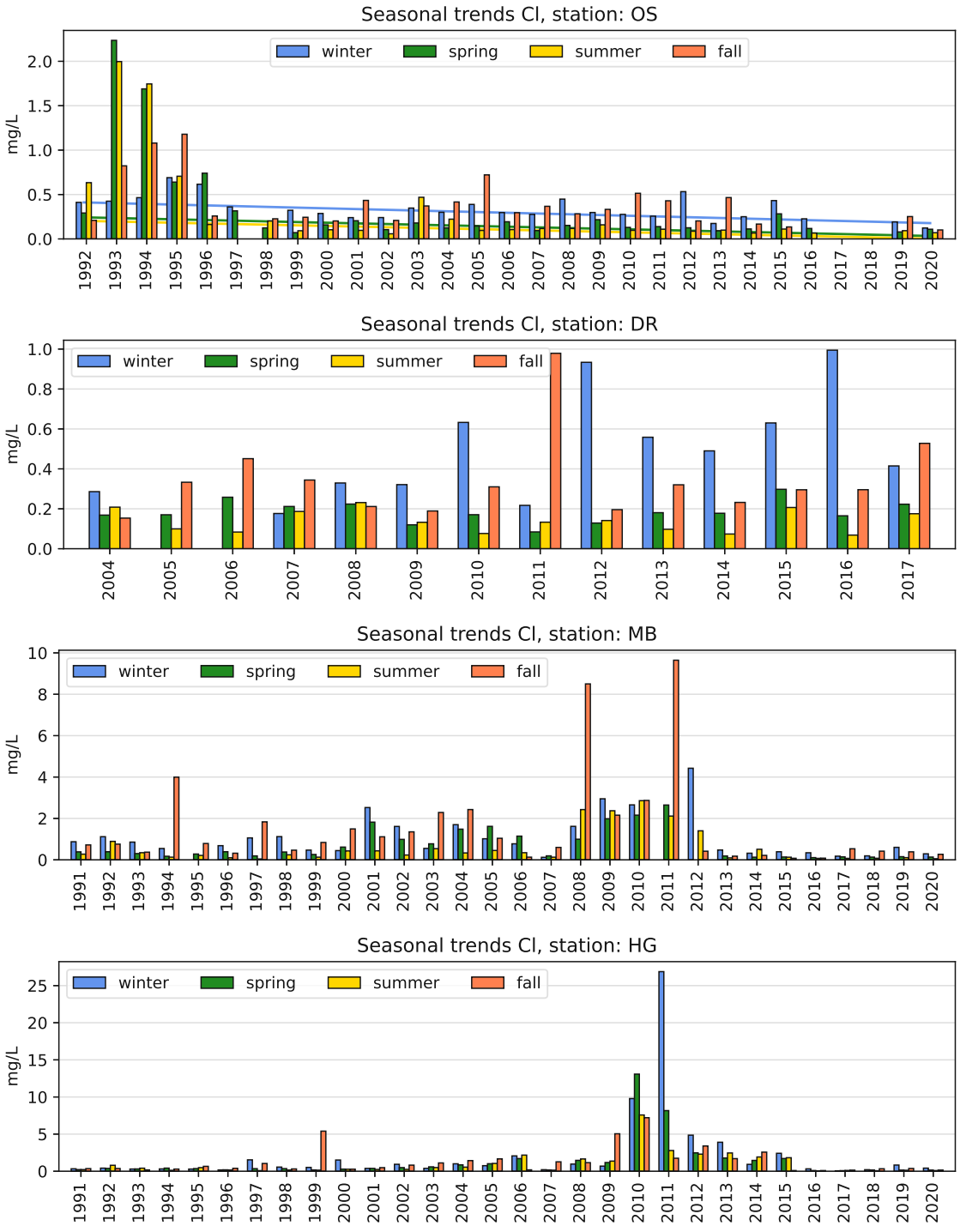


(a) Höfen, Niederndorferberg, Innervillgraten, Haunsberg

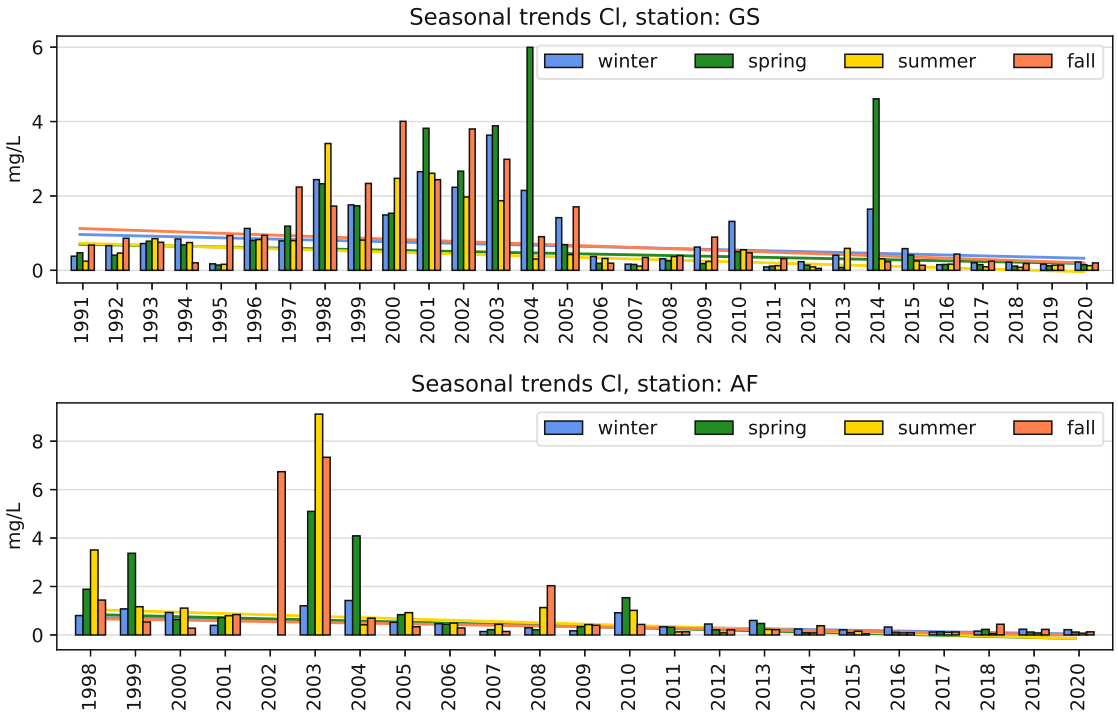
Figure D.7: Separated seasonal chlorid concentration trends



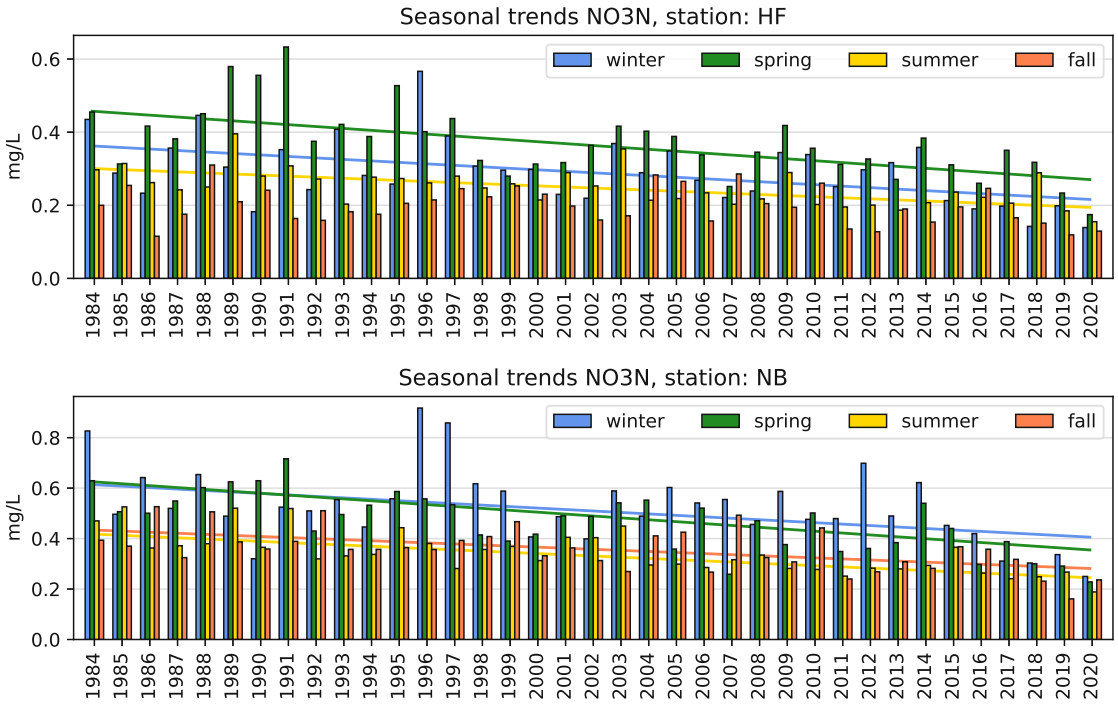
(b) Fig. D.7 (cont.): Werfenweng, Sonnblick, Litschau, Lunz



(c) Fig. D.7 (cont.): Ostrong, Drasenhofen, Masenberg, Hochgößnitz

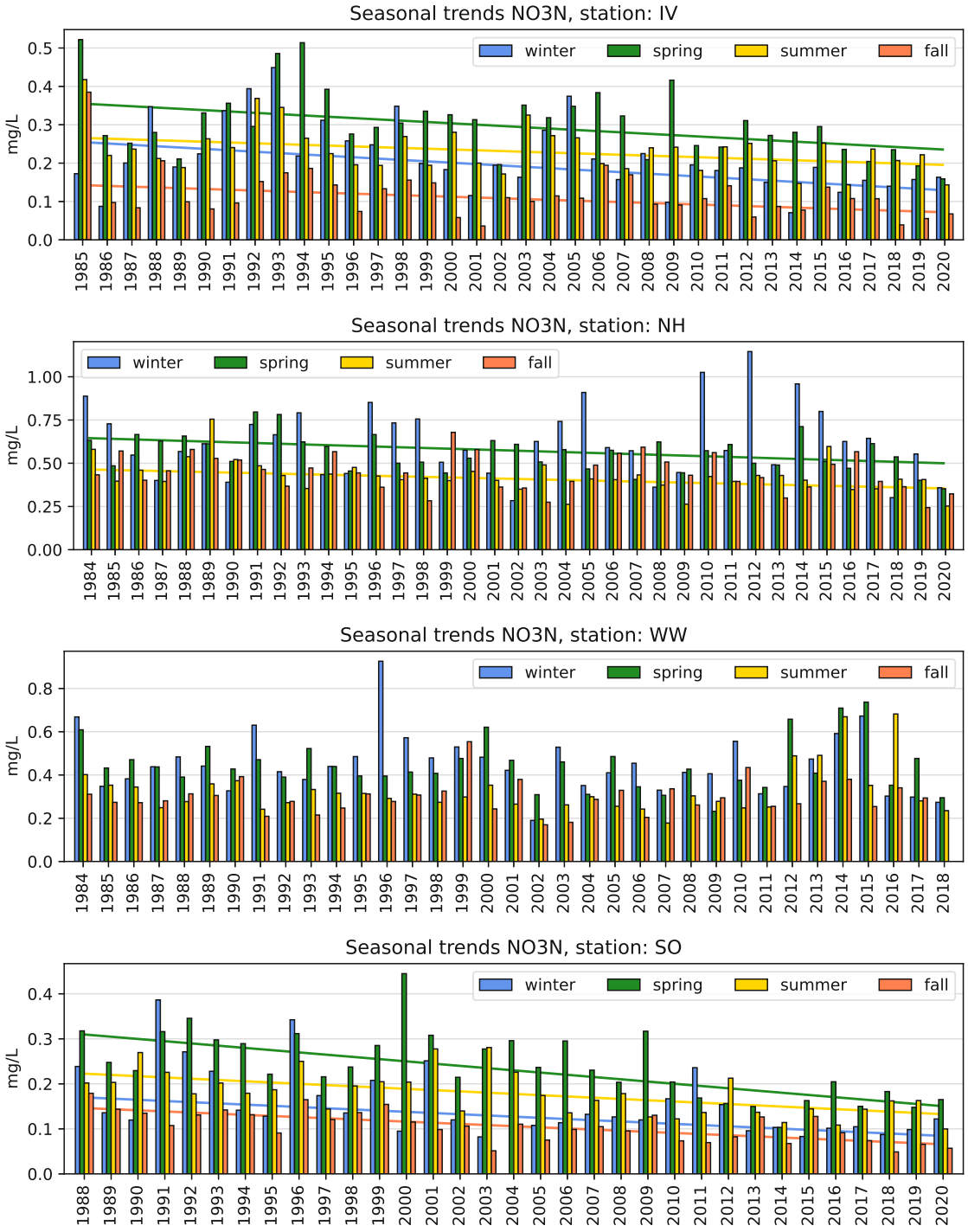


(d) Fig. D.7 (cont.): Grundlsee, Arnfels

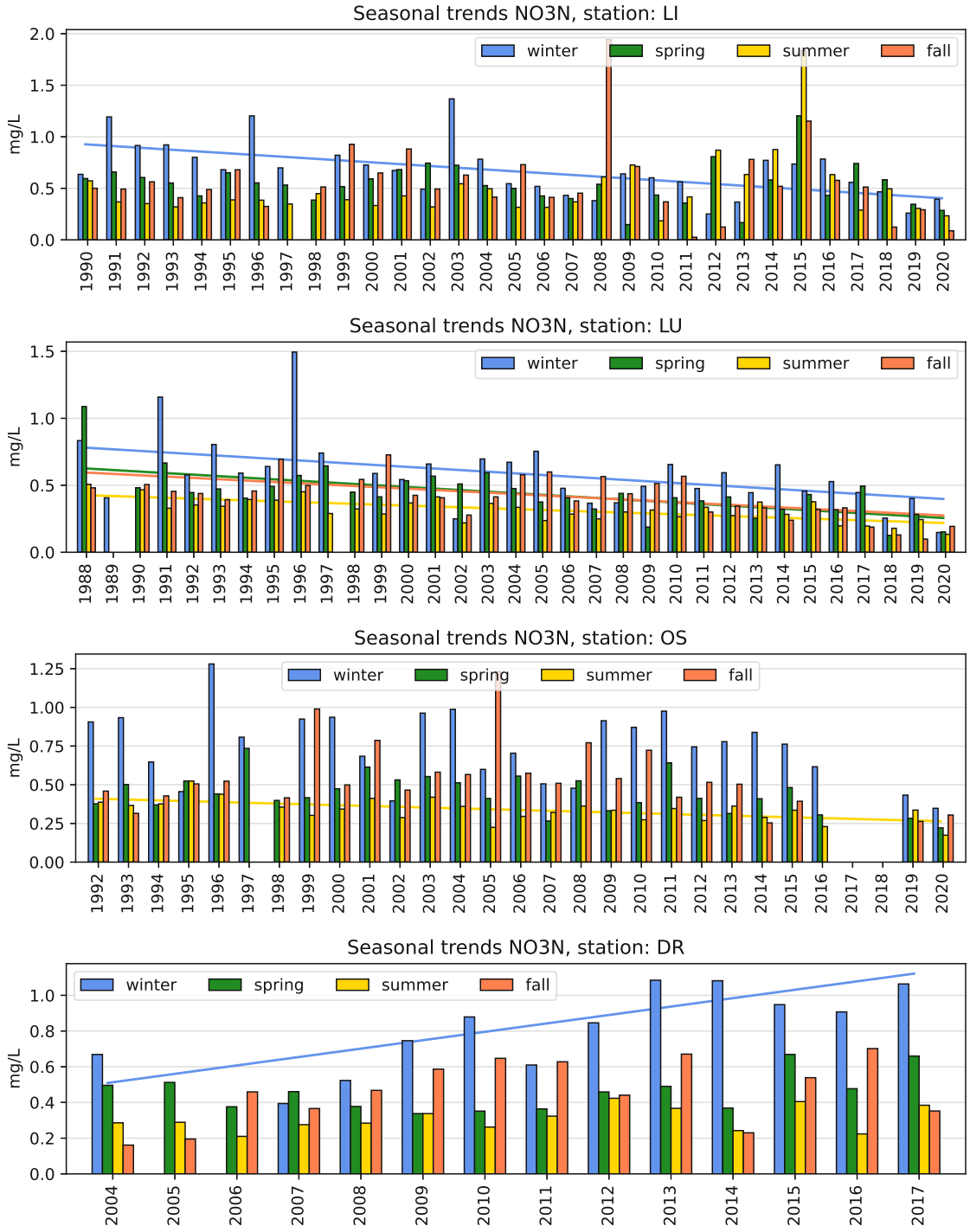


(a) Höfen, Niederndorferberg

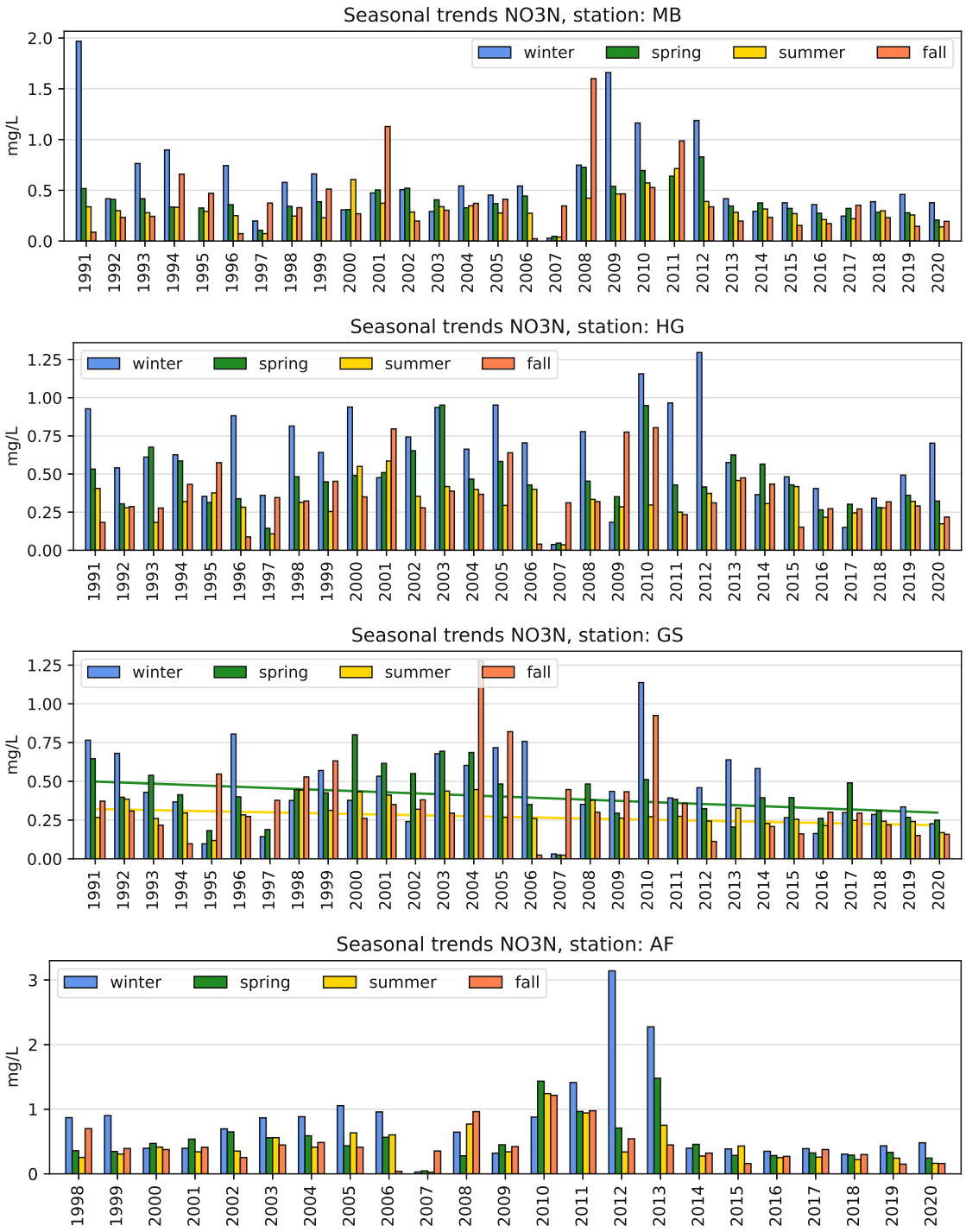
Figure D.8: Separated seasonal oxidized nitrogen concentration trends



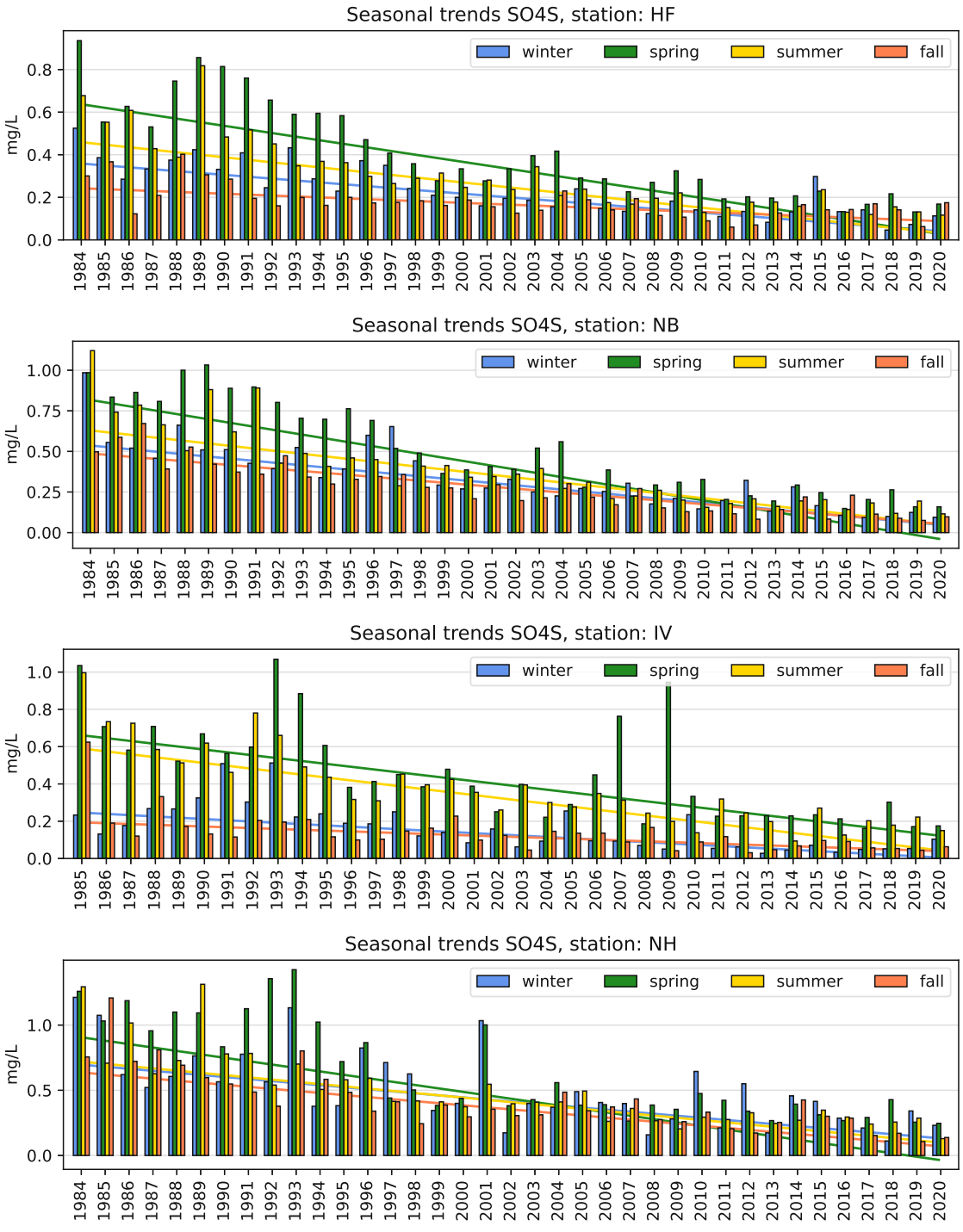
(b) Fig. D.8 (cont.): Innervillgraten, Haunsberg, Werfenweng, Sonnblick



(c) Fig. D.8 (cont.): Litschau, Lunz, Ostrong, Drasenhofen

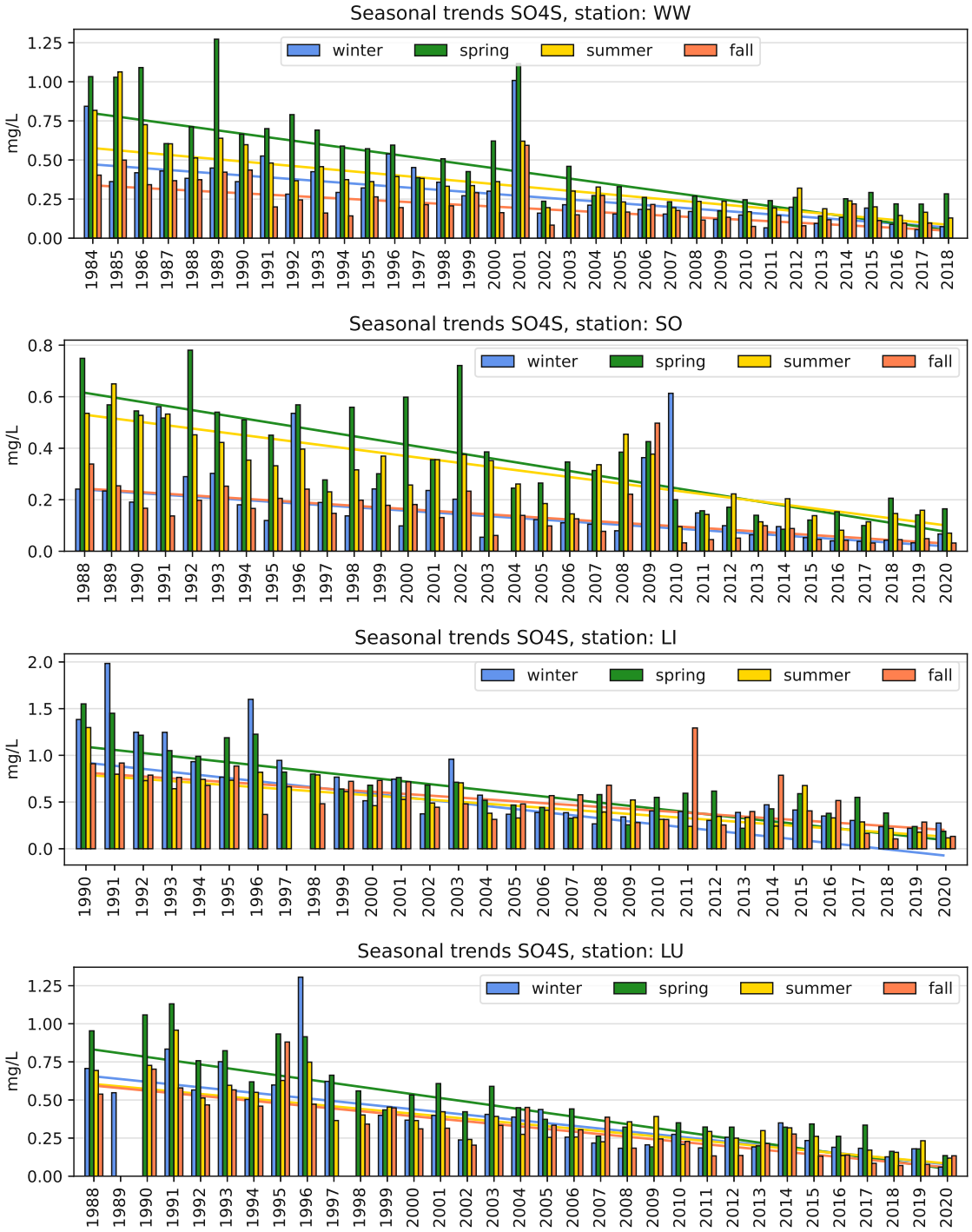


(d) Fig. D.8 (cont.): Masenberg, Hochgößnitz, Grundlsee, Arnfels

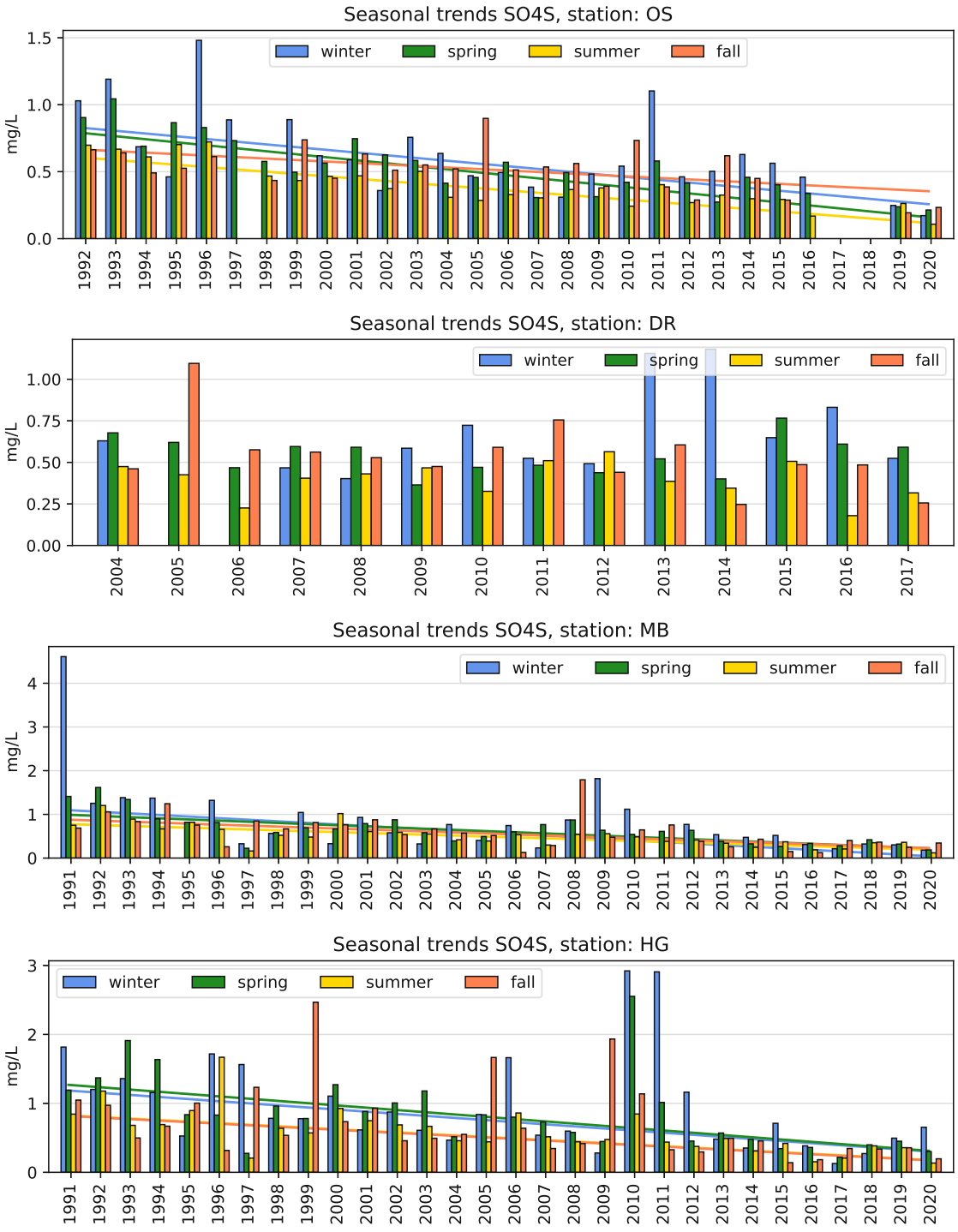


(a) Höfen, Niederndorferberg, Innervillgraten, Haunsberg

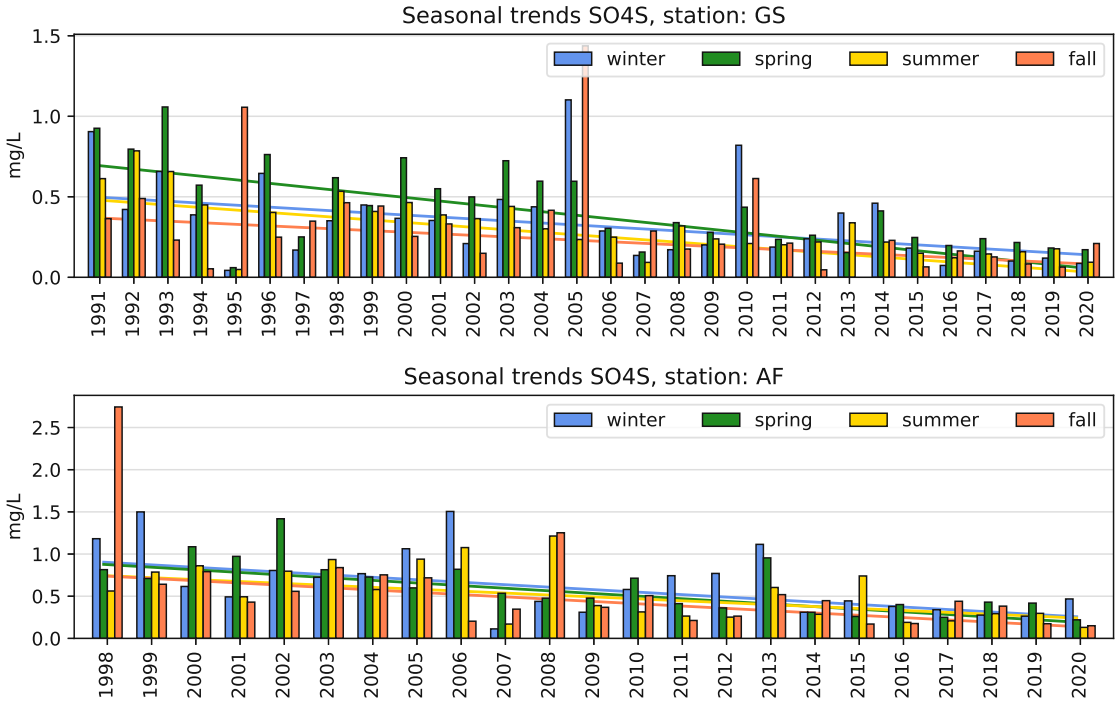
Figure D.9: Separated seasonal sulfur concentration trends



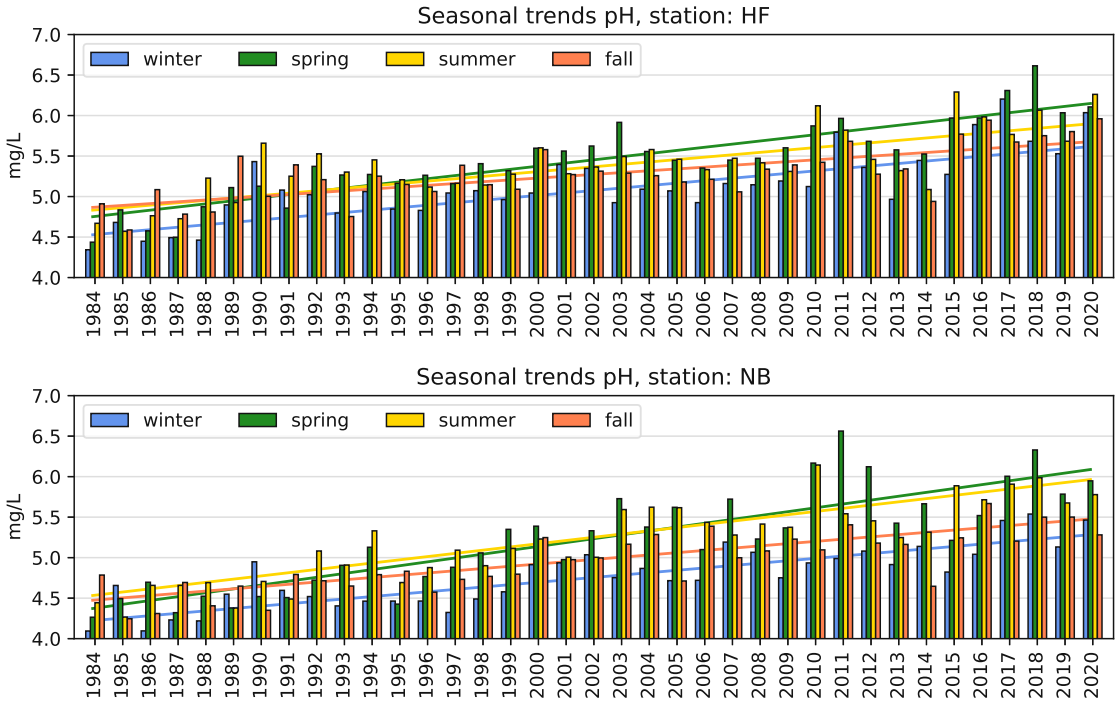
(b) Fig. D.9 (cont.): Werfenweng, Sonnblick, Litschau, Lunz



(c) Fig. D.9 (cont.): Ostrong, Drasenhofen, Masenberg, Hochgößnitz

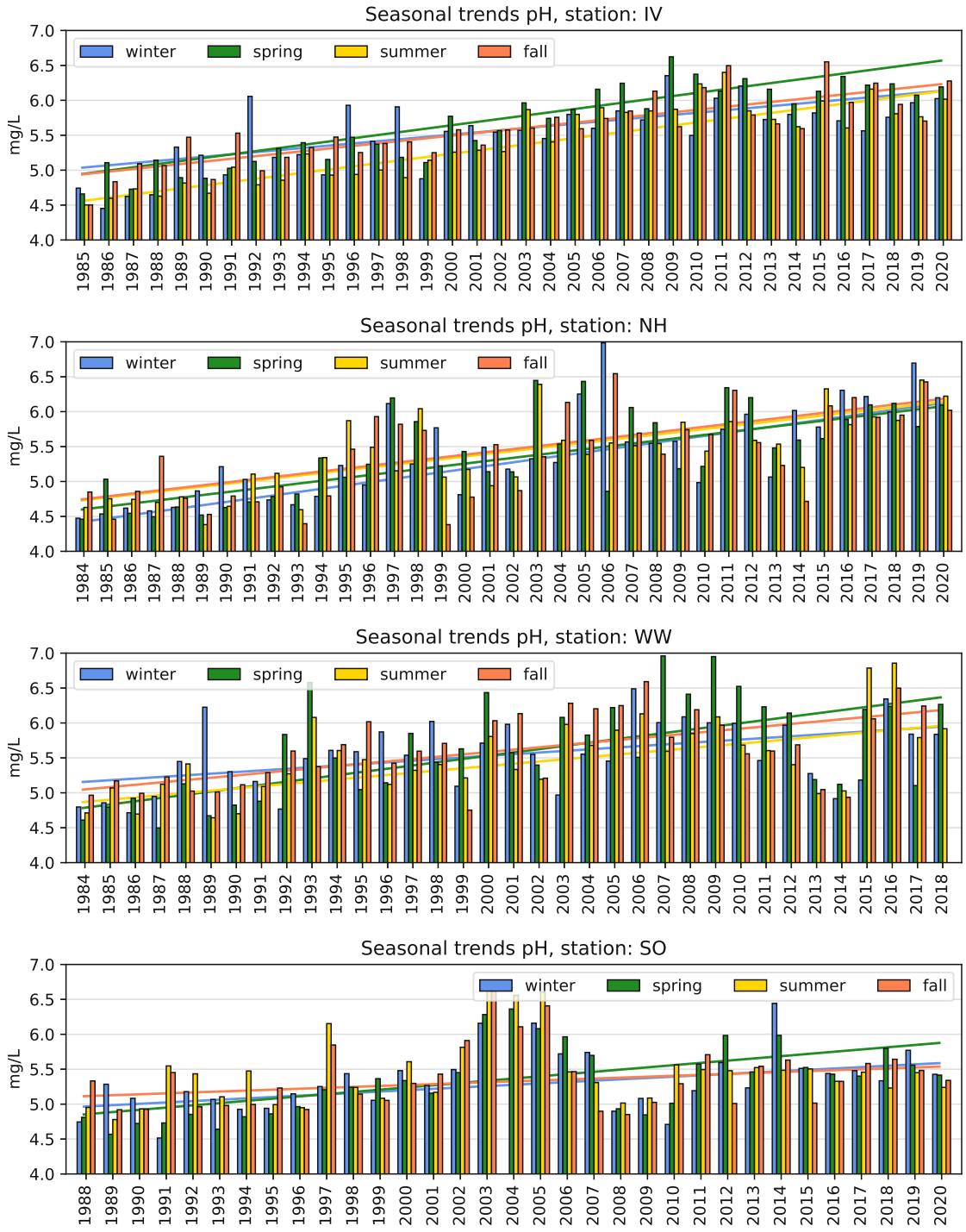


(d) Fig. D.9 (cont.): Grundlsee, Arnfels

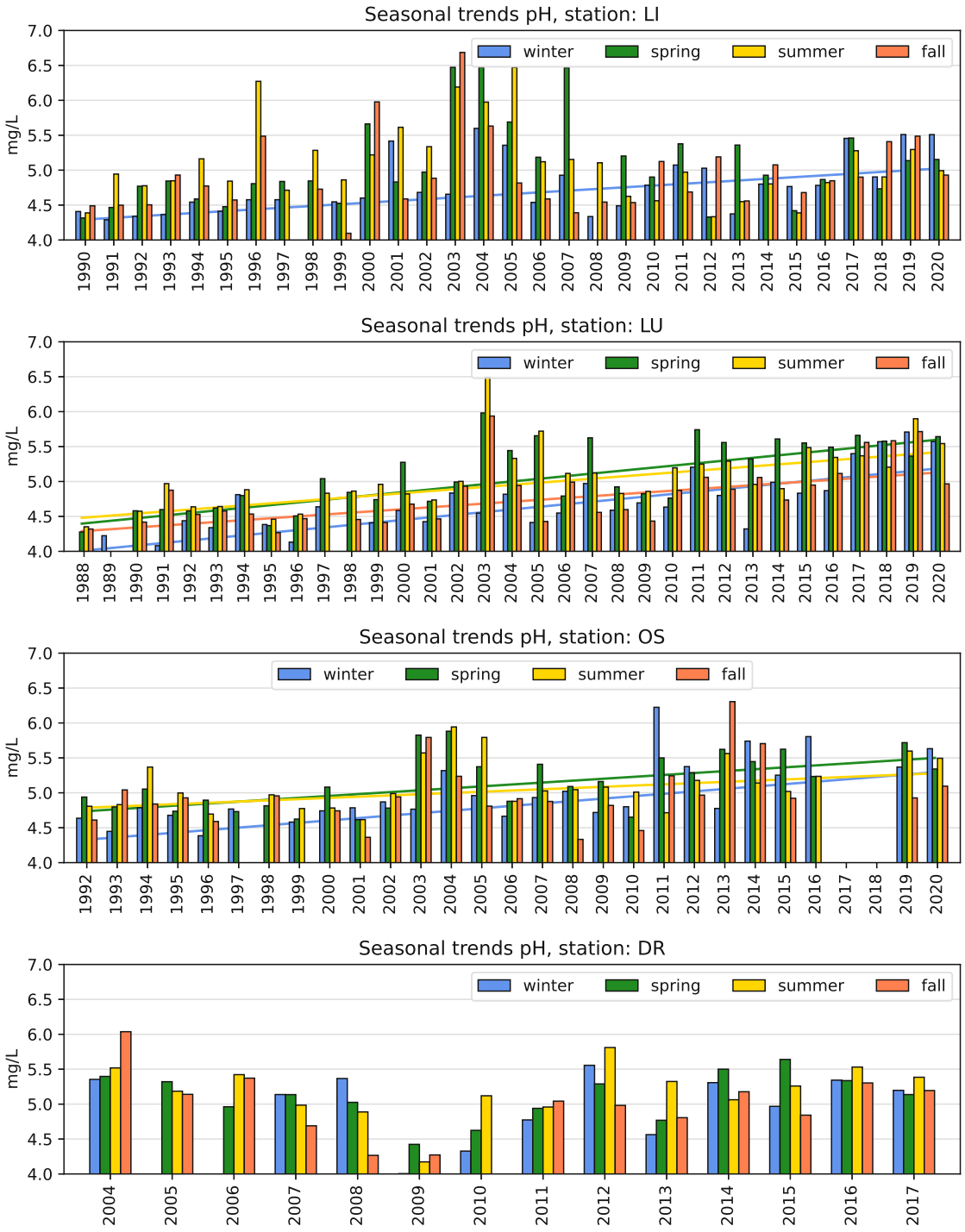


(a) Höfen, Niederndorferberg

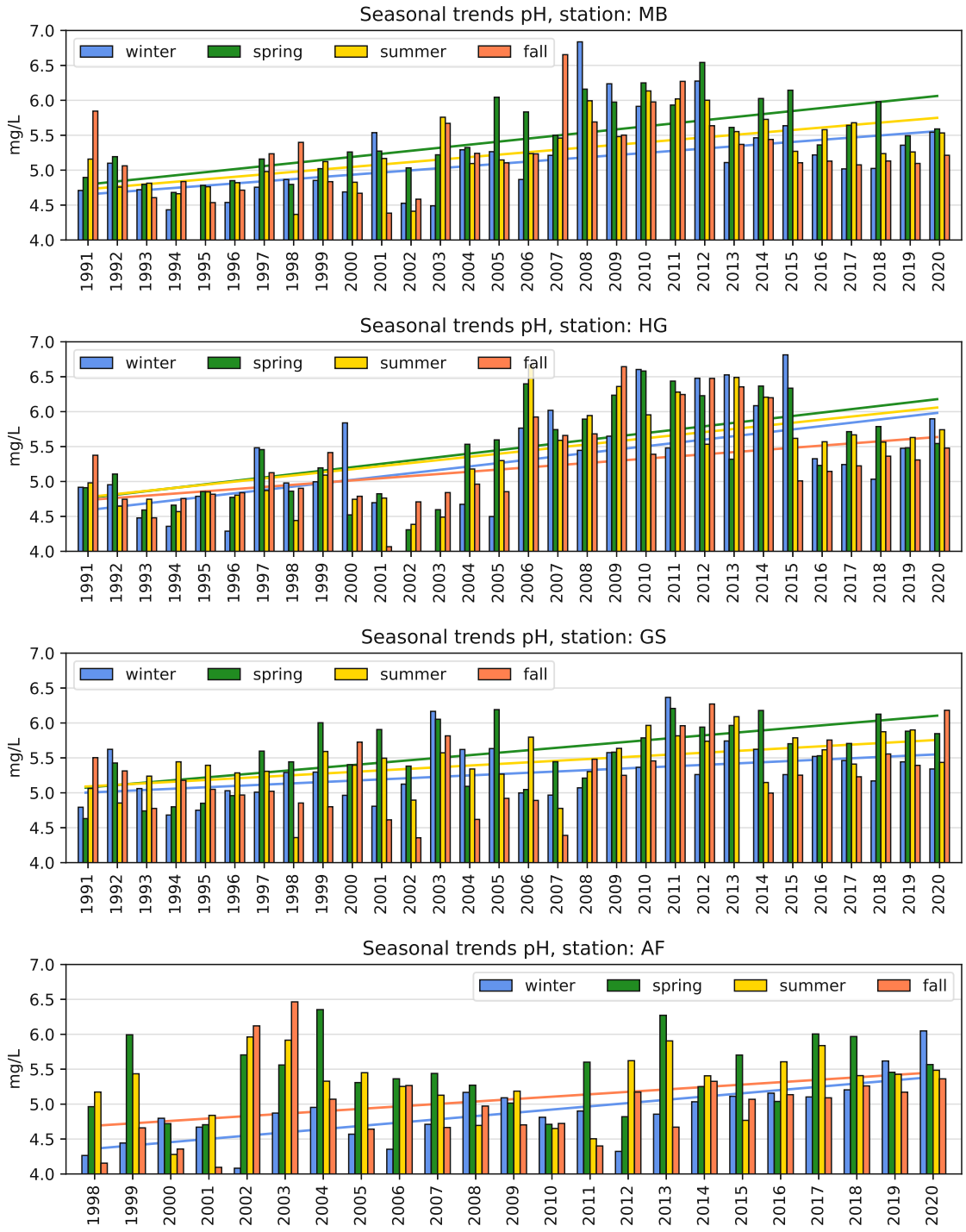
Figure D.10: Separated seasonal pH value trends



(b) Fig. D.10 (cont.): Innervillgraten, Haunsberg, Werfenweng, Sonnblick



(c) Fig. D.10 (cont.): Litschau, Lunz, Ostrong, Drasenhofen



(d) Fig. D.10 (cont.): Masenberg, Hochgößnitz, Grundlsee, Arnfels

Statement of originality

I hereby confirm that I have written the accompanying thesis by myself, without contributions from any sources other than those cited in the text and acknowledgements. This applies also to all graphics, drawings, maps and images included in the thesis.

The work has not been presented in the same or a similar form in other examination procedures in Austria or abroad.

Vienna, February 28, 2022

Peter Redl