

Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze

Abschlussbericht

Andreas Wunsch, M.Sc.
Dr. Tanja Liesch
Karlsruhe, September 2020

Institut für Angewandte Geowissenschaften, Abteilung Hydrogeologie
KIT - Die Forschungsuniversität in der Helmholtz-Gemeinschaft

Karlsruher Institut für Technologie
Die Forschungsuniversität in der Helmholtz-Gemeinschaft

Institut für Angewandte Geowissenschaften
Hydrogeologie

Autoren:	Wunsch, Andreas Liesch, Tanja, Dr.
Auftraggeber:	Bundesanstalt für Geowissenschaften und Rohstoffe (BGR)
Auftragsnummer:	20-4500109468
Datum:	September 2020

Inhaltsverzeichnis

Abbildungsverzeichnis		IV
Tabellenverzeichnis		VI
Abkürzungs- und Symbolverzeichnis		VII
1 Einleitung		1
2 Identifizierung von Referenzmessstellen		3
2.1 Zielsetzung		3
2.2 Datengrundlage und Preprocessing		5
2.2.1 Grundwasserdaten		5
2.2.2 Preprocessing		5
2.2.3 Untersuchungsgebiete		8
2.3 Methoden		10
2.3.1 Zeitreihen-Features		10
2.3.2 Self-Organizing-Maps		15
2.3.3 DS2L Clusteralgorithmus		16
2.3.4 Ensemblemodellierung		18
2.3.5 Cluster-Validierungs-Indizes		20
2.3.6 Referenzmessstellenauswahl		21
2.3.7 Methodensynthese: Workflow		23
2.4 Ergebnisse		25
2.4.1 Übersicht		25
2.4.2 Großraum 3: Oberrheingraben mit Mainzer Becken und nordhessischem Tertiär		27
2.4.3 Großraum 1: Nord- und mitteldeutsches Lockergesteinsgebiet		31
2.4.4 Großraum 2: Rheinisch-Westfälisches Tiefland		33
2.4.5 Großraum 4: Alpenvorland		34
2.4.6 Großraum 5: Mitteldeutsches Bruchschollenland		35
2.4.7 Großraum 6: West- und süddeutsches Schichtstufen- und Bruchschollenland		37
2.4.8 Großraum 7: Alpen		38
2.4.9 Großraum 8: West- und mitteldeutsches Grundgebirge		39
2.4.10 Großraum 9: Südostdeutsches Grundgebirge		40
2.4.11 Großraum 10: Südwestdeutsches Grundgebirge		41
2.5 Zusammenfassung und Diskussion der Referenzmessstellenauswahl		42
3 Grundwasserstandsvorhersage an Referenzmessstellen		46
3.1 Zielsetzung		46

3.2	Datengrundlage	47
3.2.1	Grundwasserdaten	47
3.2.2	Klimadaten	48
3.3	Methoden	50
3.3.1	Inputparameter	50
3.3.2	KNN-Modell	50
3.3.3	Fehlerkriterien zur Beurteilung der Vorhersagegüte	53
3.3.4	Vorhersageevaluierung	55
3.4	Ergebnisse und Diskussion	56
3.4.1	Wochenvorhersagen	57
3.4.2	Monatsvorhersagen	66
3.4.3	Jahreszeitenvorhersagen	72
3.4.4	Diskussion der Modellanpassungen und mögliche Fehlerquellen	78
3.4.5	Operationelle Vorhersagegüte	80
3.5	Zusammenfassung der Methodenentwicklung zur Grundwasserstandsvorsage	86
4	Vergleich mit numerischen Grundwasserströmungsmodellen	89
4.1	Zielsetzung	89
4.2	Material und Methoden	89
4.2.1	Künstliche Neuronale Netze	89
4.2.2	Grundwassermodelle des Oberrheingrabens Süd und Nord	90
4.2.3	Grundwassermodell Lenzen an der Elbe	93
4.2.4	Vergleichsstrategie	93
4.3	Ergebnisse und Diskussion	94
4.3.1	Vergleich mit den Modellen Oberrheingraben Süd und Nord	94
4.3.2	Vergleich mit Grundwassermodell Lenzen an der Elbe	100
4.4	Zusammenfassung Modellvergleich	100
5	Regionalisierung des mittleren Grundwasserflurabstands	103
5.1	Zielsetzung	103
5.2	Material und Methoden	103
5.2.1	Datengrundlage	103
5.2.2	Regionalisierung	104
5.3	Ergebnisse und Diskussion	105
5.4	Zusammenfassung der Regionalisierung des mittleren Grundwasserflurabstands	114
6	Zusammenfassung und Ausblick	115
	Literatur	118
	Anhang	125

Abbildungsverzeichnis

2.1	Originaldatensatz, Lage der Messstellen	6
2.2	Preprocessing	7
2.3	Ausreißerbereinigung	8
2.4	Hydrogeologische Raumgliederung	10
2.5	DS2L-Algorithmus	17
2.6	Ensemblemodellierung	18
2.7	Bipartite Graphen	20
2.8	Workflow	24
2.9	Lage von RM und verbleibenden Messstellen	26
2.10	Potentielle Eignung der Teilräume für eine spätere Vorhersage	45
3.1	GW-Zeitreihe mit Inhomogenität	47
3.2	GW-Interpolation	48
3.3	Güte der EPISODES-Klimavorhersage	49
3.4	NARX Architekturen	51
3.5	HP-Optimierung Datenteilung	53
3.6	Zusammenhang Inputkombinationen und Flurabstand	57
3.7	Gesamtüberblick Vorhersagegüte Wochenvorhersagen	57
3.8	Vergleich open-loop/closed-loop Wochenvorhersagen	59
3.9	Gesamtüberblick Vorhersagegüte Wochenvorhersagen je Monat	60
3.10	Korrelationsmatrix Fehlerkriterien Wochenvorhersagen	61
3.11	Übersichtskarte Wochenvorhersagen PI1 alle RM	62
3.12	Übersichtskarte Wochenvorhersagen PI1 alle Mst.	63
3.13	PI1 nach Startmonat gesplittet	64
3.14	Übersichtskarte Wochenvorhersagen PI1 Juni	65
3.15	Gesamtüberblick Vorhersagegüte Monatsvorhersagen	66
3.16	Korrelationsmatrix Fehlerwerte Monatsvorhersagen	67
3.17	Gesamtüberblick Vorhersagegüte Monatsvorhersagen je Startmonat	68
3.18	PI4op nach Startmonat gesplittet	69
3.19	Übersichtskarte Monatsvorhersagen PI4op alle RM	70
3.20	Übersichtskarte Monatsvorhersagen PI4op alle Mst.	71
3.21	Gesamtüberblick Vorhersagegüte Jahreszeitenvorhersagen	72
3.22	Korrelationsmatrix Fehlerkriterien Jahreszeitenvorhersagen	73
3.23	Gesamtüberblick Vorhersagegüte Jahreszeitenvorhersagen je Startmonat	74
3.24	PI12op nach Startmonat gesplittet	75
3.25	Übersichtskarte Jahreszeitenvorhersagen PI12op alle RM	76
3.26	Übersichtskarte Jahreszeitenvorhersagen PI12op alle Mst.	77
3.27	Übersicht Güte PI1, PI4op,PI12op nach Grundwasserleiter	79
3.28	Gesamtüberblick Güte EPISODES-basierter Jahreszeitenvorhersagen	80
3.29	Güte-Vergleich Jahreszeitenvorhersagen	81
3.30	Übersichtskarte Jahreszeitenvorhersagen PI12op - EPISODES	82

3.31	EPISODES Vorhersagen Instabilitäten	83
3.32	Güte-Vergleich Jahreszeitemvorhersagen 2	84
3.33	Saisonalität operat. Jahreszeitemvorhersagen	84
3.34	Vorhersagekonfidenz für 2007 an BB_34426014	85
3.35	Zeitliche Konfidenzbereichsanteile BB_34426014	87
4.1	Modellgebiete und Vergleichsmessstellen Oberrheingraben Süd und Nord .	92
4.2	Modellgebiet Lenzen an der Elbe	93
4.3	Überblick Fehler Oberrheingraben	95
4.4	Vergleich Vorhersagegüte an Messstelle BW_104-114-5	97
4.5	Vergleich Vorhersagegüte an Messstelle BW_710-256-3	98
4.6	Vergleich Vorhersagegüte an Messstelle BW_138-019-0	99
4.7	Vergleich Vorhersagegüte an Messstelle BB_29341721	101
5.1	Datengrundlage	104
5.2	Scatterplot GR1	107
5.3	Scatterplot GR2	107
5.4	Scatterplot GR3 ohne R33	107
5.5	Co-Kriging GW-Oberfläche GR1	108
5.6	Vergleich mit GW-Isolinien Brandenburg	109
5.7	Vergleich mit GW-Isolinien Mecklenburg-Vorpommern	111
5.8	Co-Kriging GW-Oberfläche GR2	112
5.9	Co-Kriging GW-Oberfläche GR3 ohne R33	113

Tabellenverzeichnis

2.1	Bundeslandkürzel	9
2.2	Anzahl der Messstellen in Großräumen	11
2.3	Feature-Übersicht	11
2.4	Cluster-Validierungs-Indizes	21
2.5	Übersicht der genutzten Features	27
2.6	Übersicht Messstellen- und Clusteranzahl	42
2.7	Messstellen- und Clusteranzahl nach Gebiet	43
3.1	Evaluierungsschemata	56
3.2	Inputparameter	56
4.1	Vergleichsmessstellen Oberrheingraben Süd und Nord	91
5.1	Kreuzvalidierungsfehler Co-Kriging Gebiete	106

Abkürzungs- und Symbolverzeichnis

ANN	Artificial Neural Network(s)
AM	Aussagemonat (Monat für den eine Vorhersage berechnet wird); alt.: Target Month
BB	Brandenburg
BE	Berlin
BW	Baden-Württemberg
BY	Bayern
Bias	systematische Abweichung
C	C-Index
CH	Calinski-Harabasz-Index
d_{norm}	normierter Abstand zum Clusterzentrum
DM	Density Merging (DS2L-Parameter)
DR	Density Refinement (DS2L-Parameter)
DS2L	Density-based simultaneous two-level algorithm
DWD	Deutscher Wetterdienst
FD	Feedback-Delay
FFN	Feedforward Netz
FR	Frankreich
GR	hydrogeol. Großraum (HYRAUM)
GWL	Grundwasserstand (Groundwater Level)
HB	Hansestadt Bremen
HE	Hessen
HH	Hansestadt Hamburg
HP	Hyperparameter
ID	Input-Delay
Isomap	Isometric Feature Mapping
KNN	Künstliche(s) Neuronale(s) Netz(e)
MR	McClain-Rao-Index
MStD	Messstellendichte in Messstellen pro 100km ²
MV	Mecklenburg-Vorpommern
NARX	Nonlinear Autoregressive Network with Exogenous Input
NI	Niedersachsen
NSE	Nash-Sutcliffe Efficiency
NTH	Neighborhood-Threshold (DS2L-Parameter)
NW	Nordrhein-Westfalen
P	Niederschlag
PBM	PBM-Index (nach: Pakhira, Bandyopadhyay und Maulik)
PCA	Principal-Component-Analysis, Hauptkomponentenanalyse

PI	Persistency Index
PI _{op}	operationeller PI
R	hydrogeol. Raum (HYRAUM)
R ²	Bestimmtheitsmaß
rBias	relative systematische Abweichung
rH	relative Luftfeuchte
RL	Ratkowsky-Lance-Index
RM	Referenzmessstelle
RMSE	Root Mean Squared Error
RNN	Rekurrentes Netz
RP	Rheinland-Pfalz
rRMSE	relativer RMSE
$\overline{R_w}$	mittlerer gewichteter Korrelationskoeffizient
SH	Schleswig-Holstein
SL	Saarland
SM	Startmonat (einer Vorhersage)
SN	Sachsen
SOM	Self Organizing Map
ST	Sachsen-Anhalt
σ	Standardabweichung
T	Temperatur
TH	Thüringen
TR	hydrogeol. Teilraum (HYRAUM)
Tsin	an Temperatur gefittetes Sinusignal
WMF	Wiener-Mehrkanalfilter

1 Einleitung

Das Projekt *Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze* (KNN) hat übergeordnet zum Ziel herkömmliche physikalisch-numerische Grundwassermodelle zur überregionalen Vorhersage von Grundwasserständen durch Techniken des maschinellen Lernens zu ersetzen, um so die Abhängigkeit von fehlenden oder unvollständigen Felddaten (z.B. Geometrien und physikalische Eigenschaften des Untergrundes) aufzulösen. Dies ist möglich, da hierbei die Fragestellung auf eine Input-Output Beziehung reduziert wird, und die zuvor aufwändige und teure Datenerhebung zur Charakterisierung des Untergrundes entfällt. Konkretes Ziel des Projektes ist die wiederkehrende Berechnung und vor allem Bereitstellung von Grundwasserstandsvorhersagen von bis zu drei Monaten in die Zukunft an über das gesamte Gebiet der Bundesrepublik Deutschland verteilten Messstellen. Dadurch soll für Öffentlichkeit und Entscheidungsträger die Ableitung von Aussagen bezüglich der Wasserverfügbarkeit für Trinkwasserversorgung und für landwirtschaftliche sowie industrielle Nutzung ermöglicht werden. Weitere Nutzungs- und Anwendungsszenarien der Vorhersagen sind denkbar, wie beispielsweise die Entwicklung von Vermeidungs- und Anpassungsstrategien für die Folgen von Grundwasserextremständen auf Ökosysteme, Infrastruktur und Nutzflächen.

Vorliegender Endbericht fasst die während der dreijährigen Projektlaufzeit erzielten Ergebnisse zusammen. Die inhaltliche Struktur orientiert sich dabei an den vorher definierten Arbeitspaketen, die jeweils in eigenständigen Kapiteln beschrieben und zusammengefasst werden. Kapitel 2 beschäftigt sich mit der Auswahl repräsentativer Referenzmessstellen aus dem zur Verfügung stehenden Gesamtdatensatz. Die Referenzmessstellen sollten dabei jeweils eine für eine größere Gruppe von Messstellen repräsentative Dynamik aufweisen. Dieser Schritt ist nötig, da eine Vorhersage an allen Messstellen für einen operativen Betrieb zu aufwändig wäre. Innerhalb einer Gruppe mit ähnlicher Dynamik (Cluster) können später Vorhersagen mit nur geringem Genauigkeitsverlust von der Referenzmessstelle auf die restlichen Messstellen des Clusters übertragen werden. Gleichzeitig erlaubt der Schritt der Auswahl von Referenzmessstellen auch ein besseres Verständnis der Grundwasserdynamik an den Messstellen, z.B. hinsichtlich möglicher Einflussfaktoren, deren Einbeziehung in die spätere Vorhersage das Ergebnis dieser verbessern könnte. Darüber hinaus können so automatisiert Messstellen aussortiert werden, an denen eine Vorhersage nicht möglich erscheint, z.B. aufgrund mangelnder Datenqualität oder häufiger unvorhersehbarer anthropogener Einflüsse wie Absenkungen für Baumaßnahmen.

Kapitel 3 beschreibt die Entwicklung, Anwendung und Validierung der eigentlichen Vorhersagealgorithmen. Ziel dieses Arbeitsschrittes war die Entwicklung einer übertragbaren, auf künstlichen neuronalen Netzen basierende Methodik zur Berechnung von Wochen-, Monats-, und Jahreszeitvorhersagen (3 Monate) an den vorher ausgewiesenen Referenzmessstellen auf Basis von beobachteten und modellierten Klimarasterdaten des Deutschen Wetterdienstes (DWD). In Kapitel 4 werden die Ergebnisse eines Ver-

gleiches der entwickelten Vorhersagemethodik mit den Daten existierender numerischer Grundwasserströmungsmodelle vorgestellt. Hierfür wurden drei numerische Modelle ausgewählt und die Zeitreihen von insgesamt 16 Messstellen vergleichend betrachtet. Die Auswahl umfasst sowohl große überregionale, als auch kleinräumig spezialisierte Modelle. In einem abschließenden Schritt (Kapitel 5) wurden auf Basis der in Kapitel 2 gesammelten und geclusterten Daten Karten des mittleren Grundwasserflurabstandes sowie der absoluten Grundwasseroberfläche in mNN für ganz Deutschland erstellt. Dabei wurde auf Co-Kriging als bewährtes Interpolationsverfahren zur Regionalisierung und auf die hydrogeologische Raumgliederung von Deutschland (BGR und SGD, 2015) zur Abgrenzung räumlicher Einheiten zurückgegriffen.

Kapitel 6 fasst die zentralen Ergebnisse des Gesamtprojektes zusammen und gibt einen Ausblick auf mögliche nachfolgende Arbeitsschritte.

2 Identifizierung von Referenzmessstellen

2.1 Zielsetzung

Die Entwicklung und Anwendung einer Methodik zur Identifizierung von Referenzmessstellen (RM) ist der erste und ein gleichzeitig essentieller Arbeitsschritt für eine repräsentative Vorhersage von Grundwasserständen, da sich hiermit der Arbeitsaufwand einer operationellen Vorhersage erheblich reduzieren lässt. Grundidee ist die stellvertretende Vorhersage einer repräsentativen Messstelle für eine ganze Gruppe, um so Aussagen für alle Messstellen der Gruppe treffen zu können. Dabei sollten die jeweiligen Referenzmessstellen weitgehend anthropogen unbeeinflusst sein, um die natürlich Grundwasserdynamik abzubilden sowie je nach Situation als maßgebend für ein zu definierendes Gebiet (z.B. Grundwasserkörper/hydrogeologischer Raum oder Teilraum) gelten. Als Datengrundlage dient ein mit freundlicher Unterstützung der geologischen Dienste der Länder neu zusammengestellter Datensatz mit Grundwasserstandszeitreihen von ca. 13.500 Messstellen und deren Stammdaten. Durch im Projektverlauf gewonnene Erkenntnisse waren Modifikationen der ursprünglich formulierten Kriterien für Referenzmessstellen notwendig. Diese werden im weiteren Verlauf detaillierter erläutert.

Das Ziel, anthropogen unbeeinflusste Messstellen auszuwählen, um die natürliche Grundwasserdynamik zu betrachten, erscheint auf den ersten Blick sinnvoll und notwendig. Dadurch kann mutmaßlich eine höhere Genauigkeit der späteren Vorhersagen erzielt werden, da diese auf natürlichen Inputparametern wie beispielsweise dem Niederschlag und der Temperatur beruhen und anthropogene Einflüsse bei der Vorhersage nicht oder nur schwer berücksichtigt werden können. Weiterhin soll dadurch gewährleistet werden, dass eine Messstelle die vorherrschende Dynamik eines größeren Gebietes wiedergibt und nicht durch lokale anthropogene Einflussfaktoren wie Entnahmen oder Infiltration dominiert wird. Da in Deutschland über weite Teile eine intensive Grundwasserbewirtschaftung betrieben wird, stellt sich hierbei jedoch die Frage, ob und wo es überhaupt gänzlich unbeeinflusste Gebiete gibt. Häufig ist zumindest eine regionale anthropogene Beeinflussung durch eine Vielzahl von Faktoren nicht auszuschließen. Nachvollziehbarerweise sind die Vorhersagen, neben beispielsweise Gebieten mit geringerer Wasserverfügbarkeit, später auch und vor allem für jene Gebiete interessant, in denen Grundwasserbewirtschaftung und -nutzung stattfindet. Hierdurch stehen die formulierten Anforderungen bereits teilweise kontradiktorisch zur Zielsetzung des Projekts. Weiterhin gibt es auch lokal begrenzte natürliche Einflussfaktoren (v.a. In-/Exfiltration an Vorflutern), die sich deutlich in der Dynamik einer Messstelle wiederfinden lassen. Demgegenüber sind anthropogene Einflüsse wie Entnahmen häufig eng an natürliche Einflüsse gekoppelt, wie beispielsweise erhöhte Entnahmeraten für die Bewässerung bei höheren Temperaturen und geringen Niederschlägen.

Die Dynamik einer Messstelle ist also eine komplexe Überlagerung aller Umgebungseinflüsse zu unterschiedlichen Anteilen, welche nur schwer oder gar nicht quantifizier- und

differenzierbar sind. Dies führt dazu, dass man an nah beieinander liegenden Messstellen mit identischer oder zumindest stark ähnlicher Geologie von Zeit zu Zeit deutliche Unterschiede in der relativen und absoluten Grundwasserdynamik beobachten kann. Umgekehrt ist es nicht selten möglich, dass weiter entfernte Messstellen eine ähnliche Dynamik aufweisen, wie man sie in diesem Maß zunächst nicht erwarten würde.

Beruhend auf der oben beschriebenen Vorgabe, dass Referenzmessstellen anthropogen unbeeinflusst sein sollten, existieren verschiedene Ansätze zur Differenzierung zwischen anthropogen beeinflussten Messstellen und solchen, die nur von natürlichen klimatischen Einflussfaktoren abhängen. Lischeid et al. (2012) nutzten dazu eine Hauptkomponentenanalyse (PCA), Böttcher et al. (2014) verwendeten mit Isomap einen mit der PCA eng verwandten Ansatz. Beide Verfahren haben den für die vorliegende Anwendung entscheidenden Nachteil, dass Daten lückenlos und über den gleichen Zeitraum vorhanden sein müssen. Dadurch wären nur ein Bruchteil der gesammelten Daten überhaupt verwendbar und ein hoher Informationsverlust müsste in Kauf genommen werden. Eine weitere Möglichkeit, den anthropogenen Einfluss an Grundwassermessstellen zu erfassen, bietet der Wiener-Mehrkanalfilter (WMF). Hierbei werden jedoch Vorkenntnisse über unbeeinflusste Prüfmessstellen und ein unbeeinflusster Kalibrierzeitraum benötigt (Bucher, 1999; Grossmann und Skowronek, 2005). Dies ist bei einer bundesweiten Auswahl von Referenzmessstellen vor allem ohne entsprechende Vorkenntnisse nicht praktikabel. Das Verfahren der klimatischen Ganglinie von Wriedt (2017) betrachtet im Gegensatz zu den bisher angesprochenen Verfahren den kausalen Zusammenhang von klimatischen Bedingungen und Grundwasserstand. Durch Translation und Tiefpassfilterung der Zeitreihe der klimatischen Wasserbilanz wird diese bestmöglich an die Grundwasserstandszeitreihe angepasst. Nachträglich erfolgt auf Basis der Residuen zwischen beiden Zeitreihen eine subjektive Einteilung in die Gruppen unbeeinflusst, beeinflusst bzw. unklar. Schwachpunkt ist bei diesem Verfahren aktuell die noch offene Anwendbarkeit vor allem auf Karstgebiete und sowohl die Datenverfügbarkeit zur Berechnung als auch die Berechnungsmethodik der klimatischen Wasserbilanz selbst. Weiterhin erscheint ein Verfahren, welches eine subjektive Einschätzung einzelner Ganglinien von einem Bearbeiter erfordert, für ein so großes Gebiet mit einer Vielzahl an Messstellen unpraktikabel.

Einen anderen Ansatz zur Auswahl repräsentativer Messstellen mit stärkerem Fokus auf der Zeitreihendynamik anstelle von anthropogener Beeinflussung wählte Wingerling (1999). Hierbei wurde eine hierarchische Clusteranalyse auf Grundwassermessstellen in Baden-Württemberg angewandt, um Messstellen mit gleicher Dynamik zu gruppieren. Auch dabei existieren strenge Anforderungen an den Ausgangsdatensatz hinsichtlich Datenlücken und Beobachtungszeitraum. Aus diesem Grund wurde von Wingerling (1999) der Datensatz auf Monatsmittelwerte aggregiert und auf zehn Jahre eingeschränkt. Auf die Problematik der anthropogenen Beeinflussung wird bspw. durch ein Aussortieren der Messstellen in der Nähe von Grundwasserentnahmen eingegangen.

Der in diesem Projekt gewählte Ansatz ähnelt am ehesten dem letztgenannten von Wingerling (1999). Der Fokus liegt klar auf dem Vergleich relativer Grundwasserdynamiken und der Gruppierung solcher mit großer Ähnlichkeit. Hierbei sind externe Einflussfaktoren, egal ob anthropogen oder natürlich, erst einmal irrelevant und fließen nur konzeptionell in das Verfahren ein. Ziel ist die Bildung von Gruppen mit homogener Dynamik. Externe Einflussfaktoren sollen sich im Idealfall in der Gruppenbildung widerspiegeln. Nach erfolgreichem Clustern der Grundwasserganglinien kann eine Beurteilung der vor-

herrschenden Dynamik durch den Bearbeiter erfolgen, indem beispielsweise die räumliche Lage in Relation zu bekannten Einflüssen (z.B. Vorfluternähe) betrachtet wird. Die eigentliche RM-Auswahl erfolgt dann für jedes (geeignete) Cluster separat. Zusätzliche Anforderung an das genutzte Verfahren ist es, möglichst wenig Voraussetzungen an den Ausgangsdatensatz zu stellen, sodass Zeitreihen unterschiedlicher Länge, mit und ohne Datenlücken und aus unterschiedlichen Beobachtungszeiträumen verarbeitet werden können. Weiterhin soll der Informationsverlust möglichst gering gehalten und so viele der verfügbaren Ganglinien wie möglich genutzt werden. Um diese Ziele zu erreichen, wurden verschiedene Methodiken kombiniert. Kernstück bildet die Anwendung von selbstorganisierenden Karten (Kohonen, 1987), sog. Self-Organizing-Maps (SOM), zum Clustern der Grundwasserstandszeitreihen. Hierbei handelt es sich um einem speziellen Typ von Künstlichen Neuronalen Netzen (KNN) und somit wie für die Vorhersage in Kapitel 3 ebenfalls um einen datengetriebenen Modellansatz aus dem Bereich des maschinellen Lernens.

2.2 Datengrundlage und Preprocessing

2.2.1 Grundwasserdaten

Der verwendete Datensatz wurde mit freundlicher Unterstützung der geologischen Dienste der Länder zusammengetragen und besteht in seiner Rohform aus Grundwasserstandsdaten von 13.492 Messstellen variabler Länge und Datenqualität. Abb. 2.1 zeigt die Lage aller verfügbaren Messstellen. Die Stammdaten umfassen Name, ID, Koordinaten und Angaben zum Grundwasserleiter. Letztere Angabe war jedoch nicht flächendeckend bzw. mit gewünschter Genauigkeit verfügbar, und auch der Umfang der weiteren übermittelten Stammdaten variiert erheblich je Bundesland.

Bei der Zusammenstellung der Datensätze der Länder wurde ursprünglich angestrebt, einheitliche Auswahlkriterien anzulegen. Aus Gründen der Datenqualität und Datenhaltung in den Bundesländern war das nicht immer möglich. Generell wurden bevorzugt Wochendaten von aktiven, in flachen Grundwasserleitern verfilterten Messstellen angefordert, die eine Mindestlänge von 4 Jahren und bestenfalls eine Länge von 30 Jahren aufweisen. Teilweise wurden deutlich darüber hinausgehend Daten von bis zu ca. 100 Jahren einzelner Messstellen geliefert, viele Messstellen weisen auch kürzere Datenerlängen als 30 Jahre auf. Es wurden noch weitere Kriterien formuliert, jedoch waren die Abweichungen von diesen und teils auch den bereits genannten Kriterien oft so vielfältig, dass man nicht von einem einheitlichen Datensatz sprechen kann. Die Überprüfung und Modifikation aller notwendigen Parameter wurde in den Arbeitsablauf eingebettet, um so möglichst einheitliche Voraussetzungen für die Bearbeitung der Untersuchungsgebiete zu schaffen.

2.2.2 Preprocessing

Die Zeitreihen wurden einem Preprocessing unterzogen, um eine Datenbank aus wöchentlichen Grundwasserstandsdaten, einheitlichem Koordinatensystem (ETRS 1989 UTM 32N), Datenformat (.txt) und Name (s.u.) zu erstellen und die generelle Datenqualität

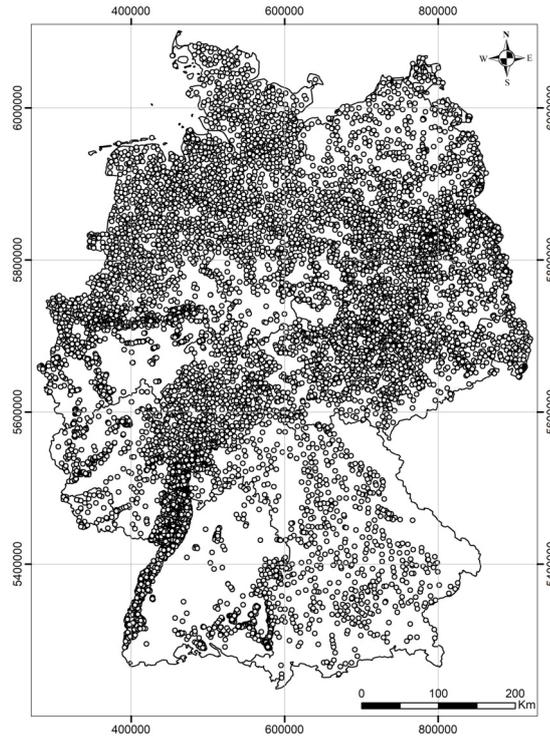


Abbildung 2.1: Lage der Messstellen des unbearbeiteten Originaldatensatzes, Koordinaten in ETRS 1989 UTM 32N

zu erhöhen. Der Ablauf dieses Preprocessings ist in Abbildung 2.2 dargestellt und wird im Folgenden genauer erläutert.

Zunächst wird die Originalzeitreihe auf Ausreißer untersucht. Dazu wird ein gleitender Mittelwert auf die Zeitreihe angewandt, und alle Werte, die einen Grenzwert von 3σ (Standardabweichung) über- oder unterschreiten, werden entfernt. Die Größe des gleitenden Fensters ist hierbei entscheidend und wurde auf Basis von Vorversuchen auf elf Werte festgelegt. Demnach werden jeweils fünf Werte vor und nach dem aktuell betrachteten Messwert einbezogen. Diese Konfiguration ist sehr konservativ gewählt, damit nicht versehentlich reale Messwerte von starken Grundwasseranstiegen, wie sie beispielsweise in Karst-Messstellen vorkommen können, entfernt werden. Der 3σ -Grenzwert ist aus der Standardnormalverteilung abgeleitet, bei der innerhalb des 3σ -Intervalls 99,73% aller Messwerte zu finden sind (Abb. 2.3a). Für die Bereinigung der Originaldaten werden benachbarte Werte unabhängig von ihrem tatsächlichen zeitlichen Abstand betrachtet. Dies führt zu einer Abhängigkeit der Grenzwerte vom Messintervall, sodass die Grenzwerte umso dichter beieinander liegen, je höher das Messintervall ist. Abb. 2.3b zeigt den Einfluss der Datendichte bzw. des Messintervalls in den Originaldaten auf den oberen ($+3\sigma$, hellblau) und unteren (-3σ , gelb) Grenzwert bei der Ausreißerbereinigung. Es ist eine deutlich erhöhte Ausreißersensitivität ab Mai 2012 zu erkennen (Abstand von gelber und blauer Linie wird kleiner), da ab diesem Zeitpunkt die Messfrequenz deutlich erhöht wurde. Im November wurden dann erfolgreich vier Ausreißer identifiziert und entfernt.

Der nächste Schritt des Preprocessings führt eine Datenaggregation bzw. Disaggregation der Messdaten durch und erzeugt Messreihen mit konstant wöchentlichem Abstand

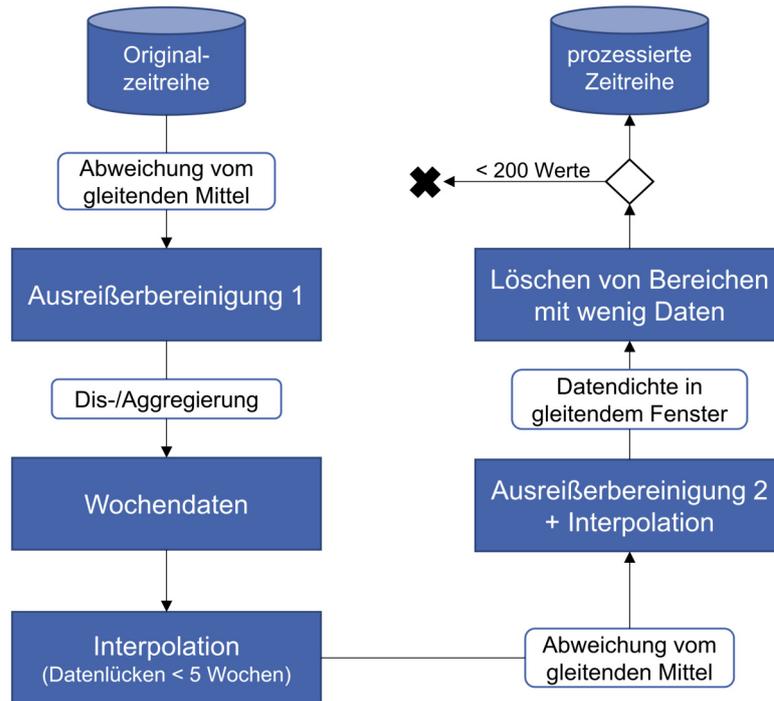


Abbildung 2.2: Workflow Preprocessing

der Messwerte. Im Falle von Disaggregation (downscaling) werden lediglich Fehlstellen erzeugt, im Falle einer Datenaggregation (upscaling) werden Montagswerte (traditionell Stichtag) ausgewählt. Von Montagen abweichende Messzeitpunkte werden dem nächstgelegenen Montag zugeordnet, falls dort nicht bereits ein Messwert vorhanden ist. Anschließend folgt eine lineare Interpolation kleinerer Messlücken bis zu einem Monat (≤ 4 Werte). Dies hält den Einfluss der Interpolation auf die Gesamtzeitreihe gering, ermöglicht jedoch einen Informationsgewinn für Abschnitte, in denen nur monatliche Messdaten vorliegen. Neben linearer Interpolation wurde auch Spline-Interpolation auf ihre Eignung hin untersucht. Diese liefert zwar häufig realistischere Ergebnisse, reagiert aber im Umfeld nicht erkannter Ausreißer überproportional stark und generiert zusätzliche Fehler. Splines können zudem einen glättenden Effekt auf die Zeitreihe haben, wodurch die Beschreibung der Zeitreihe durch Features (s. Kapitel 2.3) verfälscht werden kann. Lineare Interpolation hingegen hat nur einen minimalen Effekt auf diese Featureberechnung.

Im Anschluss ist eine zweite Ausreißerbereinigung im Workflow vorgesehen. Dies ist nötig, da durch die Skalierung auf Wochendaten sich die Werte des gleitenden Mittels eines Fensters und damit die Grenzwerte für Ausreißer ändern. Abb. 2.3 zeigt diese Änderung, wenn Originaldaten (2.3c) und Wochendaten mit identifizierten Fehlstellen (2.3d) der Ausreißerbereinigung unterzogen werden. Der umgekehrte Fall, also der Ausreißer wird erst nach Skalierung und Interpolation auf Wochendaten erkannt, ist ebenfalls möglich. Direkt einhergehend mit der zweiten Ausreißerbereinigung werden diese einzelnen entfernten Werte wieder linear interpoliert, um keine zusätzlichen Datenlücken zu erzeugen.

Im letzten Preprocessing-Schritt wird ein gleitendes Fenster mit einer Größe von zwei Jahren auf die Zeitreihe angewandt, um nach dem Preprocessing noch immer lückenhafte

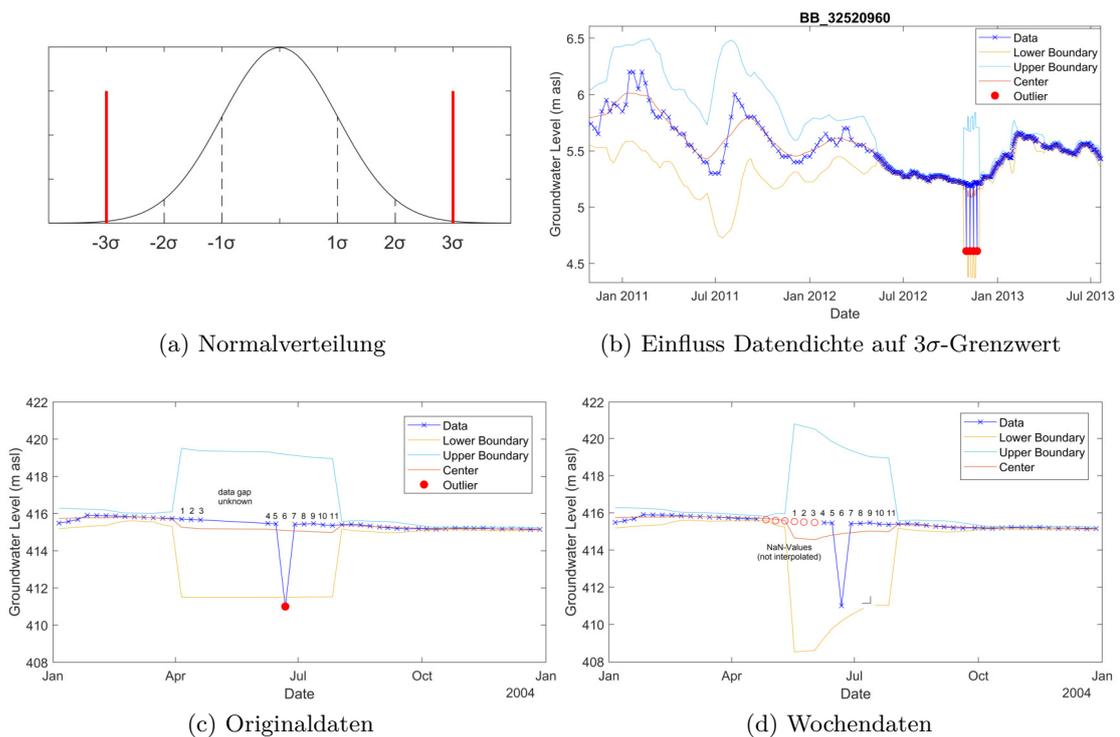


Abbildung 2.3: (a) Herleitung der Grenzwerte für die Ausreißerbereinigung gemäß Pre-processing Workflow (Abb. 2.2) sowie Sensitivität gegenüber Datendichte/Messintervall (b) und Fehlstellen (c,d)

Abschnitte auszusortieren. Bereiche, in denen der Anteil der Datenlücken mehr als 25% beträgt, werden entfernt. Dieser Schritt wurde so modifiziert, dass im Übergangsbereich zwischen lückenhaften und lückenlosen Bereichen nicht unnötig viele Daten gelöscht werden. Wenn die Anzahl der Messwerte nach all diesen Schritten weniger als 200 beträgt, wird die Zeitreihe verworfen.

Anschließend werden alle verbleibenden Zeitreihen nach folgendem Schema umbenannt: *Bundeslandkürzel_Bundesland-ID* (vgl. Messstelle in Abb. 2.3b). Die Kürzel der Bundesländer sind in Tabelle 2.1 aufgeführt, die Bundesland-ID entspricht dem im jeweiligen Bundesland verwendeten Format zur Messstellenbezeichnung.

2.2.3 Untersuchungsgebiete

Die Nutzung einer Vorgliederung nach (hydro-)geologischen Gesichtspunkten ist sinnvoll, da prinzipiell ein Kausalzusammenhang der Gangliniendynamiken innerhalb der zu untersuchenden Gebiete hergestellt werden soll. Auf diese Weise kann die angewandte, rein mathematisch basierte Methode mit hydrogeologischen Informationen ergänzt werden, und die Gangliniengruppierungen lassen in der Folge mehr Interpretation zu, als die der isoliert betrachteten mathematischen Ähnlichkeit. Als Grundlage für die räumliche Gliederung der untersuchten Gebiete zur Referenzmessstellenauswahl dient die hydrogeologische Raumlagerung von Deutschland (Ad-hoc-Arbeitsgruppe Hydrogeologie, 2016).

Tabelle 2.1: Bundeslandkürzel für die Benennung der prozessierten Messstellendaten

Bundesland	Kürzel
Brandenburg	BB
Berlin	BE
Baden-Württemberg	BW
Bayern	BY
Hansestadt Bremen	HB
Hessen	HE
Hansestadt Hamburg	HH
Mecklenburg-Vorpommern	MV
Niedersachsen	NI
Nordrhein-Westfalen	NW
Rheinland-Pfalz	RP
Schleswig-Holstein	SH
Saarland	SL
Sachsen	SN
Sachsen-Anhalt	ST
Thüringen	TH

Hierbei wurde das Gebiet der Bundesrepublik Deutschland unter Berücksichtigung von hydrogeologischen, geologischen, wasserchemischen und morphologischen Gesichtspunkten in hydrogeologische Großräume, Räume und Teilräume untergliedert (Abb. 2.4). Dies hat den großen Vorteil, dass die Ausweisung der Einheiten bundesweit nach einheitlichen Kriterien erfolgte und die Grenzziehung bundeslandübergreifend konsistent ist. Dies wäre beispielsweise bei der Verwendung von Grundwasserkörpern nicht der Fall. Dass auch bei der hydrogeologischen Raumgliederung mit unterschiedlicher Generalisierung gearbeitet wurde, ist deutlich erkennbar, wenn man die sehr feine Grenzziehungen in Baden-Württemberg mit denen anderer Bundesländer vergleicht. Dennoch bildet die hydrogeologische Raumgliederung eine gute Basis für dieses Projekt.

Als Ausgangspunkt für die Ausweisung von Referenzmessstellen dient aus verschiedenen Gründen die Einteilung in hydrogeologische Großräume. Die Messstellendichte richtet sich in Deutschland häufig nach der Relevanz eines Gebietes für die Grundwasserbewirtschaftung und ist dementsprechend unterschiedlich ausgeprägt (vgl. Abb. 2.1). Bei Betrachtung auf Großraumbene ist überwiegend gewährleistet, dass ausreichend viele Messstellen für die Anwendung eines datengetriebenen Modelles wie SOM vorhanden sind. Ausnahmen sind lediglich die Großräume 7 und 10 (vgl. Tabelle 2.2). Auf Raum- oder sogar Teilraumbene wären etliche leere Flecken unvermeidbar, da teilweise schlichtweg keine oder zu wenig Messstellen in den entsprechenden (Teil-) Räumen vorhanden sind. Weiterhin ist durch die Einteilung in Räume oder Teilräume keinesfalls das komplette Gebiet als hydrogeologisch homogen einzustufen, da es sich bei der Raumgliederung um eine zweidimensionale Generalisierung von komplexen dreidimensionalen Strukturen handelt. Es ist beispielsweise davon auszugehen, dass in der überwiegenden Anzahl der Räume, auch wenn diese nicht als Porenaquifere beschrieben werden, zumindest kleinräumige fluviatile Porenaquifere vorhanden sind. Messstellen können also auch grenzübergreifend ähnliche Dynamiken aufweisen. Von der Großraumgliederung wird bei der Modellierung nur in wenigen Fällen abgewichen und dies an entsprechender Stelle näher

Hydrogeologische Großräume und Räume in Deutschland

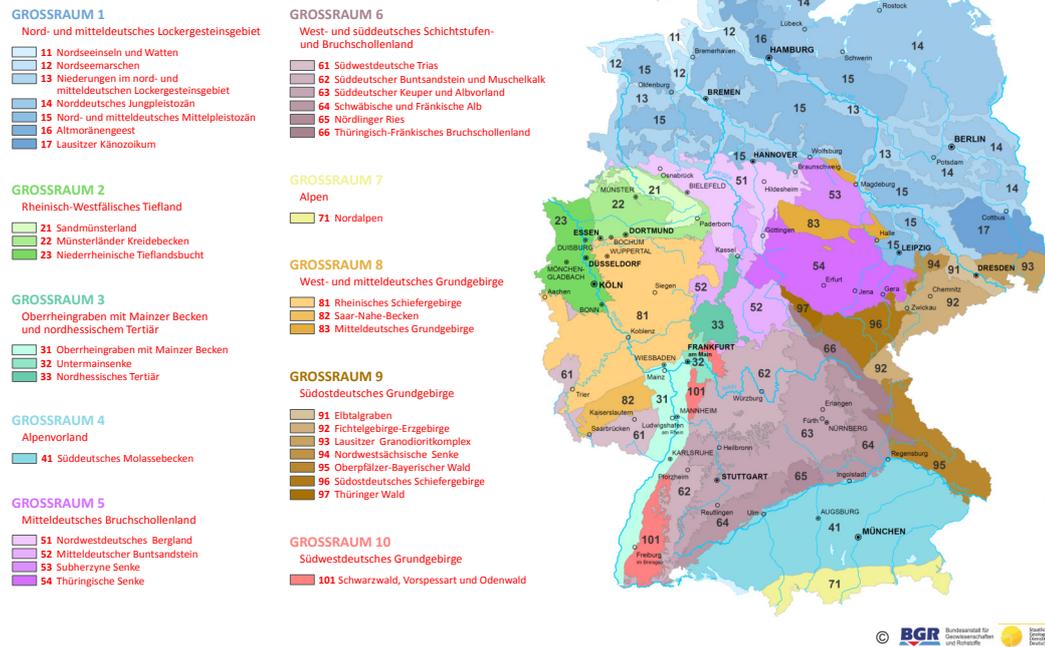


Abbildung 2.4: Hydrogeologische Raumgliederung von Deutschland aus Ad-hoc-Arbeitsgruppe Hydrogeologie (2016) und verfügbar unter: www.bgr.bund.de/hyraum

erläutert. Tabelle 2.2 fasst die Anzahl der Messstellen in den jeweiligen Großräumen vor und nach dem Preprocessing zusammen. An dieser Stelle sei angemerkt, dass die Anzahl der Messstellen nach dem Preprocessing von der später für das Clustering verwendeten Anzahl abweicht. Dies ergibt sich daraus, dass bei der Eingrenzung des Untersuchungszeitraums (siehe Kapitel 2.3.7) in der Regel noch weitere Messstellen wegfallen, die im gewählten Zeitraum keine oder zu wenige Daten aufweisen. Die Anzahl der jeweils für das Clustering verwendeten Messstellen findet sich in Kapitel 2.4.

2.3 Methoden

2.3.1 Zeitreihen-Features

Das Clustern von Zeitreihen ist nicht trivial. Zentrales Problem ist die Frage nach einer geeigneten Metrik, um die Ähnlichkeit der Zeitreihen zu beurteilen. Die häufig für Clusternanwendungen genutzte euklidische Distanz ist hierfür ungeeignet, da diese nicht auf Permutationen der Zeitreihenwerte reagiert und daher die serielle Abhängigkeit von Zeitreihenwerten außer Acht lässt (Caiado et al. 2015). Neben direktem und modellbasiertem Clustern von Zeitreihen ist featurebasiertes Clustern ein Ansatz zur Gruppierung ähnlicher Zeitreihen. Dieser wurde auch für die vorgestellte Methodik gewählt, da die direkte Verarbeitung von Zeitreihen mit Self-Organizing-Maps prinzipiell möglich ist, je-

Tabelle 2.2: Verteilung der Messstellen auf die hydrogeologischen Großräume

Großraum	Anzahl Messstellen	
	Originaldatensatz	Preprocessed
1	6.023	5.881
2	1.027	1.021
3	2.497	2.462
4	650	649
5	1.376	1.367
6	662	655
7	17	17
8	730	722
9	435	435
10	75	75
Summe	13.492	13.284

doch hierfür, wie für viele andere Methoden auch, lückenlose Zeitreihen eines identischen Zeitraumes mit gleicher Anzahl von Zeitschritten notwendig wären. Um dieses Problem zu umgehen, wird jede Zeitreihe zunächst mittels sogenannter Features beschrieben. Diese Features sind dazu entworfen, die Aspekte der Grundwasserdynamik zu beschreiben und dienen ihrerseits dann wiederum als Input für die SOM. Featurebasiertes Clustern von Zeitreihen mit SOM wurde beispielsweise von Wang et al. (2005, 2006, 2004) durchgeführt, die Features zur allgemeinen Beschreibung von Zeitreihen zusammengestellt haben. Auf die Berechnung und die zugrunde liegenden Überlegungen der einzelnen grundwasserspezifischen Features wird in den folgenden Abschnitten eingegangen. Tab. 2.3 nimmt bereits eine Übersicht aller Features vorweg.

Tabelle 2.3: Übersicht der Features zur Beschreibung der Zeitreihendynamik

Name(n)	Daten	Zweck
RangeRatio	orig.	überlagernde langperiodische Signale
LongestRecession	orig.	(unnatürlich) lang fallende Grundwasserstände
std(d/dx), F_{dif}	orig.	Variabilität (Häufigkeit)
Periodicity, $P52_{corr}$	orig.	Anteil des Jahresgangs an der Variabilität
Skewness, Schiefe	orig.	Inhomogenitäten, Ausreißer, häufige Extremwerte
SEM	orig.	Streuungsmaß
jumps	z-normiert	Inhomogenitäten, Variabilität
ExVals	orig.	Rauigkeit, hochfrequente Variabilität
SeasonalBehaviour	z-normiert	Zeitpunkt des Maximums im Jahresgang
YearlyVariance, Var_y	z-normiert	Variabilität, Periodizität
diffsum	z-normiert	Variabilität (Betrag)
m_{up} / m_{down}	orig.	Steig- und Fallverhalten, Symmetrie von GW-Peaks

Die Features sind so konzipiert, dass sie entweder die Grundwasserdynamik beschreiben oder dass eine Eigenschaft der Zeitreihe detektiert wird, die so auffällig ist, dass entsprechende Zeitreihen zusammen gruppiert werden sollen. Hierbei handelt es sich um eine Sammlung aller Features, aus denen im weiteren Verlauf nach mathematischen und hy-

drogeologischen Gesichtspunkten geeignete Kandidaten vorausgewählt werden (vgl. Abschnitt 2.3.7). Dies ist notwendig, da es prinzipiell unterschiedliche Arten der Grundwasserdynamiken gibt, die sich mit verschiedenen Features unterschiedlich gut beschreiben lassen. So unterscheidet sich z.B. die typische Dynamik in einem Porengrundwasserleiter ganz erheblich von der in einem Karstgrundwasserleiter, was in der Regel andere Features zur Beschreibung erfordert. Der für jedes Gebiet jeweils final eingesetzte Feature-Satz wird in Tabelle 2.5 dargestellt.

Feature: RangeRatio

Das Feature *RangeRatio* betrachtet das Verhältnis der mittleren jährlichen Spannweite ($\overline{rng_y}$) zur maximalen Spannweite (rng_{max}) der Grundwasserstandszeitreihe. Primärer Zweck dieses Features ist die Unterscheidung von Ganglinien mit und ohne überlagerndes langperiodisches Signal.

$$RangeRatio = \frac{\overline{rng_y}}{rng_{max}} \quad (2.1)$$

Neben langperiodischen Schwankungen führen auch deutliche Ausreißer oder abschnittsweise Absenkungen/Erhöhungen zu niedrigen *RangeRatio*-Werten. Zur Berechnung von $\overline{rng_y}$ werden nur die Spannweiten von Jahren gemittelt, für die maximal 16 Fehlstellen vorliegen. So wird gewährleistet, dass die jährliche Spannweite mit hoher Wahrscheinlichkeit sowohl das Jahresmaximum als auch das Jahresminimum abbildet.

Feature: LongestRecession

Das Feature *LongestRecession* sucht den längsten Abschnitt einer Zeitreihe TS ohne positive Steigungswerte. Dies ermöglicht es, Zeitreihen zu identifizieren, die Abschnitte mit (unnatürlich) lang fallenden Grundwasserständen aufweisen. Hierbei kann sogar der Jahresgang vollständig überprägt sein.

$$LongestRecession = \left(\frac{d}{dt}TS(t), \frac{d}{dt}TS(t+1), \dots \right)_{max}, \quad \forall \frac{d}{dt}TS(t) \leq 0 \quad (2.2)$$

Feature: F_{dif}

Das Feature F_{dif} beschreibt die Häufigkeit hoher Änderungsraten in einer Zeitreihe TS und ist deshalb ein Maß für die Variabilität. Berechnet wird dafür die Standardabweichung σ der 1. Ableitung.

$$F_{dif} = \sigma \left(\frac{d}{dt}TS(t) \right) \quad (2.3)$$

Feature: Periodizität

Das Feature Periodizität ist als Maß für den Anteil der jährlichen Periodizität an der Variabilität einer Zeitreihe konzipiert. Hierfür wird zunächst die mittlere jährliche Periodizität ($TS_{periodic}$) durch einfache Mittelung korrespondierender Werte der Einzeljahre extrahiert. Zu diesem Zweck wird eine Periodenlänge von 52 (Anzahl Wochen/Messwerte

in einem Jahr) angenommen und alle Werte im Abstand 52 gemittelt. Da ein Jahr immer aus 52 Wochen besteht, je nach Verteilung der Wochentage aber 53 Montage und damit 53 Werte pro Jahr vorhanden sein können, wird an dieser Stelle ein kleiner Fehler in Kauf genommen. Den schlussendlichen Feature-Wert erhält man durch Berechnung des Pearson R zwischen TS und $TS_{periodic}$:

$$P52_{corr} = corr(TS, TS_{periodic}) \quad (2.4)$$

Feature: Schiefe

Dieses Feature beschreibt die Schiefe der Werteverteilung einer Zeitreihe TS mit Länge n . Die Schiefe beschreibt allgemein die Asymmetrie einer Wahrscheinlichkeitsverteilung und zeigt an, ob und wie stark die Verteilung linksschief (negativ) oder rechtsschief (positiv) ist. Die Berechnung erfolgt nach:

$$Schiefe = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^3 \quad (2.5)$$

Die Schiefe ist zum Einen gut geeignet, um nach Ausreißern oder längeren Absenken/Hochständen zu gruppieren, zum Anderen lässt sich gut die Dynamik von Messstellen mit häufigen und schnell wieder abklingenden Extremwerten beschreiben.

Feature: Standardfehler - SEM

Der Standardfehler ist ein auf die Größe der Stichprobe, hier die Länge n der Zeitreihe, normiertes Maß für die Varianz. Der Standardfehler heißt auch Standard Error of the Mean (SEM) und berechnet sich nach:

$$SEM = \frac{\sigma}{\sqrt{n}} \quad (2.6)$$

In der Regel ist der Standardfehler gut mit der absoluten Spannweite einer Zeitreihe korreliert und deshalb ein wichtiges Feature, um Ganglinien mit auffälligen Amplituden zu gruppieren.

Feature: Jumps

Das Feature $Jumps$ soll Inhomogenitäten in Form von Sprüngen in der Zeitreihe detektieren. Dazu wird die Mittelwertänderung aufeinanderfolgender Jahre betragsweise betrachtet. Analog zu $RangeRatio$ werden nur Jahre einbezogen, die maximal 16 Fehlwerte aufweisen, damit der Mittelwert den Jahresgang realistisch repräsentiert. Die Berechnung von $Jumps$ erfolgt im Gegensatz zu den bisher genannten Features auf Basis der z-normierten Daten (vgl. Tab.2.3), da Inhomogenitäten skalenunabhängig und nur relativ zur Dynamik der einzelnen Zeitreihe betrachtet werden sollen. Wenn J die Zeitreihe der Jahresmittelwerte ist, dann folgt:

$$Jumps = \left(\frac{\left| \frac{d}{dt} J(t) \right|_{max}}{\left| \frac{d}{dt} J(t) \right|} \right)^2 \quad (2.7)$$

Die Normierung der maximalen Mittelwertänderung durch die mittlere Änderung aufeinanderfolgender Jahre ermöglicht die Unterscheidung zwischen echten Inhomogenitäten und regelmäßig stattfindenden Änderungen. Um große *Jumps*-Werte hervorzuheben, erfolgt abschließend eine Quadrierung der Werte.

Feature: ExVals

Das Feature *ExVals* beschreibt die Rauigkeit einer Zeitreihe, indem die Anzahl der Hochpunkte (Peaks), normiert auf die Zeitreihenlänge n berechnet wird.

$$ExVals = \frac{\left\{ TS(t) \mid \frac{d}{dt}TS(t) = 0 \wedge \frac{d^2}{dt^2}TS(t) < 0 \right\}}{n} \quad (2.8)$$

Durch diesen Wert werden insbesondere kleine, hochfrequente Schwankungen mit geringen Amplituden erfasst, Ganglinien also, die nahezu zwischen jedem Wert einen Richtungswechsel absolvieren und nur übergeordnet einer anderen Periodizität oder einem Trend folgen.

Feature: SeasonalBehaviour

Das Feature *SeasonalBehaviour* dient der Unterscheidung der Zeitreihen nach der Position von Jahresmaximum und -minimum. Hierfür werden monatsweise alle korrespondierenden Werte der Zeitreihe gemittelt und ein mittlerer Jahresgang SB der Messstelle bestehend aus zwölf Werten generiert. Dieser wird mit dem Verlauf einer Modellkurve MC (Sinuskurve mit Maximum im März und Minimum im September) korreliert (Pearson R). Um auch die normierte Amplitude miteinzubeziehen, erfolgt darüber hinaus eine Berechnung der L_2 -Norm (euklidischer Abstand) beider Kurven SB und MC . Ist die Korrelation beider Kurven kleiner Null, so wird die L_2 -Norm zwischen SB und der inversen Modellkurve MC_{inv} (Minimum und Maximum vertauscht) berechnet. Als Grundlage für *SeasonalBehaviour* dienen z-normierte Daten, um einen sinnvollen Abgleich von SB und MC zu ermöglichen.

$$R = corr(SB, MC) \quad (2.9)$$

$$SeasonalBehaviour = \frac{R}{d} \begin{cases} d = \|SB - MC\|, & \forall R \geq 0 \\ d = \|SB - MC_{inv}\|, & \forall R < 0 \end{cases} \quad (2.10)$$

Feature: YearlyVariance

Das Feature *YearlyVariance* (Var_y) ist der Median der jährlichen Varianz μ einer z-normierten Zeitreihe. Die Berechnung erfolgt erneut nur für Jahre J mit weniger als 17 Fehlstellen. Dadurch wird eine Kombination aus Periodizität und Varianz der Zeitreihe erfasst.

$$Var_y = \widetilde{\mu}(J) \quad (2.11)$$

Oft ist die Korrelation zum Feature *Periodizität* hoch, da stark jahresperiodische Zeitreihen auch im Mittel, relativ gesehen, hohe jährliche Varianzen aufweisen. Umgekehrt jedoch können auch hohe Var_y Werte erzielt werden, wenn keine starke regelmäßige Jahresperiodizität ausgeprägt ist. Der Vergleich von Var_y -Werten unterschiedlicher Zeitreihen erfolgt am sinnvollsten auf Basis von z-normierten Daten.

Feature: *diffsum*

Das Feature *diffsum* erfasst auf neue Weise die Variabilität der Zeitreihe. Es handelt sich bei diesem Feature um den Median der Summe der Beträge aller Ableitungen. Ein sinnvoller Vergleich unterschiedlicher Zeitreihen ist vor allem für normierte Daten möglich.

$$\sum_{i=1}^n \left(\left| \frac{d}{dt} x_i \right| \right) \quad (2.12)$$

Auch wenn wie bei F_{dif} die Variabilität erfasst wird, so korrelieren beide Features in der Regel fast nicht. Ersteres Feature erfasst die Häufigkeit starker Steigungen, letzteres den mittleren Betrag der Steigungen.

Feature: m_{up} / m_{down}

Das Feature m_{up}/m_{down} beschreibt die Steig- und Fall-Dynamik einer Zeitreihe. Es handelt sich hierbei um das Verhältnis der Mittelwerte jeweils aller positiven und negativen Ableitungen im Betrag. Hier wird also die Symmetrie der An- und Abstiege erfasst, und Ganglinien, deren Peaks ein langes Tailing aufweisen (z.B. im Karst), können so charakterisiert werden.

$$m_{up}/m_{down} = \frac{\overline{\left\{ \frac{d}{dt} x(t) \right\}_+}}{\left| \left\{ \frac{d}{dt} x(t) \right\}_- \right|} \quad (2.13)$$

2.3.2 Self-Organizing-Maps

Bei Self-Organizing-Maps (SOM) handelt es sich um unüberwacht lernende Künstliche Neuronale Netze (KNN), die zurückgehen auf Kohonen (1987, 1990). Die Anwendung von SOM ist weit verbreitet in der explorativen Datenanalyse, statistischen Datenanalyse sowie ganz allgemein bei der Mustererkennung (Cluster- und Klassifikationsaufgaben) und in den unterschiedlichsten Disziplinen (Finanzwesen, Biologie und Medizin, Telekommunikation, Autonome Mobilität, u.a.) (Kohonen, 2014).

SOM sind eine nicht-lineare Projektion mehrdimensionaler Daten auf eine regelmäßige Gitter-Oberfläche aus Neuronen. Jedes Neuron kann dabei als ein lokales Mittelwertmodell der ihm zugeordneten und umgebenden Datenpunkte verstanden werden. Weiterhin erhalten SOM während des Trainings die Topologie des Ausgangsdatensatzes und sind deshalb ebenfalls eine Daten-Komprimierungsmethode. Diese Verbindung von lokaler Datenmittelung bei gleichzeitiger Topologieerhaltung ist das Alleinstellungsmerkmal von SOM gegenüber anderen Clusteralgorithmen und Projektionsmethoden (Kohonen,

2014). Die Beziehung der Neuronen zueinander repräsentiert schlussendlich die Beziehung der Daten im Originaldatensatz. Durch die klaren Nachbarschaftsbeziehungen der Neuronen kann das Gitter einfach als zweidimensionale Fläche dargestellt werden, auch wenn die SOM an sich weiterhin mehrdimensionale Informationen beinhaltet. Dies eröffnet u.a. besonders einfache Visualisierungsmöglichkeiten für schwer darstellbare mehrdimensionale Daten.

Das Training einer Self-Organizing-Map läuft grob nach folgendem Schema ab. Zunächst werden die Neuronen des Netzes zufällig oder gezielt initialisiert und anschließend das erste Inputsignal präsentiert. Das Neuron, welches dem Input am ähnlichsten ist, ist der Gewinner (winner-takes-it-all Prinzip), auch Best-Matching-Unit *BMU* genannt und wird einer Lernregel entsprechend in Richtung des Inputs verschoben (im mehrdimensionalen Raum). Zusätzlich werden auch benachbarte Neuronen dem Grad ihrer Nachbarschaftsbeziehung zum Gewinner-Neuron entsprechend weniger stark in die Richtung des Inputs verschoben. Diese Trainingsschritte können sequenziell oder als Batch-Training durchgeführt werden. Für eine detaillierte Beschreibung der Algorithmen und der Trainingsschritte sei an dieser Stelle beispielsweise auf Vesanto (2000) oder Vesanto (2002) verwiesen.

Für die Anwendung der SOM wird Matlab 2018b (Mathworks Inc., 2018) in Kombination mit der SOM-Toolbox (Vesanto, 2000, 2005) verwendet. Der Input-Datensatz wird zunächst für jedes Feature einzeln z -normiert, damit ähnliche Wertebereiche der Inputparameter für die Anwendung der SOM vorliegen. Die Anzahl der Neuronen der SOM wird entweder über die Formel $5\sqrt{n}$ (Vesanto, 2005) oder alternativ über die vierfache Anzahl (Toolbox Optionen „normal“ und „big“) festgelegt. Die Form der SOM ist asymmetrisch, und das Seitenverhältnis ist engst-möglich angelehnt an das Verhältnis der ersten beiden Hauptkomponenten des Ausgangsdatensatzes (Vesanto, 2005). Dadurch wird die Erhaltung der Topologie während des Trainings vereinfacht. Initialisiert wird die SOM linear, entsprechend einer Fläche, die entlang der ersten beiden Hauptkomponenten im Eingaberaum aufgespannt wird. Dadurch wird das Training beschleunigt und das Ergebnis reproduzierbar. Des Weiteren wird auf den Batch-Trainingsalgorithmus und auf eine hexagonale Gittergeometrie zurückgegriffen.

2.3.3 DS2L Clusteralgorithmus

Obwohl das Training von Self-Organizing-Maps selbst bereits ein Clustering der Daten darstellt, ist die Interpretation der trainierten SOM und die Ableitung von Clustern nicht trivial. Eine Möglichkeit bietet der *density-based simultaneous two-level* (DS2L)-Algorithmus von Cabanes et al. (2012). Die Grundidee dieses Clusteralgorithmus ist ein sogenanntes SOM-Enrichment. Dabei handelt es sich um eine SOM, für die während des Trainings zusätzliche Informationen generiert werden, welche dann zur Clusterdefinition herangezogen werden. Konkret werden ein Dichte- und ein Nachbarschaftswert berechnet. Die Dichte gilt dann als Maß für die Datendichte, die ein Neuron umgibt, der Nachbarschaftswert ermöglicht die Bewertung der Nachbarschaftsbeziehungen zweier Neuronen. Dieser ist umso höher, je häufiger zwei benachbarte Neuronen *BMU* (vgl. Abschnitt 2.3.2) und *second-BMU* eines Inputs sind. Die Cluster werden dann so festgelegt, dass der Kernbereich eines Clusters durch hohe Dichte gekennzeichnet ist und die Clusterränder Bereichen mit niedriger Dichte, unbelegten Neuronen oder sehr schwachen

Nachbarschaftsbeziehungen entsprechen. Vorteile dieser Methode sind die automatische Bestimmung der Clusteranzahl auf Basis der enriched-SOM und die Handhabbarkeit von nicht-linear separierbaren und hyper-sphärischen Clustern. Weiterhin kann auch mit überlappenden Clustern und Rauschen noch ein Ergebnis erzielt werden.

Der Algorithmus arbeitet in vier Schritten, welche in Abb. 2.5 dargestellt sind. In der von Cabanes et al. (2012) veröffentlichten Form werden immer alle Schritte vollzogen. In der hier genutzten Version wurde der Algorithmus so modifiziert, dass dem Bearbeiter einige Möglichkeiten zur Verfügung stehen, das Clusterergebnis zu beeinflussen. Vor allen Dingen kann so grob gesteuert werden, ob eine sinnvolle Clusteranzahl erzielt wird. Schritt 1 nutzt die Nachbarschaftswerte, um klar separierte Cluster zu trennen. Hier wurde der Schwellenwert NTH (*Neighbourhood-THreshold*) als aktiver Steuerparameter eingeführt. Höhere NTH -Werte führen, vereinfacht ausgedrückt, zu mehr Clustern. In Schritt 2 und 3 werden lokale Dichtemaxima gesucht und mit Hilfe einer *watersheds*-Methode jedem Cluster ein Bereich zugeordnet, jedoch nur innerhalb der in Schritt 1 definierten Grenzen. Hierfür wird der Dichtewert als dreidimensionale Oberfläche interpretiert und analog zu einer realen Topographie der Verlauf von theoretischen Wasserscheiden berechnet. Die Schritte 2 und 3 sind über den Parameter DR (*Density-Refinement*) steuerbar. Schritt 4 nutzt ein Ähnlichkeitskriterium basierend auf Yue et al. (2004), um mutmaßlich zusammengehörige Cluster zusammenzulegen (Parameter DM - *Density-Merging*).

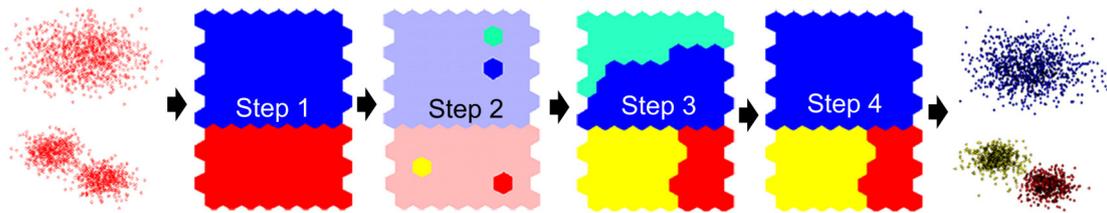


Abbildung 2.5: Übersicht über die Clusterschritte des DS2L-Algorithmus, geändert nach Cabanes et al. (2012)

Alle Schritte können über die Steuerparameter genutzt oder ausgelassen werden (vgl. Anhang A). In der Praxis hat sich gezeigt, dass vor allem NTH und DM sehr nützlich beim Einstellen einer sinnvollen Gruppierung sein können. Die Schritte 2 und 3 (Parameter DR) werden nahezu immer durchgeführt. Obwohl recht schnell intuitiv die Ähnlichkeit innerhalb eines Clusters optisch erfasst werden kann, sollte an dieser Stelle unbedingt mit hydrogeologischem Sachverstand beurteilt werden, welche Gruppierungen als sinnvoll erachtet werden können. Die Steuerparameter werden im weiteren Verlauf als DS2L-Parameter bezeichnet. Die Originalskripte des DS2L-Algorithmus wurden freundlicherweise von G. Cabanes bereitgestellt und von den Autoren wie beschrieben modifiziert.

Als zusätzlicher DS2L-Parameter wurde außerdem $HMTH$ - *hits merging threshold* etabliert, welcher ermöglicht, alle Cluster mit weniger als $HMTH$ Messstellen zusammenzuführen. Dies ist bei hohen Clusteranzahlen und vielen kleinen Clustern nützlich, da die Übersichtlichkeit erhöht wird, sollte jedoch nur bei Einzelanwendungen des Clustering und nicht in Verbindung mit Ensembles genutzt werden. Entsprechend besitzt der Parameter $HMTH$ bei allen in diesem Bericht vorgestellten Ergebnissen den Wert 0.

2.3.4 Ensemblemodellierung

Der Begriff Ensemblemodellierung beschreibt die Generierung eines Ergebnisses aus den Aussagen vieler verschiedener Modelle anstelle eines einzelnen Modelles. Dies ist vor allem im Kontext der Klassifikation ein etabliertes Verfahren und findet auch im Bereich des Clusterings seit einigen Jahren häufiger Anwendung. Trotz des erhöhten Zeit- und Rechenaufwandes von Ensemblemodellierungen überwiegen die Vorteile. Durch die Kombination vieler Modelle kann die Qualität und Konsistenz (Robustheit) eines Clusterergebnisses erheblich verbessert werden (Alqurashi und Wang, 2018; Ghosh und Acharya, 2011; Vega-Pons und Ruiz-Shulcloper, 2011). Besonders im Anwendungsbereich des Clusterings, für das im Gegensatz zu Klassifikationsaufgaben keine Lösung vorgegeben werden kann und demnach weder die Anzahl der Cluster bekannt ist noch Validierungsmethoden wie Kreuzvalidierung angewandt werden können, ist die Verwendung von Ensembles nicht trivial. Wie in Abb. 2.6 dargestellt, wird der Ensemble-Clusterprozess in der Regel in zwei Hauptschritte eingeteilt (Vega-Pons und Ruiz-Shulcloper, 2011):

1. Generierung der Ensemblemitglieder
2. Anwendung einer sog. *Consensus-Function* zur Aggregierung der Ergebnisse

Vor allem die Auswahl einer geeigneten Consensus-Function, die die Einzelergebnisse der Ensemblemitglieder zusammenführt und so die Qualität des Clusterergebnisses verbessern kann, bestimmt über den Erfolg eines Clusterensembles (Vega-Pons und Ruiz-Shulcloper, 2011).

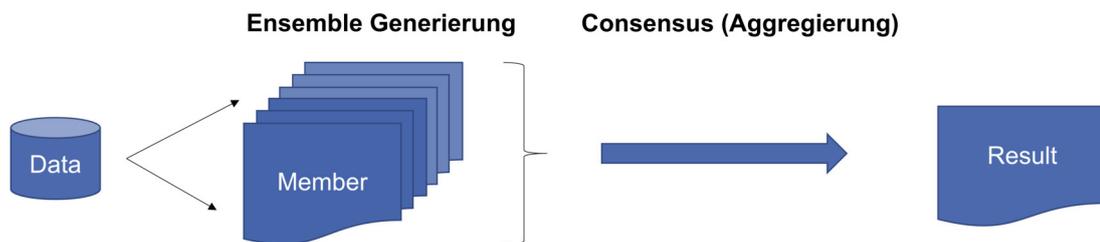


Abbildung 2.6: Schritte einer Ensemblemodellierung

Für den 1. Schritt, die Generierung der Ensemble-Mitglieder, kommen verschiedene Methoden in Frage. Ziel ist es, bei allen Ansätzen eine ausreichend hohe Diversität zu erzeugen, um die Konsistenz und Robustheit des Ensembles zu erhöhen, wenngleich es noch strittig ist, ob eine möglichst hohe oder nur mittlere Diversität anzustreben ist (Alqurashi, 2017). Um dieses Ziel zu erreichen, können variable Modellierungsansätze gewählt oder auch verschiedene Clusteralgorithmen genutzt werden. Unterschiedliche Initialisierungswerte oder Clusterparameter (hier: z.B. DS2L-Parameter) eines Clusteralgorithmus können ebenso genutzt werden, um ein Ensemble zu erstellen. Nützlich sind zudem subsampling und resampling Techniken, die den zu clusternden Datensatz manipulieren. Diese Möglichkeit wurde auch für die hier vorgestellte Methodik gewählt.

delete-d-Jackknife

Bei Jackknifing handelt es sich um eine resampling-Methode, bei der der zu untersuchende Datensatz systematisch oder zufällig reduziert wird. Im Gegensatz zum ebenfalls weitverbreiteten Bootstrapping geschieht dies jedoch ohne Zurücklegen der gewählten Probe. Jedes Element kommt also nur einmal im Datensatz vor. Im Falle von *delete-d-Jackknifing* werden zufällig für jedes Member d Proben aus dem Datensatz gelöscht. Als Richtlinie für die Größe von d bei n Stichproben gilt nach Sinharay (2010):

$$d > \sqrt{n} \quad (2.14)$$

Damit nicht beliebig viele Daten beim Jackknifing ausgelassen werden, wurde die maximale Größe von d auf 10% der Datensatzgröße n begrenzt, für kleine Datensätze, für die $0.1 \times n < \sqrt{n}$ gilt, wird $d = \lceil \sqrt{n} \rceil$ gesetzt.

Voting Consensus

Schritt 2 einer Ensemblemodellierung ist die Ergebnisaggregation mittels Consensus-Funktion. Für den vorliegenden Anwendungsfall fiel die Wahl auf das sog. *Voting*, auch bekannt als *direct approach* oder *relabeling approach*. Der genutzte Algorithmus ist von Shestakov (2017) auf Basis von Ayad und Kamel (2010) geschrieben. Die Grundidee von Voting ist es, eine Permutation von Cluster-Labels zu finden, die die Übereinstimmung zwischen zwei Clusterergebnissen (bestehend aus Cluster-Labels) maximiert. Dies resultiert in zwei Hauptproblemen, die dabei vom Algorithmus gelöst werden müssen (Vega-Pons und Ruiz-Shulcloper, 2011):

1. Zuordnungsproblem (weitere Bezeichnungen hierfür sind *assignment problem*, *matching problem*, *relabeling problem*, *label-correspondence problem*)
2. Aggregierungsproblem (eigentliches Voting)

Das Zuordnungsproblem ist nichts anderes als ein bipartites (gewichtetes) Matchingproblem aus der Graphentheorie (Abbildung 2.7). Zur Lösung wird der Ungarische-Algorithmus oder Hungarian-Algorithm (Kuhn, 1955; Munkres, 1957) verwendet, ein Algorithmus zur allgemeinen Lösung bipartiter Zuordnungsprobleme. Dieser ist auch bekannt unter Kuhn-Munkres-Algorithmus. Das genutzte Skript wurde verfasst von Buehren (2011). Die Komplexität des Zuordnungsproblems ist $O(k!)$, wird jedoch durch die Verwendung des Hungarian Algorithm reduziert auf $O(k^3)$, wobei k der Clusteranzahl der Consensus-Partitionierung entspricht (Ayad und Kamel, 2010).

Die Lösung des Aggregierungsproblems ist das eigentliche Voting. Hierfür wird für jedes Sample, in unserem Fall Messstellen, das am häufigste gewählte Label (majority voting) als Lösung ausgewählt (Hanczar und Nadif, 2013). Cluster-Labels können neben Zahlen auch jede andere Art von Bezeichnung sein. Voting weist eine geringe Startwertabhängigkeit auf, da es sich um einen von der ersten Clusterlösung ausgehenden iterativen Prozess handelt. Im Sinne der im nächsten Abschnitt 2.3.5 dargestellten Validierungsindizes beträgt die Variation der Indizes für die Voting-Lösungen meist wenige Prozent.

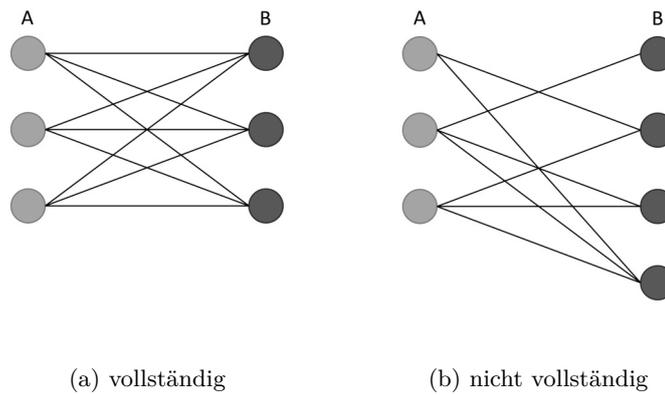


Abbildung 2.7: vollständige (a) und nicht vollständige (b) bipartite Graphen mit Punkt-
mengen A und B

Beim Voting selbst spielen diese Indizes keine Rolle, sie dienen lediglich nachträglich der Beurteilung.

2.3.5 Cluster-Validierungs-Indizes

Gemeinhin unterscheidet man Cluster-Validierungs-Indizes, die zur qualitativen Beurteilung eines Clusterings herangezogen werden, nach den zur Verfügung stehenden Informationen in drei Gruppen: intern, extern und relativ (Jain und Dubes, 1988).

1. intern: Beurteilung, wie gut ein Clusterergebnis die Daten repräsentiert, auf Basis der beim Clustern selbst zur Verfügung stehenden Daten
2. extern: Beurteilung, wie gut ein Clusterergebnis mit einer vordefinierten Struktur übereinstimmt
3. relativ: Vergleich zweier oder mehrerer Strukturen, oft auf Basis von internen Indizes

Die Entwicklung und Beurteilung von Validierungsindizes ist noch immer ein aktives Forschungsthema. Dies führt zur Entwicklung von immer neuen Indizes, wodurch die Auswahl geeigneter Indizes erschwert wird. Mit unterschiedlichen Indizes sind auch unterschiedliche Schlussfolgerungen möglich, da jeder Index explizite und implizite Annahmen trifft und so für verschiedene Anwendungen unterschiedlich gut geeignet ist. Selten sind diese Annahmen wirklich gut dokumentiert oder gänzlich eindeutig. Da für reale Datensätze in der Regel kein *bestes Clusterergebnis* existiert und die scheinbar optimale Lösung je nach Index variiert, wurden zur Beurteilung der Clusterergebnisse fünf verschiedene Indizes herangezogen. Diese sind in Tabelle 2.4 zusammengefasst. Auf eine ausführliche Herleitung der Indizes wird an dieser Stelle verzichtet und auf die gute Dokumentation des *R*-Pakets *clusterCrit* (Desgraupes, 2018) verwiesen. Auszüge davon inklusive der Formeln zur Berechnung sind in Anhang B zusammengestellt. Das Paket *clusterCrit* wurde auch zur Berechnung der Indizes in *R* genutzt.

Die Indizes CH, MR und PBM beziehen in ihre Berechnung sowohl die Kompaktheit der Cluster als auch deren Separierung ein, RL und C-Index betrachten jeweils nur die

Tabelle 2.4: Zusammenstellung der verwendeten Cluster-Validierungs-Indizes und deren Zielwerte

Index	Abkürzung	Zielwert	Referenz
Calinski-Harabasz	CH	Max	Caliński und Harabasz (1974)
McClain-Rao	MR	Min	McClain und Rao (1975)
PBM	PBM	Max	Pakhira et al. (2004)
Ratkowsky-Lance	RL	Max	Ratkowsky und Lance (1978)
C-Index	C-Index	Min	Hubert und Schultz (1976)

Separierung bzw. die Kompaktheit der Cluster. Gemäß der Angabe in der Spalte Zielwert von Tabelle 2.4 sollen die Werte für CH, PBM und RL maximiert werden, für MR und C-Index wird eine Minimierung des Indexwertes angestrebt.

2.3.6 Referenzmessstellenauswahl

Die Grundidee für die Referenzmessstellenauswahl ist die Auswahl einer Messstelle, deren Dynamik für ein gewisses Gebiet repräsentativ ist, das heißt, dass sich die Grundwasserdynamik an anderen Messstellen des Gebiets ähnlich wie in der Referenzmessstelle verhält und damit z.B. Vorhersagen von Grundwasserständen in der Referenzmessstelle auch auf den Rest des Gebiets übertragbar sind. In erster Näherung kann man hier annehmen, dass das erste Gesetz der Geographie gilt, d.h. dass sich die Dynamik in räumlich nahe beieinander liegenden Messstellen ähnlicher verhält, als bei Messstellen, die weiter entfernt sind. Darüber hinaus haben neben der räumlichen Lage jedoch zahlreiche weitere Einflussfaktoren, die sich zum Teil aus der komplexen, dreidimensionalen Natur hydrogeologischer Aspekte ergeben, einen Einfluss auf die Grundwasserdynamik. Selbst wenn man nur flache Messstellen betrachtet, die alle im gleichen Grundwasserleiter verfiltert sind, so können sich z.B. aus verschiedenen Flurabständen, Durchlässigkeiten der ungesättigten Zone oder Grundwasserneubildungsraten, die alle einen Einfluss auf die Dämpfung von Niederschlagssignalen haben, sehr unterschiedliche Grundwasserdynamiken ergeben.

Mit dem gewählten Ansatz des Clusterings von Grundwasserganglinien wird die räumliche Lage zunächst außer Acht gelassen. Das Clustering erfolgt rein auf Basis der mit den Features beschriebenen Grundwasserdynamik der Zeitreihen. Von daher können sich im Idealfall Cluster bilden, in denen Messstellen gruppiert werden, die auch räumlich nahe beieinander liegen und daher tatsächlich die Grundwasserdynamik eines (räumlich zusammenhängenden) Gebiets widerspiegeln. Darüber hinaus bilden sich jedoch auch Cluster, in denen Ganglinien gruppiert werden, die sich sehr ähnlich verhalten, die jedoch räumlich nicht nahe beieinander liegen. Das kann z.B. bei Messstellen der Fall sein, die von ihrer hydrogeologischen Eigenschaft her hauptsächlich auf den hydrologischen Jahresgang und nicht kurzfristig auf Niederschlagsereignisse oder sonstige lokale Einflussfaktoren reagieren. Weiterhin werden sich je nach Datenlage Cluster bilden, in denen Ganglinien gruppiert werden, die sich sehr speziell verhalten und die keiner anderen Ganglinie ähnlich sind. Die „Ähnlichkeit“, aufgrund derer die Ganglinien in das gleiche Cluster gruppiert werden, kann dann z.B. in der Gemeinsamkeit von Datenlücken, Sprüngen oder im Preprocessing nicht als solche erkannten Ausreißern liegen. Die letztgenannten Cluster sind prinzipiell weder zur Auswahl von Referenzmessstellen

noch für eine spätere Vorhersage von Grundwassermessstellen geeignet und können in der Regel verworfen werden.

Für die beiden anderen Clustertypen kann angenommen werden, dass sich die Messstellen bzw. Ganglinien innerhalb des Clusters ähnlich verhalten, und dass damit die Auswahl einer Referenzmessstelle, die später für die Vorhersage der anderen Messstellen des Clusters (unabhängig von deren räumlicher Lage) genutzt werden kann, sinnvoll ist.

Dabei wird das Clusterzentrum von der Ganglinie gebildet, deren beschreibende Features prinzipiell als am repräsentativsten für alle anderen Ganglinien des Clusters angesehen werden können. Mit steigendem Abstand zum Clusterzentrum sinkt auch die Ähnlichkeit zu diesem. Ob und wie stark die Ähnlichkeit dabei noch ausgeprägt ist, hängt dabei von der Kompaktheit des Clusters ab. Der Abstand zum Clusterzentrum wird über ein normiertes Distanzmaß d_{norm} beurteilt. Hierbei wird nach Gleichung 2.15 die euklidische Distanz einer Messstelle zum zugehörigen Clusterzentrum d mit der mittleren (Median) Distanz aller Messstellen zum Datensatzzentrum D_m normiert. Dies ermöglicht einen Vergleich von Datensätzen auf unterschiedlichen Skalen.

$$d_{norm} = \frac{d_1}{D_m} \quad (2.15)$$

Da das Clustern auf der Ähnlichkeit der die Ganglinien beschreibenden Features (vgl. Abschnitt 2.3.1) beruht, für die spätere Vorhersage aber eine Ähnlichkeit der Ganglinien hinsichtlich ihrer tatsächlichen Werte relevant ist, eignet sich für die Referenzmessstellenauswahl die Distanz zum Clusterzentrum nur bedingt. Ursache dafür ist, dass die Beschreibung von Ganglinien über die gewählten Features zwar prinzipiell geeignet ist, Ähnlichkeiten in der Dynamik zu beschreiben, jedoch die tatsächlichen Werte dabei außer Acht lässt. So kann es im Extremfall passieren, dass zwei Ganglinien mit identischen Feature-Werten beschrieben werden können und damit im gleichen Cluster und sehr nahe beieinander gruppiert werden, die eigentlichen Werte der Zeitreihe aber praktisch nicht miteinander korrelieren. In der Praxis ist das auf Grund des Clusters innerhalb hydrogeologischer Einheiten (Großräume bzw. Räume) nicht zu erwarten, dennoch hat eine Korrelationsanalyse innerhalb der gebildeten Cluster gezeigt, dass die Messstellen nahe des Clusterzentrums nicht zwingend am stärksten mit allen anderen Messstellen des Clusters korrelieren. Daher wurde für die Referenzmessstellenauswahl ein anderer Weg gewählt. Innerhalb eines Clusters wurde jede Ganglinie mit jeder anderen korreliert (Pearson-Korrelation). So erhält man eine Korrelationsmatrix mit Pearson-R und zugehörigen p-Wert (Signifikanz). Da man bei der Korrelation der Ganglinien davon ausgeht, dass prinzipiell ein Kausalzusammenhang besteht, wurden die p-Werte lediglich zur Gewichtung der Korrelations-Koeffizienten genutzt, nicht dazu, nicht signifikante Korrelationen auszusortieren. Hierbei wurde nach Gleichung 2.16 ein mit der Signifikanz gewichteter Korrelations-Koeffizient berechnet:

$$R_w = R * (1 - p) \quad (2.16)$$

Anschließend wurde für jede Ganglinie der Mittelwert der gewichteten Korrelations-Koeffizienten berechnet und die Ganglinien nach absteigendem $\overline{R_w}$ sortiert. So wird gewissermaßen das Clusterzentrum umdefiniert, so dass die Messstelle, welche im Schnitt am besten mit allen anderen korreliert, im (neuen) Clusterzentrum zum Liegen kommt.

Auch bei relativ homogenen Clustern tritt oft der Fall auf, dass bei Ganglinien am Rande des Clusters die Korrelation zu den restlichen Ganglinien des Clusters stark abfällt. Es handelt sich dabei um Ganglinien, die einen speziellen Verlauf zeigen ohne aber besondere Auffälligkeiten wie große Sprünge oder ähnliches. Sie passen letztendlich in kein Cluster wirklich gut und werden dann dem Cluster zugeordnet, dem sie noch am ähnlichsten sind. Da diese Messstellen für die Vorhersage nicht in Frage kommen, werden sie aussortiert, indem das Cluster ab einem bestimmten Schwellenwert von $\overline{R_w}$ zum Rand hin gekappt wird. Der Schwellenwert muss dabei von Cluster zu Cluster vom Bearbeiter individuell festgelegt werden, da die $\overline{R_w}$ -Werte stark von der Gesamt-Homogenität des Clusters abhängen. An der Ganglinedynamik ist jedoch meist gut erkennbar, welche Messstellen betroffen sind. Bei den meisten Clustern hat sich ein Schwellenwert von 0,5 bis 0,6 als geeignet erwiesen. Nach Kappung des Clusters wird für die verbleibenden Messstellen der Mittelwert der gewichteten Korrelations-Koeffizienten neu berechnet und die Ganglinien wiederum nach absteigendem $\overline{R_w}$ sortiert. So erhält man quasi eine bereinigte Homogenität des Clusters. Durch die Kappung der Cluster kann der Fall eintreten, dass die Cluster nicht mehr nach der Anzahl der beinhalteten Messstellen sortiert sind. Auf eine Um-Nummerierung wurde jedoch verzichtet.

Im weiteren Verlauf muss nun noch festgestellt werden, welche Messstelle als Referenzmessstelle in Frage kommt. Aus mathematischer Sicht sollte diese möglichst nah am Clusterzentrum liegen, d.h. einen möglichst hohen $\overline{R_w}$ -Wert aufweisen, um die restlichen Messstellen des Clusters ausreichend gut zu repräsentieren. Da die RM später zur Vorhersage genutzt werden sollen, sind darüber hinaus aber noch weitere Anforderungen, wie etwa lange und lückenlose Zeitreihen notwendig. Wie lange die Zeitreihen mindestens sein müssen, ist nur schwer definierbar. Daher wird von Fall zu Fall entschieden und bei gleichen $\overline{R_w}$ -Werten eine längere Zeitreihe im direkten Vergleich bevorzugt. Eine Liste der von den Autoren zunächst vorgeschlagenen Referenzmessstellen für jedes Cluster findet sich in Anhang C. Eine Beurteilung durch die Lokalexperten der Länder ergab keine Einwände, die den Austausch einzelner Referenzmessstellen nötig gemacht hätte, denn grundsätzlich könnte für die meisten Cluster eine benachbarte Ganglinie nahe des Clusterzentrums als Ersatz gewählt werden.

2.3.7 Methodensynthese: Workflow

Das für dieses Projekt entwickelte Auswahlverfahren für Referenzmessstellen ist eine Kombination aller bisher erläuterten Methoden. Abb. 2.8 fasst alle wichtigen Arbeitsschritte zusammen. Der erste Schritt ist die Erstellung eines Datensatzes für das zu untersuchende Gebiet aus der Gesamtdatenbank. Auch wenn die Gesamtmethodik prinzipiell in der Lage ist, alle Arten von Datenlücken oder Zeitreihenlängen zu verarbeiten, so haben diese Zeitreihenparameter einen teils deutlichen Einfluss auf das Clusterergebnis. Aus diesem Grund werden zunächst die Zeitreihen eines Gebietes gesichtet und ein sinnvoller Untersuchungszeitraum, eine Mindestlänge der Zeitreihen und ein maximaler Datenlückenanteil festgelegt. Für die beiden letzteren Parameter wurde für alle untersuchten Gebiete gleichermaßen eine Mindestlänge von 312 Werten und ein Anteil der Datenlücken von maximal 25% definiert. Prinzipiell sollten diese Werte jedoch an die Datenverfügbarkeit und -qualität eines Datensatzes angepasst werden. Die Zeitreihenlänge beträgt demnach mindestens sechs Jahre, wenn keine Datenlücken vorhanden sind. Prinzipiell werden alle Zeitreihen aussortiert, die keine Daten nach dem 1. Januar

2010 aufweisen, da diese als gesichert inaktiv angesehen werden. Das Jahr 2010 wurde hierbei erneut äußerst konservativ gewählt, da teilweise unklar war, ob neuere Daten möglicherweise vorliegen und nur (noch) nicht im gesammelten Datensatz vorhanden sind. Auf die jeweils festgelegten Untersuchungszeiträume wird im Ergebnisteil näher eingegangen, prinzipiell wurde jedoch immer ein Zeitraum gesucht, in dem die Mehrheit der Messstellen Daten aufweist.

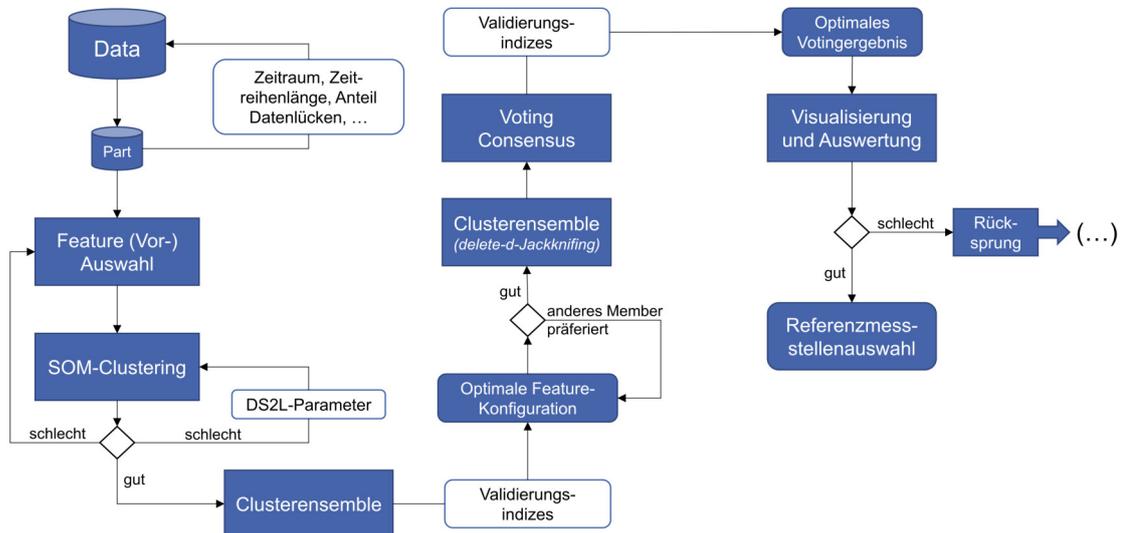


Abbildung 2.8: Workflow aller Arbeitsschritte zur Referenzmessstellenauswahl aus einer Datenbank von Grundwasserganglinien

Eine Feature-Vorauswahl ist notwendig, da der Workflow nur eingeschränkt dazu geeignet ist, aus einem Satz beliebiger Features die bestmögliche Kombination zu extrahieren. Diese beste Konfiguration beruht vollständig auf dem Ranking der fünf Cluster-Validierungs-Indizes, und diese wiederum können nicht unterscheiden, ob ein Feature Messstellen nach Dynamik gruppiert oder aufgrund anderer Merkmale (z.B. Messstellen mit schlechter Datenqualität wie etwa Sprünge). Rein mathematisch können somit gute Clusterergebnisse gefunden werden, die jedoch nicht dem Anspruch homogener Grundwasserdynamiken genügen. Vom hydrogeologischen Standpunkt ist die Notwendigkeit einer Feature-Auswahl gut nachvollziehbar, da sich die Dynamiken in unterschiedlichen Grundwasserleitertypen ganz grundsätzlich unterscheiden. Es ist also nur logisch, dass für unterschiedliche Gebiete auch verschiedene Feature-Kombinationen notwendig sind, um die Dynamik zu beschreiben.

Weiterhin sind darauffolgend einer oder mehrere Clusterdurchgänge nötig, um sinnvolle DS2L-Parameter (vgl. Abschnitt 2.3.3) auszuwählen, die dann im nächsten Schritt zur Erstellung des Clusterensembles dienen. Dieses Clusterensemble dient zur Bestimmung der optimalen Featurekonfiguration. Hierzu zählt sowohl die Auswahl der Features, als auch deren Reihenfolge. Letztere hat nur einen untergeordneten, jedoch einen nachweisbaren Einfluss auf das Clusterergebnis. Dies liegt an der Initialisierung der SOM (vgl. Abschnitt 2.3.2) auf Basis der Hauptkomponenten des Datensatzes. Diese ändern sich, wenn die Reihenfolge der Features, also die Dimensionen des Datensatzes vertauscht werden. Zur Erstellung des Ensembles werden alle möglichen Kombinationen zunächst ohne Berücksichtigung der Reihenfolge zusammengestellt. Dafür wird der Binomial-

koeffizient gemäß Gleichung 2.17 gelöst, wobei n die maximale Anzahl der zur Auswahl stehenden Features darstellt.

$$\binom{n}{k}, \quad \forall k \in \mathbb{N} : 1 \leq k \leq n \quad (2.17)$$

Um nun den Einfluss der Feature-Reihenfolge zu berücksichtigen, werden für jede mögliche Kombination aus 2.17 je nach Gebietsgröße unterschiedlich viele Permutationen gerechnet. Die Anzahl richtet sich nach dem dadurch erzeugten Rechenaufwand, liegt in der Regel aber zwischen 10 und 20. Als maximale Ensemblegröße, die mit realistischem Zeitaufwand bearbeitet werden kann, hat sich ca. 10.000 herauskristallisiert. Im nächsten Schritt ist ein Eingreifen des Bearbeiters möglich. Oft sind unterschiedliche Featurekonfigurationen nahezu gleich bewertet, enthalten aber unterschiedliche Features. Hier kann der Bearbeiter bei Bedarf eine nicht auf Platz 1 gerankte Konfiguration bevorzugt behandeln. Wenn eine optimale Konfiguration festgelegt wurde, wird, wie in Abschnitt 2.3.4 beschrieben, mittels *delete-d-Jackknife* ein Ensemble mit i.d.R. 5000 Mitgliedern erzeugt. Anschließend wird ein mehrfaches Voting durchgeführt, um der leichten Startwertabhängigkeit des Votings gerecht zu werden. Wie oft dieses konkret durchgeführt wird, wurde nach Rechenaufwand bestimmt, da der Zeitaufwand mit dritter Potenz zur Clusteranzahl multipliziert mit der Anzahl der Jackknife-Ensemblemitglieder zunimmt. Die genauen Ensemblegrößen zur Bestimmung der Feature-Konfiguration, des Jackknife-Ensembles sowie die Anzahl der Voting-Berechnungen sind in Anhang A aufgeschlüsselt. Die Votings werden erneut mit Hilfe der Cluster-Validierungs-Indizes in ein Ranking eingeordnet und anschließend ein optimales Votingergebnis ausgegeben. Dieses wird vom Bearbeiter hauptsächlich mittels Clustervisualisierungen beurteilt. Ist das Ergebnis nicht zufriedenstellend, kann zu mehreren Punkten im Workflow zurückgesprungen werden. Im Falle einer positiven Beurteilung kann mit der Referenzmessstellenauswahl fortgefahren werden.

2.4 Ergebnisse

2.4.1 Übersicht

In den nachfolgenden Kapiteln werden die Ergebnisse der Clusterings sowie die Auswahl möglicher Referenzmessstellen in den einzelnen Großräumen im Detail vorgestellt. Eine kurze hydrogeologische Beschreibung der Großräume, soweit sie zum Verständnis und für die Interpretation der Ergebnisse erforderlich ist, wurde weitgehend aus der Veröffentlichung Regionale Hydrogeologie von Deutschland der Ad-hoc-Arbeitsgruppe Hydrogeologie (2016) übernommen, auf welche für eine ausführlichere Darstellung verwiesen wird.

Großraum 3 (Oberrheingraben mit Mainzer Becken und nordhessischem Tertiär) nimmt in mehrfacher Hinsicht eine Sonderstellung ein und wird daher auch an den Anfang gestellt. Die weiteren Großräume folgen entsprechend ihrer Nummerierung in der Hydrogeologischen Raumgliederung von Deutschland.

Eine Zusammenfassung der Ergebnisse für alle Großräume sowie eine Diskussion und Bewertung hinsichtlich der im weiteren Verlauf anstehenden Vorhersage von Grundwas-

serständen findet sich in Kapitel 2.5. Hier sind in tabellarischer Form die Anzahl der Cluster pro Gebiet, die Anzahl verbleibender (zur Vorhersage potentiell geeigneter) und verworfener Cluster sowie Messstellen dargestellt. Die räumliche Verteilung von potentiellen Referenzmessstellen sowie zur Vorhersage potentiell geeigneter Messstellen sind in der Karte in Abbildung 2.9 dargestellt.

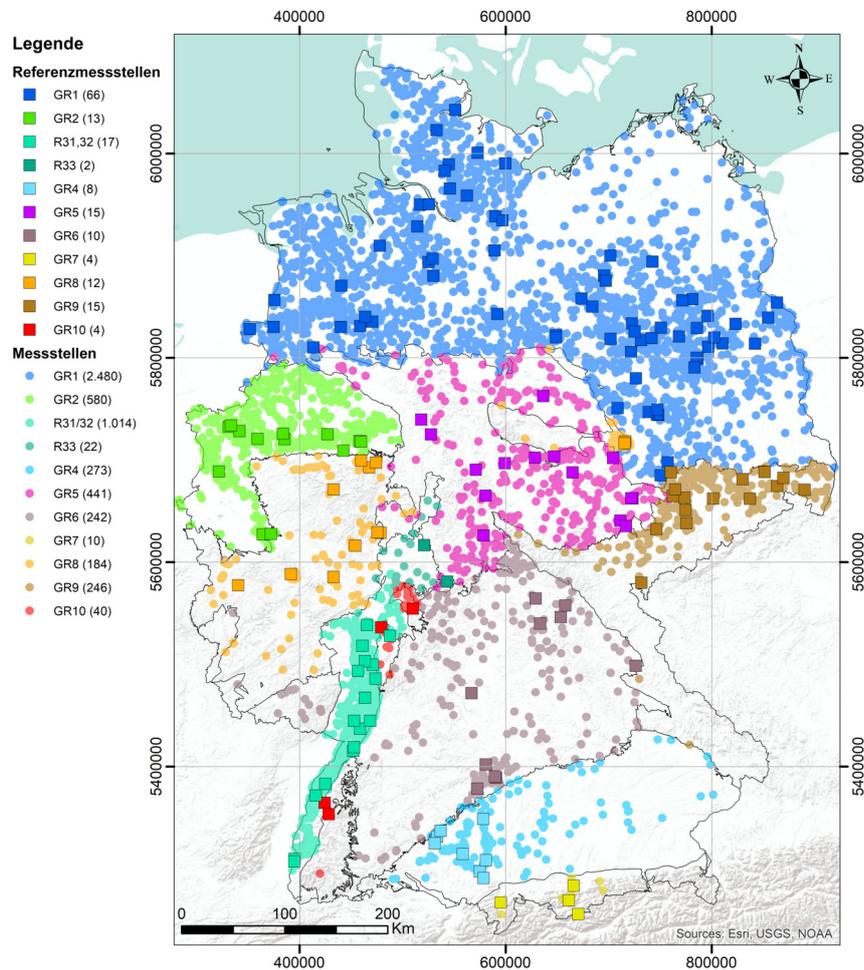


Abbildung 2.9: Lage und Anzahl von Referenz- und verbleibenden (zur späteren Vorhersage geeigneten) Messstellen, aufgeschlüsselt nach Clustergebieten.

Eine Liste aller Referenzmessstellen findet sich in Anhang C, in Anhang D ist für jedes Clustergebiete die Verteilung der Features pro Cluster als Boxplots dargestellt, wobei sich die Verteilung auf das Clusterergebnis vor der Kappung (siehe Abschnitt 2.3.6) bezieht.

Detaillierte Ergebnisse für jedes Cluster (Karten der Messstellenverteilungen und gestapelter Darstellung aller Ganglinien sowie für jedes Cluster maximale und mittlere R_w -Werte) wurden auf Grund der Vielzahl als elektronischer Anhang bereitgestellt. Dieser erhält darüber hinaus die Ergebnisse als georeferenzierte Daten.

Im Folgenden werden für jeden Großraum die Ergebnisse zusammenfassend beschrieben, wobei auf Grund der Vielzahl der gebildeten Cluster nur auf ausgewählte (in der Regel

große und/oder von der Dynamik her interessante Cluster) näher eingegangen wird. Tabelle 2.5 zeigt die für jedes Clustergebiet verwendeten Features.

Tabelle 2.5: Cluster und Messstellenanzahl der Clustergebiete mit Angabe der zur Erstellung genutzten Features.

	GR 1	GR 2	R 31,32	R 33	GR 4	GR 5	GR 6	GR 7	GR 8	GR 9	GR 10
Cluster	237	17	22	4	11	43	27	7	31	26	5
Messstellen	5.291	888	1.774	80	624	1.175	567	16	436	405	47
RangeRatio	x		x				x				x
LongestRecession			x				x				
F_{dif}					x	x	x				
$P52_{corr}$	x	x		x	x	x	x		x	x	
Schiefe	x	x	x		x	x	x	x	x	x	x
SEM	x	x			x	x	x	x	x		x
jumps			x		x		x				x
ExVals	x			x	x	x				x	
SeasonalBehaviour	x	x	x			x		x			
Var_y	x	x	x	x		x				x	
diffsum					x						x
m_{up}/m_{down}					x						

2.4.2 Großraum 3: Oberrheingraben mit Mainzer Becken und nordhessischem Tertiär

Der hydrogeologische Großraum Oberrheingraben mit Mainzer Becken und nordhessischem Tertiär ist gekennzeichnet durch die aktive, mit känozoischen Lockersedimenten gefüllte Senkungsstruktur des Rheintals sowie die angrenzenden Gebiete mit tertiären Sedimenten.

Prägend ist hierbei der Raum 31, Oberrheingraben mit Mainzer Becken, der hauptsächlich aus quartären Lockergesteinen aufgebaut ist, die gleichzeitig einen der wasserwirtschaftlich wichtigsten Porengrundwasserleiter in Deutschland darstellen. Hydrogeologisch gesehen ist in Raum 31 eine deutliche Nord-Süd-Gliederung erkennbar. Die Gesamtmächtigkeit der Lockersedimente nimmt von Süden nach Norden zu. Ab etwa der Mitte schalten sich zunehmend geringdurchlässige Zwischenhorizonte ein, die allerdings nicht überall flächenhaft ausgeprägt sind. Nördlich von Karlsruhe führen diese zu einer Gliederung in mehrere hydraulisch getrennte Grundwasserstockwerke. Auch lateral gesehen lassen sich unterschiedliche Bereiche ausweisen, die sich in der weiteren Untergliederung in Teilräume widerspiegeln. Neben einer Zunahme der Mächtigkeit zur Grabenmitte hin variieren auch die Durchlässigkeiten, vor allem im Bereich von Schwemmfächern aus den angrenzenden Randgebirgen. Eine etwas spezielle Situation herrscht im Mainzer Becken, wo tertiäre Kalksteine, Sande und Mergel eine Abfolge verschiedener Grundwasserleiter bilden. Dennoch wird das Mainzer Becken dem Raum 31 zugerechnet. Von der Grundwasserdynamik lassen sich vier Großgruppen unterscheiden. Die Messstellen nahe des Rheins sind deutlich von dessen Dynamik beeinflusst. Unterschiede in Nord-Süd-Richtung zeigen sich auch hier, was aber vor allem der Wasserhaltung mit Staustufen

geschuldet ist. Die Messstellen nahe des Grabenrands sind von Randzuflüssen geprägt, wobei sich hier ebenfalls Unterschiede zwischen dem südlichen Bereich (Randzuflüsse aus dem Schwarzwald) und nördlichen Bereich (Kraichgau und Odenwald) zeigen. Die Messstellen auf der Niederterrasse zeigen ebenfalls oft eine ähnliche Dynamik, die jedoch lokal von Vorflutern und Entnahmen überprägt wird. Speziell im Norden (Rhein-Neckar-Raum und Hessisches Ried) ist der Oberrheingraben stark von wasserwirtschaftlichen Eingriffen (hauptsächlich Entnahmen, zum Teil aber auch Wasserhaltung mit Infiltration) geprägt. Die Grundwasserdynamik auf der linksrheinischen Seite verhält sich vor allem Richtung in des Pfälzer Waldes oft etwas anders, was hauptsächlich auf Unterschiede in der Sedimentfüllung zurückzuführen ist. Ein weiterer Grund, vor allem im Bereich der Niederterrasse, dürfte die in Rheinland-Pfalz in den letzten Jahren stärker ausgeprägte Bewässerungslandwirtschaft sein.

Die wasserwirtschaftliche Bedeutung des Großraums 3 spiegelt sich auch in der Messstellendichte wider. Der Großraum weist insgesamt 2.497 Messstellen auf, von denen nach dem Preprocessing und der Eingrenzung des Untersuchungszeitraums Daten aus 1.854 Messstellen verblieben. Diese verteilen sich auf die Bundesländer Baden-Württemberg, Rheinland-Pfalz und Hessen. Die wasserwirtschaftliche Bedeutung, zusammen mit der Tatsache, dass den Bearbeitern des Projekts zu diesem Gebiet zahlreiche Literatur, Erfahrungen aus früheren Untersuchungen sowie Kontakte zu regionalen Experten zur Verfügung standen, was die Einschätzung und Bewertung der Ergebnisse erleichtert, führte zur Wahl von Großraum 3 als Pilotgebiet zur Entwicklung und Optimierung der Methodik. Im Zuge dessen hat sich herausgestellt, dass eine getrennte Betrachtung von Raum 33 (Nordhessisches Tertiär) zu einer deutlichen Verbesserung des Clusterergebnisses im restlichen Gebiet führte, was vermutlich auf die doch deutlich abweichende hydrogeologische Situation zurückzuführen ist. Raum 32 (Untermainsenke) bildet den nordöstlichen Randbereich des Oberrheingrabens, mit gegenüber Raum 31 verminderten Sedimentmächtigkeiten. Aufgrund seiner hydrogeologischen Ähnlichkeit wurde Raum 32 mit Raum 31 zusammengefasst.

Raum 31 und 32

Die für das Clustering zusammengefassten Räume 31 und 32 weisen insgesamt 1.774 Messstellen auf, die in 22 Cluster gruppiert wurden. Das Clustering beruht dabei auf sechs Features (Tabelle 2.5). Von den 22 Clustern wurden 17 Cluster identifiziert, die sich für die Vorhersage von Messstellen anhand ausgewählter Referenzmessstellen eignen. Die restlichen 5 Cluster wurden verworfen, da die Korrelationen der Ganglinien in ihnen untereinander als zu gering eingestuft wurde und deren Ähnlichkeit z.B. auf der Gemeinsamkeit einer schlechten Datenqualität mit Lücken oder großen Sprüngen besteht.

Das größte Cluster (Cluster 1) umfasst dabei 201 Messstellen, dessen zentrale Messstelle im Schnitt mit einem $\overline{R_w}$ -Wert von 0,81 mit allen anderen Messstellen korreliert. Da diese Messstelle von der Datenqualität als ausreichend für die Vorhersage angesehen wird, wurde sie als Referenzmessstelle definiert. Die Messstellen des Clusters zeichnen sich in ihrer Dynamik vor allem durch einen gut ausgeprägten Jahresgang bei gleichzeitig sonst wenig Variabilität und einem leicht verschobenen Jahresmaximum bzw. -minimum aus. Von ihrer Lage her befinden sich die meisten Messstellen auf der Niederterrasse.

Die Messstellen im Clusterzentrum konzentrieren sich auf der rechtsrheinischen Seite etwa auf den Raum um Offenburg. Die Messstellen eher am Clusterrand treten mehr im Süden bei Freiburg und in Rheinland-Pfalz auf. Insgesamt ist das Cluster sehr homogen. Die ersten 50-100 Ganglinien zeigen einen fast identischen Verlauf mit einer mittleren Korrelation zu allen anderen Ganglinien von deutlich $> 0,75$.

Das zweitgrößte Cluster (Cluster 2, 95 Messstellen) verhält sich relativ ähnlich, mit einer etwas geringen RangeRatio und YearlyVariance. Auch von der räumlichen Verteilung finden sich Ähnlichkeiten zu Cluster 1, wobei hier mehr Messstellen Richtung östlichem Grabenrand liegen. Die Homogenität ist insgesamt vergleichbar zu Cluster 1.

Cluster 3 (90 Messstellen) fasst vor allem die Messstellen mit einem sehr gering bis kaum ausgeprägten Jahresgang zusammen, bei gleichzeitig längeren Zeiträumen mit fallenden Grundwasserständen (Feature LongestRecession). Räumlich konzentrieren sich die Messstellen vor allem auf den Bereich nördlich von Karlsruhe und hier vor allem auf die rechtsrheinische Seite. Insgesamt ist das Cluster bis Messstelle 50 sehr homogen, bei einem maximalen $\overline{R_w}$ -Wert von 0,84.

Cluster 4 (103 Messstellen) verhält sich wiederum sehr ähnlich wie Cluster 3, sowohl von der Dynamik, die sich nur noch stärker durch einen fehlenden Jahresgang und auch sonst eine geringe Variabilität auszeichnet, als auch in der räumlichen Lage. Hier konzentrieren sich die Messstellen noch stärker auf den nördlichen Bereich, etwa zwischen Heidelberg und Darmstadt. Hier dürfte die Grundwasserhaltung im Hessischen Ried die Dynamik bestimmen. Ziemlich ähnlich verhält sich zudem noch Cluster 6 (27 Messstellen), bei etwas geringerer Schiefe und leicht verschobenem Jahresmaximum/-minimum, das sich auch in der räumlichen Lage mit Cluster 4 deckt. Die Homogenität beider Cluster ist mit einem maximalen $\overline{R_w}$ -Wert von 0,88 für Cluster 4 und 0,90 für Cluster 6 sehr hoch.

Cluster 5 umfasst 120 Messstellen, die sich mit wenigen Ausnahmen am Clusterrand in Rhein-Nähe konzentrieren, vor allem nördlich von etwa Offenburg an. Sie zeichnen sich durch einen starken Einfluss des Rhein-Wasserstands aus, was sich in einer hohen Variabilität mit meist überprägtem Jahresgang und gleichzeitig kaum längeren Perioden von fallenden Grundwasserständen zeigt. Räumlich fällt weiterhin eine Nord-Süd-Ausrichtung auf, wobei die Messstellen in Clusterzentrum eher im Norden liegen und die Messstellen Richtung Süden immer mehr zum Clusterrand wandern. Insgesamt ist das Cluster sehr homogen (max. $\overline{R_w} = 0,85$). Ein weiteres Cluster mit Rhein-nahen Messstellen ist Cluster 13, wo sich die Messstellen im Zentrum noch mehr auf den nördlichen Bereich konzentrieren und dort auch ein Gebiet nördlich des Neckar-Zuflusses abdecken, das in Cluster 5 ausgespart blieb. Cluster 16 und 19 enthalten Messstellen entlang des südlichen Oberrheins zwischen Basel und Offenburg, die durch die Wasserhaltung mit Staustufen in diesem Bereich eine etwas andere Dynamik aufweisen.

Cluster 7 mit 83 Messstellen ähnelt von der Dynamik Cluster 5, mit jedoch etwas geringer ausgeprägter RangeRatio und YearlyVariance, dafür etwas längeren Perioden fallender Grundwasserstände. Auch hier herrschen Messstellen in der Rheinniederung vor, vor allem auf Höhe von Karlsruhe, mit einem im Schnitt aber etwas größeren Abstand zum Rhein.

Cluster 8 zeichnet sich durch eine insgesamt eher geringe Homogenität aus (maximaler $\overline{R_w}$ -Wert von 0,53). Bei vielen der Ganglinien ist mutmaßlich eine anthropogene Beein-

flussung erkennbar, was sich teils auch auf Grund der räumlichen Lage vermuten lässt. Daher wurde dieses Cluster als Ganzes verworfen.

Es folgen weitere Cluster mit unterschiedlichen Charakteristika, von denen hier noch Cluster 16, 18 und 22 auf Grund ihrer Besonderheit erwähnt werden sollen. Cluster 16 umfasst 12 Messstellen entlang des südlichen Oberrheins, die meisten im Bereich des Kulturwehres Breisach. Sie zeichnen sich durch eine extrem positive Schiefe aus. In der Ganglinie zeigt sich das durch viele Extrem-Ausschläge nach oben. Ganz ähnlich verhält sich Cluster 19, wenn auch etwas weniger stark ausgeprägt. Es dürfte sich dabei um Messstellen im Bereich von Überflutung-Poldern handeln. Cluster 18 und 22 umfassen typische Messstellen mit geringer Datenqualität, charakterisiert durch große Sprünge und eine sehr geringe oder sehr hohe Schiefe, die untereinander nicht korrelieren und für die Vorhersage verworfen werden müssen.

Insgesamt ist das Clusterergebnis von Raum 31 und 32 als sehr gut zu bewerten, was auf Grund der relativ einheitlichen hydrogeologischen Verhältnisse auch zu erwarten war. Die sieben größten Cluster umfassen nach der Bereinigung im Schnitt jeweils mehr als 100 Messstellen mit ähnlicher Dynamik, und es bilden sich nur wenige kleine Cluster aus, abgesehen von denen, die Messstellen mit geringer Datenqualität gruppieren. Insgesamt erscheinen 1.014 Messstellen der ursprünglichen 1.774 Messstellen als zur Vorhersage geeignet, 760 Messstellen mussten verworfen werden.

Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,80. Die Abdeckung des gesamten Raums mit Referenzmessstellen und Messstellen, die auf Grund der Dynamik in den Referenzmessstellen vorhergesagt werden können, ist praktisch flächendeckend, mit wenigen Ausnahmen ganz im Süden und Richtung westlichem Rand von Rheinland-Pfalz.

Raum 33

In Raum 33 stellt sich Situation etwas anders dar. Hydrogeologisch gesehen herrscht im südlichen Teil (Teilraum Vogelsberg) ein mehrschichtiges Grundwasserstockwerkssystem in den tertiären Vulkaniten vor. Im nördlichen Teil (Niederhessische Senke) liegt ein tertiärer Porengrundwasserleiter über einem Kluftgrundwasserleiter aus Mittlerem Buntsandstein. Beide sind hydraulische durch den Oberen Buntsandstein (Röt) getrennt. Die Messstellendichte ist gegenüber Raum 31 und 32 wesentlich geringer, etwa um den Faktor 8. Neben den komplexen hydrogeologischen Verhältnissen dürfte dies die Ursache sein, dass sich die Ganglinien der insgesamt 80 Messstellen untereinander wenig ähneln und sich daher nur schwer gruppieren lassen. Das Clustering beruht auf drei Features (Tabelle 2.5). Die 80 Messstellen teilen sich auf vier Cluster auf, von denen zwei Cluster (Cluster 2 und 3 mit insgesamt 46 Messstellen) auf Grund ihrer Inhomogenität verworfen werden müssen. Cluster 1 mit bereinigt 16 Messstellen zeigt zwar eine vergleichsweise geringere Korrelation der Ganglinien untereinander (max. $\overline{R_w}$ -Wert von 0,68), wurde aber dennoch behalten, ebenso Cluster 4 (6 Messstellen, max. $\overline{R_w}$ -Wert von 0,65). In beiden Clustern sind die Messstellen über weite Teile des Raums verstreut, so dass die Ähnlichkeiten weniger auf eine räumliche Nähe zurückzuführen sein dürften. Vermutlich dürfte es sich um Messstellen mit ähnlichen hydraulischen Bedingungen handeln. Ob sich hier eine sinnvolle Vorhersage realisieren lässt, wird in Kapitel 2 dargelegt.

2.4.3 Großraum 1: Nord- und mitteldeutsches Lockergesteinsgebiet

Mit 6.023 Messstellen, von denen nach dem Preprocessing und der Eingrenzung des Untersuchungszeitraums noch 5.291 verblieben, ist Großraum 1 mit Abstand der mit den meisten Messstellen, gleichzeitig aber auch räumlich der größte. Er ist dabei von einer sehr unterschiedlich verteilten Messstellendichte geprägt, mit einer sehr hohen Dichte beispielsweise in den Stadt-Staaten und einer eher geringen Dichte im Bundesland Mecklenburg-Vorpommern. Großraum 1 umfasst den größten Teil des Norddeutschen Tieflands, und seine südöstlichsten Anteile reichen bis nach Sachsen. Gekennzeichnet ist er durch das Auftreten mächtiger quartärer und tertiärer Lockergesteinsfolgen, die hydrogeologisch ergiebige Porengrundwasserleiter bilden und in weiten Teilen eine große Bedeutung für die Wasserversorgung haben.

Großraum 1 wird in der Hydrogeologischen Raumgliederung in sieben zum Teil jedoch recht kleine Räume untergliedert. Größere Flächen werden vor allem von Raum 14 (Norddeutsches Jungpleistozän) und 15 (Nord- und mitteldeutsches Mittelpleistozän) eingenommen. Raum 13 (Niederungen im nord- und mitteldeutschen Lockergesteinsgebiet) stellen ehemalige Urstromtäler und Entwässerungsrinnen dar, entlang denen noch heute die Flussniederungen von Elbe, Weser, Ems und weiteren kleineren Flüssen verlaufen. Er ist zudem teilweise durch Moorbildungen gekennzeichnet. Die Räume 11 (Nordseeinseln und Watten) und 12 (Nordseemarschen), sowie 16 (Altmoränengeest) und 17 (Lausitzer Känozoikum) sind von ihrer räumlichen Ausdehnung her eher von untergeordneter Bedeutung.

Der flächenhaft größte Raum 14, dessen westliche und südwestliche Verbreitungsgrenze etwa die maximale Ausdehnung der Weichsel-Vereisung nachzeichnet, ist durch glazialtektonisch zum Teil intensiv gestauchte Moränenzüge mit kleinräumig inhomogenem Schichtaufbau gekennzeichnet. Dies führt hydrogeologisch gesehen zu einer oft komplexen Stockwerksgliederung aus grundwasserführenden glaziofluviatilen Sanden und bindigen Trennschichten der Grundmoränen. Raum 15 bildet den zweitgrößten Raum, der auch als Geest bezeichnet wird. Es handelt sich dabei um eine glaziale Aufschüttungslandschaft aus an der Oberfläche anstehenden Geschiebemergeln und glaziofluviatilen Sanden der Saale- und Elstervereisung. In Untergrund sind Sande und Tone des Tertiärs verbreitet. Auch in Raum 15 herrscht weitestgehend eine komplexe Stockwerksgliederung aus Grundwasserleitern und -geringleitern vor, die zudem lokal durch Stauchendmoränen mit stark gestörter Lagerung überprägt ist.

Da die beiden größten Räume 14 und 15, sowie die Räume 16 und 17 hydrogeologisch gesehen sehr ähnlich sind und die verbleibenden Räume eine nur geringe Fläche einnehmen, die im Falle von Raum 13 zudem räumlich stark zergliedert ist, wurde für das Clustering der Großraum als Einheit verwendet.

Beim Clustering ergeben sich erwartungsgemäß auf Grund des großen Gebiets mit zahlreichen Messstellen sehr viele Cluster, insgesamt 237. Davon sind die ersten zwölf größere Cluster mit mehr als 100 Messstellen, die vor der Bereinigung fast 2.000 und damit mehr als ein Drittel aller Messstellen vereinen. Cluster 13 bis 30 umfassen jeweils 100 bis 39 Messstellen und bilden damit mittlere Cluster. Ab Cluster 119 umfassen die Cluster jeweils weniger als 10 Messstellen. Diese Aufteilung spiegelt auch gut die hydrogeologische Situation wieder. Prinzipiell herrschen in weiten Teilen des Großraums zwar ähnliche hydrogeologische Verhältnisse mit mächtigen quartären und tertiären Lockergesteinsfol-

gen vor, was sich in der Bildung einiger großer Cluster mit vielen Messstellen ähnlicher Dynamik zeigt. Dagegen gibt es lokal aber oft auch komplexe, inhomogene Bereiche, vor allem dort, wo glaziotektonische Überprägungen zu stark gestörten Lagerungsverhältnissen geführt haben. Dies zeigt sich wiederum in dem Auftreten zahlreicher sehr kleiner Cluster, die jedoch oft in sich recht homogen sind. Für ein sinnvolles Clustering waren sieben Features nötig (Tabelle 2.5). Deren Verteilung zeigt im Gegensatz zu den meisten anderen Großräumen eine starke Variabilität zwischen den Clustern, was dafür spricht, dass prinzipiell sehr unterschiedliche Typen von Ganglinien vertreten sind.

Bei den großen Clustern zeichnet sich Cluster 1 mit nach der Bereinigung 99 verbleibenden Messstellen durch insgesamt geringe Variabilität und kaum ausgeprägten Jahresgang aus. Die Messstellen konzentrieren sich dabei hauptsächlich auf den östlichen Teil des Großraums. Innerhalb des Clusters ist die Homogenität mit einer maximalen Korrelation von 0,84 gut, allerdings wurde dieses Cluster, das vor der Kappung 320 Messstellen enthielt, stark bereinigt, da die Messstellen am Clusterrand sehr stark inhomogen wurden. Cluster 2 grenzt sich gegenüber Cluster 1 durch einen deutlich ausgeprägten Jahresgang ab. Obwohl hier die Messstellen über den gesamten Großraum verteilt sind, ist die Homogenität des Clusters mit einem maximalen $\overline{R_w}$ -Wert von 0,80 relativ hoch, vor allem wenn man die Anzahl der Messstellen (218) berücksichtigt.

Cluster 3, bei dem viele der 213 Messstellen in Berlin liegen, zeichnet sich durch eine geringe Homogenität aus, was sich in einer negativen Schiefe und leicht erhöhter Rauigkeit äußert. Viele der Messstellen dürften stark anthropogen beeinflusst sein. Auf Grund der geringen Homogenität wurde dieses Cluster verworfen. Cluster 4 mit 188 Messstellen zeichnet sich ähnlich wie Cluster 2 durch einen deutlich ausgeprägten Jahresgang aus, zeigt darüber hinaus aber auch noch eine höhere Rauigkeit. Die Messstellen konzentrieren sich hauptsächlich auf den westlichen und nördlichen Bereich des Großraums. Cluster 5, 13 und 14 umfassen ebenso wie Cluster 3 Messstellen mit starken Beeinflussungen oder geringer Datenqualität und wurden ebenfalls verworfen.

Die folgenden Cluster mit mittlerer Größe weisen zum Teil hohe, zum Teil aber auch geringe Homogenitäten auf. Die mit hohen Homogenitäten sind räumlich oft konzentriert auf ein Gebiet, zum Teil aber auch erstaunlich weit verstreut. Besonders erwähnenswert erscheint Cluster 39 mit 30 Messstellen, bis auf zwei im Bereich der Lüneburger Heide gelegen. Alle weisen ein sehr spezielles Verhalten auf, jährliche wiederkehrende starke Abfälle mit eher kontinuierlichen Wiederanstiegen, verstärkt ab dem Jahr 2003.

Insgesamt ist mit einem mittleren $\overline{R_w}_{max}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung von 0,74 das Clusterergebnis nicht ganz so gut wie in Raum 31/32, was aber auf Grund der teils komplexeren, glazial überprägten geologischen Verhältnisse nicht verwundert. Obwohl von den insgesamt 237 Clustern 171 verworfen werden mussten, teils wegen zu großer Inhomogenität, teils wegen starker anthropogener Beeinflussung oder schlechter Datenqualität, teils aber auch nur wegen einer zu geringen Messstellenanzahl, so wird mit den Messstellen der verbleibenden 66 Cluster (2.480 verbleibende Messstellen von ursprünglich 5.291) dennoch eine weitgehend gute Abdeckung des Großraums für die spätere Vorhersage erzielt. Ausnahme bilden lediglich der Raum 14, was entweder an der hier komplexeren hydrogeologischen Situation als in den anderen Räumen liegen könnte, oder aber auch an der insgesamt eher geringen Messstellendichte im Bundesland Mecklenburg-Vorpommern, welches sich weitgehend mit Raum 14 deckt, sowie der Raum 17.

2.4.4 Großraum 2: Rheinisch-Westfälisches Tiefland

Das Rheinisch-Westfälische Tiefland ist einer der kleineren Räume, weist jedoch mit 1.027 Messstellen (888 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) eine recht hohe Messstellendichte auf. Großraum 2 ist besonders durch sehr mächtige tertiäre zyklische Lockergesteinsfolgen aus Sanden, Tonen und Braunkohleflözen gekennzeichnet, die mehrere Grundwasserstockwerke ausbilden. Er wird in drei Räume untergliedert. Raum 21 (Sandmünsterland) umfasst ein Gebiet mit mächtigen fluviatilen und glazialen Ablagerungen, entlang der Ems und oberen Lippe, die weitgehend zusammenhängende Porengrundwasserleiter von 10-40 m Mächtigkeit bilden und nur lokal durch tonige Moränensedimente in Stockwerke gegliedert sind. Raum 21 ist geologisch gesehen in das Münsterländer Kreidebecken (Raum 22) eingebettet, eine mit kreidezeitlichen Ablagerungen gefüllte Muldenstruktur, welche im Norden durch den Teutoburger Wald, im Osten durch das Eggegebirge und im Süden durch das Rheinische Schiefergebirge begrenzt wird. Hydrogeologisch und überregional bedeutsam ist in Raum 22 vor allem der bis zu 300 Meter mächtige Porengrundwasserleiter aus den lockeren bis halb verfestigten Fein- und Mittelsanden (sog. Halterner Sande). Entlang des Rheins sowie westlich davon erstreckt sich die Niederrheinische Tieflandsbucht (Raum 23). Bei dieser handelt es sich um einen in Bruchschollen zerlegten Senkungsraum mit Sedimentmächtigkeiten von bis zu 1.200 Metern. Hydrogeologisch gesehen ist die Niederrheinische Tieflandsbucht aus einer Wechselfolge von Grundwasserleitern und -geringleitern aufgebaut. Erstere enthalten ergiebige Grundwasservorkommen von hoher wasserwirtschaftlicher Bedeutung. Wichtig zu erwähnen sind hier die Flächen des rheinischen Braunkohletagebaus, wo die natürlichen Grundwasserverhältnisse durch großräumige Absenkungen von teils weit über 100 Metern stark gestört sind. Auch im Bereich stillgelegter Tagebaue, die zur Rekultivierung mit Abraum und Kraftwerksaschen verfüllt wurden, bleiben die Verhältnisse nachhaltig gestört, da die Auffüllungen eine deutlich geringere Durchlässigkeit als im ursprünglichen Zustand aufweisen.

Beim Clustering wurden 17 Cluster gebildet, basierend auf einer Auswahl von 5 Features (Tabelle 2.5). Von den 17 Clustern mussten lediglich 4 ganz verworfen werden. Das größte Cluster (Cluster 1) weist 292 Messstellen auf und ist dafür von der Dynamik her erstaunlich homogen, mit einem maximalen $\overline{R_w}$ -Wert von 0,82. Die räumliche Verteilung der Messstellen erstreckt sich über Raum 21 und 22 ohne klare Tendenz. Die Messstellen weisen einen ausgeprägten Jahresgang auf. Cluster 2 ist von der Anzahl mit 81 Messstellen deutlich kleiner. Die meisten Messstellen liegen entlang des Rheins, mit einer Konzentration ganz im Süden des Großraums, und sind von dessen Dynamik beeinflusst. Der maximalen $\overline{R_w}$ -Wert von 0,88 zeigt eine hohe Homogenität des Clusters an.

Cluster 3 bis 5 sind sich in ihrer Dynamik untereinander ähnlich, mit räumlich leicht anders gelagerten Schwerpunkten. Cluster 6 mit 58 Messstellen wurde verworfen, da hier viele der Ganglinien Sprünge, Datenlücken oder allgemein einen wenig plausiblen Verlauf zeigten. Räumlich liegen hier viele der Messstellen im Rheinischen Braunkohletagebauegebiet und sind mutmaßlich von der dortigen Wasserhaltung stark überprägt. Von Cluster 7 bis 11 folgen noch weitere mittelgroße Cluster mit zwischen 10 und 44 Messstellen, die alle eine unterschiedliche, jedoch in sich homogene Dynamik aufweisen (maximale $\overline{R_w}$ -Werte um 0,70-0,75). Ab Cluster 12 nimmt die Messstellenzahl auf unter

10 ab, wofür eine Voraussage mittels Referenzmessstelle wahrscheinlich vom Aufwand her nicht mehr sinnvoll ist.

Insgesamt erscheinen 580 der ursprünglichen 888 Messstellen als zur Vorhersage geeignet, 308 Messstellen mussten verworfen werden. Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,77. Räumlich gesehen werden vor allem die Räume 21 und 22 sehr gut von mutmaßlich vorhersagbaren Messstellen abgedeckt, sowie das Gebiet entlang des Rheins. Etwas anders ist die Situation im Raum 23. Hier dürfte eine flächenhafte Vorhersage schwierig werden. Gründe hierfür liegen in der geringen Messstellendichte, der komplexen geologischen Situation und der zumindest bereichsweise starken anthropogenen Beeinflussung durch die mit dem Tagebau verbundenen großräumigen Grundwasserhaltungsmaßnahmen.

2.4.5 Großraum 4: Alpenvorland

Das Alpenvorland ist ein aktives Vorlandbecken südlich der Donau, das auch als Süddeutsches Molassebecken bezeichnet wird. Es ist mit Sedimenten unterschiedlicher Korngrößenverteilung aus den umliegenden Gebirgen verfüllt, die weitgehend ergiebige Grundwasserleiter bilden, in welche häufig Grundwassergeringleiter eingeschaltet sind. Unter dem Molassebecken stellt der nach Süden abtauchende Oberjura (Malm) einen auch wasserwirtschaftlich bedeutenden tiefen Karstgrundwasserleiter dar. Dieser führt mit zunehmender Tiefenlage nach Süden hin Mineral- und Thermalwasser. Der Großraum 4 bildet zugleich den Raum 41 (Süddeutsches Molassebecken), der wiederum in vier Teilräume untergliedert ist. Obwohl diese zum Teil hydrogeologisch sehr unterschiedlich sind (von quartären Talschottern bis glazial stark überprägten, komplexen Moränensedimenten) wurde im Sinne eines einheitlichen Vorgehens auf eine Untergliederung in Teilräume verzichtet und somit der Großraum 4 bzw. Raum 41 als Ganzes betrachtet.

Insgesamt lagen Daten aus 650 Messstellen (624 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) vor, die sich recht gleichmäßig über das Gebiet verteilen. Insgesamt zeigt das Alpenvorland im Vergleich zu anderen Großräumen jedoch eine der niedrigsten Messstellendichten.

Von den 11 gebildeten Clustern wurden 3 auf Grund zu geringer Homogenität verworfen. Die Messstellen in den verbleibenden Clustern liegen fast ausschließlich entlang von Flüssen und sind in unterschiedlichem Maße von deren Dynamik beeinflusst. Das größte Cluster 1 mit 75 verbleibenden Messstellen konzentriert sich räumlich hauptsächlich auf den westlichen Teil des Großraums, im Bereich der Schotterplatten der südlichen Donauzuflüsse.

Cluster 2 mit 120 Messstellen, die über den gesamten mittleren und östlichen Teil des Großraums räumlich verteilt liegen, wurde auf Grund geringer Homogenität aussortiert. Hier finden sich vor allem Ganglinien mit Trends, Sprüngen, über den Zeitraum wechselnder Dynamik oder unplausiblem Verlauf.

Cluster 3 mit 66 Messstellen verhält sich von der Dynamik ähnlich wie Cluster 1, mit Messstellen vor allem entlang der Donau und von Günz und Mindel, ebenso Cluster 4 (47 Messstellen) mit einer starken räumlichen Konzentration von Messstellen im Zentrum auf ein kleines Gebiet entlang der Iller. Beide zeigen gegenüber Cluster 1 eine etwas höher ausgeprägte Variabilität und höhere Schiefe. Letzteres könnte auf Hochwasser-Einflüsse

hindeuten. Die Messstellenzahlen der weiteren verbliebenen Cluster variieren zwischen 7 und 36, wobei die Messstellen des größten Clusters (Cluster 6) räumlich am weitesten verstreut liegen, gleichzeitig weist das Cluster die geringste Homogenität auf ($\overline{R_{w\ max}}$ -Wert = 0,73). Die Cluster 7 bis 10 weisen alle nur wenige, räumlich konzentriert liegende Messstellen auf, bei gleichzeitig hoher Homogenität ($\overline{R_{w\ max}}$ -Werte zwischen 0,80 und 0,85).

Insgesamt verbleiben nur 273 der ursprünglich 624 Messstellen zur Vorhersage, und mit 351 Messstellen mussten mehr als die Hälfte verworfen werden. Obwohl der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung mit 0,80 recht hoch ist, muss man bemerken, dass viele Messstellen des Großraums, vor allem solche, die nicht an Flüssen gelegen sind, auf Grund ihrer unspezifischen Dynamik verworfen wurden. Neben der komplexen, glazial stark überprägten hydrogeologischen Situation dürfte hier auch anthropogene Beeinflussung eine Rolle spielen. Verbunden mit der schon vorher geringen Messstellendichte führt dies zu einer unbefriedigenden Abdeckung mit vorhersagbaren Messstellen vor allem im mittleren und östlichen Bereich des Großraums.

2.4.6 Großraum 5: Mitteldeutsches Bruchschollenland

Das Mitteldeutsche Bruchschollenland, auch als Thüringische und Subherzynische Senken bezeichnet, ist durch das flächenhafte Vorkommen tektonisch beanspruchter mesozoischer Einheiten gekennzeichnet, die mäßig bis teilweise sehr ergiebige Kluft- bzw. Kluft-/Poren- und Kluft-/Karstgrundwasserleiter bilden. Für die Wasserwirtschaft bedeutend sind vor allem die Grundwasserleiter der Trias im Zentrum des Thüringer Beckens sowie die des Perms am Beckenrand und in Osthessen.

Großraum 5 wird in vier Räume untergliedert. Der im Nordwesten liegende Raum 51 (Nordwestdeutsches Bergland) ist durch mesozoische Sedimentgesteine, die zum Teil gefaltet und an zahlreichen Störungen zerbrochen sind, gekennzeichnet, was zu einem uneinheitlichen und komplizierten hydrogeologischen Bau führt. Im Raum 52 (Mitteldeutscher Buntsandstein) überwiegen Schichten des Buntsandsteins und des Zechsteins, welche relativ ergiebige Kluftgrundwasserleiter ausbilden. Tektonisch gesehen sind Grabenstrukturen und eine intensive bruchtektonische Überprägung charakteristisch, was zu den erhöhten Wasserwegsamkeiten führt. Räumlich gesehen ist Raum 52 nicht zusammenhängend sondern wird durch Ausläufer von Raum 33 getrennt. Raum 53 (Subherzynische Senke) liegt im Nordosten des Großraums umgeben von Raum 83, dem Mitteldeutschen Grundgebirge. In der Senkenstruktur wurden mesozoischen Gesteine abgelagert, die jedoch durch das Auftreten von tertiären Randsenken und teilweise Salzsätteln in verschiedene Schollen getrennt sind. Raum 54 (Thüringische Senke) nimmt den größten Flächenanteil des Großraums 5 ein. In seinen Muldenstrukturen treten Gesteine des Zechsteins und der Trias auf, die hauptsächlich Kluft- (/Poren-)grundwasserleiter ausbilden. Insgesamt werden dem Großraum auch zahlreiche Messstellen zugerechnet, die in oberflächennahen quartären Talfüllungen und damit Porengrundwasserleitern verfiltert sind und sich damit von der Grundwasserdynamik anders verhalten.

Mit 1.376 Messstellen (1.175 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) ist die Messstellendichte vergleichsweise hoch, wovon nur Raum 51 auszunehmen ist. Aufgrund des hydrogeologisch ähnlichen (komplexen) Aufbaus der Räume

erschien eine Unterteilung des Großraums für das Clustering nicht zweckmäßig, weshalb das Mitteldeutsche Bruchschollenland als Ganzes betrachtet wurde.

Das Clustering ergab eine Gruppierung in insgesamt 43 Cluster auf Grund von 7 verwendeten Features (Tabelle 2.5). Von den 43 Clustern mussten allerdings 28 verworfen werden, so dass nur 8 für die Auswahl von Referenzmessstellen geeignete Cluster verblieben. Die verworfenen Cluster zeigen fast alle eine sehr geringe Variabilität, zum Teil längerfristige Anstiege über mehrere Jahre hinweg (Cluster 5), keinen deutlichen Jahresgang oder insgesamt ein eher unspezifisches Verhalten. Dies zeigt auch die vergleichsweise große Anzahl der Cluster, wovon etwa ein Drittel weniger als 10 Messstellen aufweisen. Auch wenn die Korrelation der Ganglinien untereinander zum Teil mit ca. 0,5-0,6 noch akzeptabel war, so dürfte dies in den meisten Fällen der geringen Variabilität geschuldet sein, und weniger einem Kausalzusammenhang, woraufhin entschieden wurde, diese Cluster dennoch zu verwerfen. Zurückzuführen ist das Verhalten auf die insgesamt komplexe und uneinheitliche Geologie, vor allem in Raum 51 und teilweise auch in Raum 54.

Das größte Cluster (Cluster 1) umfasst 91 verbleibende Messstellen, vor allem in Raum 54 gelegen, viele davon entlang von Flüssen, mutmaßlich in oberflächennahen Porengrundwasserleitern verfiltert und von der Dynamik der Oberflächengewässer beeinflusst. Hierauf deutet z.B. die erhöhte positive Schiefe hin. Mit einem $\overline{R_{w\ max}}$ -Wert von 0,79 ist es eines der homogeneren Cluster des Großraums. Cluster 2 (63 Messstellen) zeigt bei einer etwas geringeren Homogenität ($\overline{R_{w\ max}}$ -Wert = 0,76) eine leichte Konzentration von Messstellen im Raum 52, bei einer etwas höheren Varianz und Periodizität der Messstellen bei gleichzeitig etwas geringerer positiver Schiefe.

Cluster 3 (74 Messstellen) hat mit einem $\overline{R_{w\ max}}$ -Wert von 0,79 von der Dynamik her eine gute Homogenität, obwohl die Messstellen räumlich über den gesamten Großraum verteilt sind. Von Cluster 1 grenzt es sich vor allem durch ein leicht verschobenes Jahresmaximum/-minimum ab. Cluster 4 und 5 mit 64 bzw. 54 Messstellen sind die beiden größten der verworfenen Cluster. Räumlich konzentrieren sich die Messstellen vor allem auf den östlichen Teil von Raum 54, der durch einen komplexen Aufbau gekennzeichnet ist. Mit einem $\overline{R_{w\ max}}$ -Wert von 0,35 bzw. 0,59 zeigen die Ganglinien nur eine geringe bis mäßige Korrelation untereinander. Die Ganglinien sind durch wenig Variabilität, einen fehlenden Jahresgang und zum Teil kürzere und längere Zeiträume mit Anstiegen oder Abfällen, welche sich aber auf einzelne oder wenige Messstellen beschränken bzw. zeitlich nicht korrelieren, gekennzeichnet.

Es folgen weitere Cluster, aller mit leicht unterschiedlicher, jedoch grundsätzlich vergleichbarer Dynamik, was auf ähnliche hydrogeologische Verhältnisse in Kluftgrundwasserleitern hindeutet. Räumlich gesehen konzentrieren sich die meisten Cluster dabei auf Raum 54. Die Homogenität variiert relativ stark, mit $\overline{R_{w\ max}}$ -Werten zwischen 0,62 und 0,80.

Der mittlere $\overline{R_{w\ max}}$ -Wert des Großraums über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,74 und bewegt sich damit in einem akzeptablen Bereich für eine spätere Vorhersage. Die Anzahl der vorhersagbaren Messstellen beträgt 441 von ursprünglich 1.175, womit fast zwei Drittel (734 Messstellen) verworfen werden mussten. Von der Abdeckung her sind die Räume 52 (vor allem entlang der Weser), 53 und 54 zufriedenstellend, lediglich der westliche Bereich von Raum 51 und nördliche Bereich von Raum 52 erscheint schwierig.

2.4.7 Großraum 6: West- und süddeutsches Schichtstufen- und Bruchschollenland

Das West- und süddeutsche Schichtstufen- und Bruchschollenland erstreckt sich über weite Teile Baden-Württembergs und Bayerns, im Norden bis nach Hessen. Weiterhin werden dem Großraum der Pfälzer Wald und Teile des Saarlands zugerechnet. Geologisch geprägt ist es durch mesozoische Sedimentgesteine, die Abfolgen von Kluft- bzw. Kluft-/Poren- und Kluft-/Karstgrundwasserleitern ausbilden. Der Großraum ist auf Grund seiner Heterogenität in 6 Räume untergliedert. Raum 61 (Südwestdeutsche Trias) wird dabei von den räumlich abgetrennten Gebieten im südwestlichen Bereich von Rheinland-Pfalz und dem Saarland gebildet. Geologisch gesehen werden diese als Pfälzer Mulde und Trierer Bucht angesprochen und durch mächtige Ablagerungen des Buntsandsteins geprägt, die ein Kluft- bzw. Kluft-/Porenporengrundwasserleitersystem darstellen. Weiterhin werden dem Raum der räumlich angrenzende Südwestdeutsche Muschelkalk und Keuper zugerechnet. Raum 62 (Süddeutscher Buntsandstein und Muschelkalk) umfasst die im Bereich des Schwarzwaldes und Odenwalds, Spessarts, Rhönvorlands und die in Franken ausstreichenden gleichnamigen Trias-Abfolgen, die hydrogeologisch Kluft- und Kluft-/Karstgrundwasserleiter bilden. Raum 63 (Süddeutscher Keuper und Albvorland) umfasst die Ausstrichgebiete des Keupers und des von Lias und Dogger aufgebauten Albvorlands. Hier herrschen Sand- und Tonsteine vor, die Kluft- und Kluft-/Porengrundwasserleiter ausbilden. Die Schwäbische und Fränkische Alb (Raum 65) sind durch mächtige Ablagerungen des Oberjura (Malm) geprägt, welcher einen der wichtigsten und ergiebigsten Karstgrundwasserleiter in Deutschland darstellt. Sie werden räumlich gesehen durch Raum 65 (Nördlinger Ries) getrennt, einem durch den Ries-Impakt kompliziert aufgebauten Bereich mit hauptsächlich Trümmernmassen und feinkörnigen Seesedimenten von nur untergeordneter Bedeutung. Den nordöstlichsten Streifen des Großraums bildet Raum 66 (Thüringisch-Fränkisches Bruchschollenland), ein von mesozoischen Gesteinseinheiten mit starker tektonischer Beanspruchung und komplexem hydrogeologischem Bau gekennzeichnetes Gebiet.

Insgesamt umfasst der flächig zweitgrößte Großraum mit 662 (567 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) nur vergleichsweise wenige Messstelle, die zudem noch sehr ungleich verteilt sind. Eine getrennte Betrachtung der Räume wurde trotz der sehr unterschiedlichen hydrogeologischen Verhältnisse verworfen, da auf die Räume sonst zum Teil nur sehr wenige Messstellen entfallen, was ein sinnvolles Clustering dieser Räume unmöglich macht.

Insgesamt wurden die Messstellen in 27 Cluster gruppiert, davon 3 größere Cluster mit mehr als 50 Messstellen, 7 mittlere mit 20-50 Messstellen und 17 Kleincluster. Das Clustering erfolgte auf Basis von 7 Features (Tabelle 2.5).

Cluster 1 mit 59 Messstellen zeichnet sich von der Dynamik her durch eine hohe Variabilität und Periodizität bei gleichzeitig stark erhöhter positiver Schiefe aus. Die Messstellen verteilen sich über den gesamten Großraum, mit leichter Tendenz des Clusterzentrums nach Osten hin. Einige der Messstellen liegen nahe an Flüssen (Donau, Main, Regnitz, Naab). Cluster 2 mit 58 Messstellen fasst Ganglinien mit insgesamt geringer Variabilität, Trends Datenlücken und allgemein eher ungewöhnlichem Verlauf zusammen und wurde auf Grund der geringen Homogenität ($\overline{R_{wmax}} = 0,44$) verworfen.

Die 37 Messstellen in Cluster 3 konzentrieren sich räumlich auf das Gebiet im Norden, entlang des Mains, wobei die meisten nicht direkt in Flussnähe liegen. Ergänzt wird das Cluster durch 5 Messstellen am Clusterrand, gelegen im räumlich abgetrennten Raum 61 im Bereich des Pfälzer Waldes. Von der Dynamik zeigt das Cluster Ähnlichkeit zu Cluster 1, bei leicht geringerer Variabilität und erhöhten Zeiten mit kontinuierlich fallenden Grundwasserständen.

Cluster 4 mit 45 Messstellen zeichnet sich bei allen Features durch ein mittleres Verhalten ohne besondere Auffälligkeiten aus. Da die Ganglinien nur eine mittlere Homogenität aufwiesen ($\overline{R_{w\ max}} = 0,55$), wurde das Cluster verworfen.

Cluster 5 (40 Messstellen) zeichnet sich vor allem durch eine hohe Variabilität aus. Obwohl die Messstellen über weite Teile des Großraum verstreut liegen, weisen sie untereinander eine vergleichsweise hohe Homogenität mit einem $\overline{R_{w\ max}}$ -Wert von 0,77 auf. Cluster 6 (25 Messstellen) ist mit einem $\overline{R_{w\ max}}$ -Wert von 0,78 ebenfalls recht homogen, wobei sich hier die Messstellen auf ein kleines Gebiet etwa nordöstlich von Ulm (im Bereich der Ostalb) konzentrieren. Sie weisen mit die höchste Variabilität aller Cluster des Großraums auf, was neben der Lage deutlich auf Messstellen im Karst hinweist.

In Großraum 6 wurden 325 der 567 Messstellen verworfen, 242 verbleiben für eine Vorhersage. Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,73, bei allerdings sehr geringer Abdeckung mit vorhersagbaren Messstellen. Vorhersagen erscheinen daher ausgenommen sehr kleiner Gebiete mit etwas erhöhter Messstellendichte nur rein punktuell möglich. Die Ursache hierfür liegt hauptsächlich in der initial schon geringen Messstellendichte, die sich auf das Clusterergebnis durchpaust, und weniger z.B. an einer komplexen Dynamik auf Grund der vorherrschenden Kluft-/Karstgrundwasserleiter.

2.4.8 Großraum 7: Alpen

Der Großraum 7 (Alpen) umfasst lediglich einen sehr schmalen Streifen am Südrand von Bayern. Geologisch gesehen wird der Großraum 7 von den Nördlichen Kalkalpen, Helvetikum, Flysch und Faltenmolasse aufgebaut. Hier herrschen Kluft- und Kluft-/Karstgrundwasserleiter vor. Weiterhin werden auch Lockergesteinsgrundwasserleiter in Tallagen angetroffen.

Mit nur 16 Messstellen nach Preprocessing und Eingrenzung des Untersuchungszeitraums war in diesem Großraum kaum ein sinnvolles Clustering möglich. Der Konsistenz halber wurde dennoch ein Versuch gemacht. Es ergaben sich hierbei 7 Cluster, davon 2 Cluster mit 3 Messstellen, der Rest mit je 2 Messstellen. Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,42. Da die Messstellen in den gebildeten Clustern von der Dynamik her ähnlich sind, funktionierte hier die Cluster-Methodik zwar prinzipiell auch, dennoch wies das homogenste Cluster nur einen $\overline{R_{w\ max}}$ -Wert von 0,45 auf. Für die Vorhersage eignet sich letztendlich keines der Cluster, womit eine Ausweisung von Referenzmessstellen (im Sinne der Referenz) wenig Sinn ergibt. Eine Vorhersage der Messstellen könnte dennoch versucht werden, da die Dynamik und Datenqualität der Messstellen dafür ausreichend erscheint.

2.4.9 Großraum 8: West- und mitteldeutsches Grundgebirge

Der Großraum 8, West- und mitteldeutsches Grundgebirge, kennzeichnet das Verbreitungsgebiet paläozoischer sedimentärer und gering metamorpher Gesteine des Rheinischen Schiefergebirges sowie Harz- und Halle-Wittenberg-Scholle. Weiterhin werden dem Großraum die Vulkanite der Eifel und des Westerwalds sowie fluviatile Ablagerungen von Fließgewässern hinzugerechnet. Den größten Bereich nimmt dabei das Rheinische Schiefergebirge (Raum 81) ein, dessen paläozoische Schichtenfolge intensiv gefaltet und schwach metamorph überprägt ist und zu dem die Vulkanite von Eifel und Westerwald zählen. Hydrogeologisch gesehen herrschen Kluftgrundwasserleiter mit geringen bis mittleren Durchlässigkeiten vor. Weitere Räume des Großraums sind das Saar-Nahe-Becken (Raum 82), mit seinen bis zu 8 km mächtigen permokarbonischen Sedimentgesteinen von eher geringer Durchlässigkeit, wo wasserwirtschaftlich bedeutende Grundwasservorkommen hauptsächlich an Störungszonen gebunden sind, sowie das räumlich getrennt liegende Mitteldeutsche Grundgebirge (Raum 83). Letzteres dient als Sammelraum für weiter im Osten zu Tage tretende größere Gebiete paläozoischer Gesteine, vor allem im Harz, die teilweise mäßig bis gut leitende Kluftgrundwasserleiter bilden, insgesamt jedoch sehr heterogen aufgebaut und wasserwirtschaftlich nur lokal von Bedeutung sind. Kleinere Vorkommen paläozoischer Gesteine, z.B. im Teutoburger Wald, werden auf Grund ihrer geringen räumlichen Ausdehnung nicht dem Großraum 8 zugerechnet.

Wie in allen anderen Festgesteins-Gebieten werden auch in Großraum 8 die Porengrundwasserleiter entlang der Fließgewässer (hier vor allem Mosel, Rhein und Ahr) dem Großraum zugerechnet, die zum Teil aus mehreren Zehner Metern mächtigen Ablagerung mit hoher Durchlässigkeit aufgebaut sind und regional wasserwirtschaftliche Bedeutung haben.

Insgesamt weist Großraum 8 mit 730 Messstellen (436 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) eine mittlere bis geringe und zudem recht ungleich verteilte Messstellendichte auf, die sich vor allem entlang der Flüsse Rhein, Mosel und Ruhr konzentriert. Leider fallen in diesem Großraum besonders viele Messstellen durch die Eingrenzung des Untersuchungszeitraums weg, wodurch sich die ohnehin schon geringe Messstellendichte weiter reduziert. Grund hierfür ist die vergleichbar geringe Datenqualität, mit Datenlücken von mehr als 25 Prozent oder einer Messdauer von kürzer als 6 Jahren.

Die verbliebenen 436 Messstellen wurden auf Grund von nur 3 Features (Tabelle 2.5) 31 Clustern zugeordnet, von denen allerdings nur 12 eine ausreichend gute Homogenität aufwiesen und 19 auf Grund zu geringer Homogenität verworfen werden mussten. Nach der Kappung der verbleibenden Cluster bleiben nur 185 Messstellen übrig, die für eine Vorhersage geeignet erscheinen, 251 wurden verworfen.

Im größten Cluster (Cluster 1) werden 35 Messstellen gruppiert, die bis auf 5 am Clusterrand alle in einem sehr kleinen Gebiet direkt am Rhein, etwas nördlich von Koblenz liegen, was zu einer hohen Homogenität mit einem $\overline{R_{w\ max}}$ -Wert von 0,87 führt.

Cluster 2 mit 20 Messstellen, die sich vor allem auf den räumlich vom restlichen Großraum abgetrennten Raum 83, und hier insbesondere auf den ganz östlichen Bereich nördlich von Halle beschränken, zeichnet sich gegenüber allen anderen Clustern des Großraums vor allem durch seine geringe Periodizität aus. Das Cluster ist in sich sehr homogen, mit einem $\overline{R_{w\ max}}$ -Wert von 0,90.

Cluster 3 mit 31 Messstellen zeigt eine zufriedenstellende Homogenität von $\overline{R_{w\ max}} = 0,69$. Die Messstellen liegen hauptsächlich über den Raum 81 verstreut, mehrheitlich rechtsrheinisch. Neben einer leicht erhöhten Periodizität zeigt Cluster 3 keine besonderen Auffälligkeiten. Cluster 4 umfasst 16 Messstellen, denen hauptsächlich die erhöhte positive Schiefe gemein ist, was auf einen Hochwassereinfluss hindeutet. Etwas weniger ausgeprägt zeigen dies auch Cluster 6 und 8 mit 10 bzw. 16 Messstellen. Cluster 11 mit 16 Messstellen, die fast alle im Nordosten von Raum 81, getrennt in zwei Hauptgruppen liegen, fällt gegenüber allen anderen durch eine stark erhöhte Streuung (Feature SEM) auf. Die eine der beiden Gruppen liegt dabei südlich der Ruhr, etwa im Bereich von Iserlohn, die andere nördlich von Brilon. Die Ursache für die starke Streuung ist unklar.

Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,73, was für eine Vorhersage noch gut vertretbar erscheint. Allerdings weist Großraum 8 durch die wenigen verbleibenden Messstellen neben Großraum 6 und 7 die geringste Messstellendichte auf, was eine flächenhafte Vorhersage mit Ausnahme kleinerer Bereich unmöglich macht.

2.4.10 Großraum 9: Südostdeutsches Grundgebirge

Das Südostdeutsche Grundgebirge erstreckt sich entlang der Grenzen Bayerns, Thüringens und Sachsens zu Tschechien. Geologisch umfasst es das Grundgebirge (Saxothuringikum und Moldanubikum) im Bereich von Thüringer Wald, Vogtland, Fichtel- und Erzgebirge, sowie Oberpfälzer und Bayerischem Wald. Es überwiegen Kluftgrundwasserleiter und -geringleiter mit nur untergeordneter wasserwirtschaftlicher Bedeutung. Daher wird an dieser Stelle auf eine Beschreibung der sieben Räume verzichtet. Eine Ausnahme bildet mit Raum 91 der Elbtalgraben, wo eiszeitliche Kiese und Sande ergiebige Grundwasserleiter bilden.

Die 435 Messstellen der Großraums (405 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) sind sehr ungleich verteilt und liegen praktisch ausschließlich in Thüringen und Sachsen. Auf eine Unterteilung des Großraums wurde daher verzichtet.

Die Messstellen wurden auf Grund von 4 Features (Tabelle 2.5) in 26 Cluster gruppiert, von denen 11 verworfen wurden und 15 erhalten blieben. Verworfen wurden hier vor allem kleinere Cluster mit weniger als 10 Messstellen, so dass insgesamt 246 Messstellen nach Aussortieren und Kappung der Cluster übrig bleiben.

Das größte Cluster (Cluster 1) mit 28 Messstellen, die mehrheitlich in Raum 91 und 93 liegen, zeigt eine gute Homogenität mit $\overline{R_{w\ max}} = 0,84$ und eine geringe jährliche Varianz. Cluster 2 mit 26 Messstellen, vor allem im Raum 93 und 94, zeigt dagegen einen deutlicher ausgeprägten Jahresgang, aber eine geringere Homogenität ($\overline{R_{w\ max}} = 0,70$). Cluster 3 kennzeichnet sich vor allem durch eine positive Schiefe, vermutlich verursacht durch Hochwassereinfluss, denn viele der Messstellen sind an Flüssen gelegen. Cluster 4 (26 Messstellen) erscheint ähnlich zu Cluster 2, mit jedoch noch stärker ausgeprägtem Jahresgang. In Cluster 8 konzentrieren sich die Messstellen entlang der Elbe in Raum 91, sowie einzelne Messstellen an kleineren Flüssen.

Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt 0,71. Neben Großraum 8 zeigt sich (abgesehen von Großraum 7) die geringste mittlere Homogenität der Cluster. Dennoch erscheint eine ausreichend gute Vorhersage

der verbleibenden Messstellen realistisch, in dem gut abgedeckten nordöstlichen Bereich auch flächenhaft.

2.4.11 Großraum 10: Südwestdeutsches Grundgebirge

Zum Südwestdeutschen Grundgebirge gehören die Verbreitungsgebiete des Kristallins im Schwarzwald, Odenwald und Spessart. Der Großraum 10 entspricht dabei dem Raum 101 (Schwarzwald, Vorspessart und Odenwald). Die Gesteine bilden Kluftgrundwasserleiter geringer Ergiebigkeit, lediglich in oberflächennahen Auflockerungszonen und quartären Talfüllungen bilden sich lokal teilweise größere Grundwasservorkommen.

Die 75 Messstellen (47 nach Preprocessing und Eingrenzung des Untersuchungszeitraums) konzentrieren sich zum größten Teil auf den nördlichen, räumlich durch die Kraichgau-Senke abgetrennten Bereich. Im südlichen Teil, dem Kristallin des Schwarzwaldes treten die wenigen Messstellen zudem teils lokal konzentriert auf.

Es wurden 5 Cluster auf Grund der Auswahl von 5 Features gebildet (Tabelle 2.5). Von den Clustern wurde nur eines (Cluster 5 mit 4 Messstellen) auf Grund zu geringer Homogenität verworfen. Im weiteren Verlauf wurden 3 weitere Messstellen am Rande der Cluster gekappt, so dass 40 Messstellen verblieben.

Das größte Cluster (Cluster 1 mit 13 Messstellen) konzentriert sich dabei auf den nördlichsten Teil des Großraum (Vorspessart), mit Ausnahme einer Messstelle im Odenwald. Mit einem $\overline{R_{w\ max}}$ -Wert von 0,72 erscheint das Cluster noch ausreichend homogen für eine Vorhersage. Cluster 2 vereint 5 Messstellen im Schwarzwald, vier im Odenwald und 3 im Vorspessart. Obwohl an der Dynamik der Ganglinien des Clusters diese Dreiteilung deutlich erkennbar ist, wurden alle dem gleichen Cluster, mit daher allerdings nur mittlerer Homogenität ($\overline{R_{w\ max}} = 0,69$) zugeordnet, da die Ähnlichkeit untereinander dennoch größer ist als mit den Ganglinien der anderen Cluster. Cluster 3 umfasst 9 Messstellen in den beiden nördlichen Teilgebieten, die sich gegenüber dem Rest durch eine erhöhte Streuung (Feature SEM) auszeichnen und untereinander mit $\overline{R_{w\ max}} = 0,76$ recht homogen sind. Die restlichen 6 Messstellen befinden sich in direkter Nachbarschaft zueinander im Schwarzwald. Ihr Verlauf ist praktisch identisch, was sich in einem $\overline{R_{w\ max}}$ -Wert von 0,80 zeigt. Da die Zeitreihenlänge mit unter 10 Jahren recht kurz ist, könnte eine Vorhersage schwierig werden.

Der mittlere $\overline{R_{w\ max}}$ -Wert über alle nicht verworfenen Cluster nach der Bereinigung beträgt immerhin 0,74, was auf eine ausreichend gute Homogenität für eine spätere Vorhersage schließen lässt. Allerdings konzentrieren sich auch hier die wenigen geeigneten Messstellen (40 von ursprünglich 47) auf enge Bereiche. Am ehesten erscheint hier noch der nördliche Teil des Großraums (Vorspessart) für eine flächenhafte Vorhersage geeignet. Im südlichen Teil (Schwarzwald) ist dagegen bestenfalls eine punktuelle Vorhersage denkbar.

2.5 Zusammenfassung und Diskussion der Referenzmessstellenauswahl

Neben der Zusammenstellung und Homogenisierung einer umfangreichen Datengrundlage von insgesamt ca. 13.500 Messstellen mit Unterstützung der geologischen Dienste der Länder umfasste die erste Projektphase hauptsächlich die Entwicklung einer geeigneten und bundesweit einheitlichen Methodik zur Auswahl von Referenzmessstellen. Hierzu wurden zunächst existierende Methoden recherchiert, getestet und bewertet. Da sich diese jedoch aus verschiedenen Gründen als nicht geeignet erwiesen haben, vor allem hinsichtlich der Automatisierbarkeit und damit Anwendbarkeit auf einen solch umfangreichen Datensatz, wurde im Verlauf des Projekts eine neue Methode entwickelt. Diese basiert auf einem speziellen Typ von Künstlichen Neuronalen Netzen, den sog. Self-Organizing-Maps (SOM). Die Methodik ähnelt dabei einem klassischen Clusteransatz. Es werden Gruppen mit Ähnlichkeiten in vorgegebenen Parametern gebildet. Ziel war es, Gruppen von Grundwasserganglinien mit ähnlicher Dynamik zu finden. Da sich das Clustering der Original-Zeitreihen aus Gründen der Datenqualität sowie des Rechenaufwands nicht realisieren ließ, wurde der Ansatz gewählt, die Dynamik der Ganglinien über sog. Features zu beschreiben. Dabei werden aus den Daten selbst beschreibende mathematische Parameter extrahiert, die später geclustert werden und im besten Fall Ganglinien gleicher Dynamik einem gemeinsamen Cluster zuordnen. Dafür war vorab eine Definition räumlicher Einheiten nötig, auf deren Basis das Clustering angewendet wird. So können die beschreibenden Features optimiert werden und der Kausalzusammenhang der Ganglinedynamik bleibt weitgehend gewährleistet. Die Wahl fiel auf die Hydrogeologischen Großräume von Deutschland.

Nach der Entwicklung und Optimierung der Methodik an Großraum 3 wurde diese schließlich auf den gesamten Datensatz angewandt. Von den 13.284 nach dem Preprocessing verbleibenden Messstellen (vgl. Tab. 2.6) mussten weitere fast 2.000 Messstellen nach der Eingrenzung der Untersuchungszeiträume entfernt werden, wobei sich diese sehr ungleich auf die Großräume verteilten. Mit den verbleibenden 11.303 Messstellen wurde für 9 der 10 Großräume ein Clustering auf Großraumebene durchgeführt. Großraum 3 wurde für das Clustering in zwei Gebiete, Raum 31/32 und Raum 33 unterteilt. Insgesamt ergeben sich über alle Großräume daraus 430 Cluster, was einem Schnitt von 26 Messstellen pro Cluster entspricht. Die Clustergröße variiert dabei jedoch extrem von weit über 100 bis unter 10 Messstellen.

Tabelle 2.6: Übersicht Messstellen- und Clusteranzahl der einzelnen Arbeitsschritte.

Messstellen Originaldaten	13.492
Messstellen n. Preprocessing	13.284
Messstellen Clustering	11.303
Cluster gesamt	430
Cluster verbleibend	166
Messstellen verbleibend	5.533
Cluster entfernt	264
Messstellen entfernt	5.770

Von diesen 430 Clustern wurden 264 Cluster auf Grund zu geringer Homogenität verworfen, 166 Cluster blieben erhalten. Aus den erhaltenen Clustern wurden weiterhin am Clusterrand gelegene Messstellen entfernt, um die Korrelation der verbleibenden Ganglinien untereinander, im Sinne eines vergleichbaren Maßes für die Homogenität und die spätere Vorhersagbarkeit zu erhöhen. Insgesamt blieben 5.533 Messstellen in 166 Clustern erhalten, die potentiell für eine Vorhersage an Hand der Referenzmessstelle des jeweiligen Clusters (die Messstelle, welche am höchsten mit allen anderen Messstellen des Clusters korreliert) geeignet sein sollten. Nimmt man den Großraum 7 aus, bei dem auf Grund nur 16 vorhandener Messstellen kein sinnvolles Clustering möglich war, so beträgt der Mittelwert der Korrelation der Referenzmessstellen ($\overline{R_{w \max, Cl}}$ - Wert) mit den restlichen Ganglinien des jeweiligen Clusters im Mittel über alle Cluster 0,74, bei Maximalwerten für einzelne Cluster von bis zu 0,90 und Maximalwerten für Großraum 4 und Raum 31/32 von 0,80. Da die berechneten Korrelationen auf längeren Zeiträumen von mindestens 10, in den meisten Fällen sogar über 30 Jahren beruhen, so ist zu erwarten, dass sich die spätere Vorhersage an einer Referenzmessstelle mit der gleichen Korrelation auf die restlichen Messstellen eines Clusters übertragen lässt.

Tabelle 2.7 zeigt eine Übersicht über die Messstellen- und Clusteranzahlen in den Großräumen, maximale und mittlere $\overline{R_w}$ -Werte sowie ursprüngliche und verbleibende mittlere Messstellendichte.

Tabelle 2.7: Übersicht Messstellen (Mst.) und Clusteranzahl (Cl.) vor (ges.) und nach (verbl.) der Entfernung der Aussortiercluster und der Bereinigung der Messstellen mit geringeren R_w -Werten innerhalb der Cluster, maximale und mittlere R_w -Werte der Cluster sowie durchschnittliche Messstellendichte (MStD.) in Messstellen pro 100 km².

	GR 1	GR 2	R 31,32	R 33	GR 4	GR 5	GR 6	GR 7	GR 8	GR 9	GR 10
Mst. ges.	5.291	888	1.774	80	624	1.175	567	16	436	405	47
Cl. ges.	237	17	22	4	11	43	27	7	31	26	5
Cl. verbl.	66	13	17	2	8	15	10	4	12	15	4
Mst. verbl.	2.480	580	1.014	22	273	441	242	10	185	246	40
Cl. entf.	171	4	5	2	3	28	17	3	19	11	1
Mst. entf.	2.811	308	760	58	351	734	325	6	251	159	7
Max. $\overline{R_w}$	0,90	0,88	0,90	0,68	0,88	0,80	0,78	0,45	0,90	0,84	0,80
Mittelwert: $\overline{R_{w \max, Cl}}$	0,74	0,77	0,80	0,67	0,80	0,74	0,73	0,42	0,72	0,71	0,74
MStD	4,1	4,7	19,0	2,5	2,0	2,9	1,0	0,3	1,3	1,5	0,9
MStD. verbl.	1,9	3,0	10,9	0,7	0,9	1,1	0,4	0,2	0,6	0,9	0,7

ges. - gesamt, verbl. - verbleibend, entf. - entfernt

Eine detaillierte Darstellung der initialen sowie verbleibenden Messstellendichte auf Ebene der Räume, Teilräume sowie der Grundwasserkörper findet sich in Anhang E. Mit den verbleibenden Messstellen wurde teilweise eine sehr gute räumliche Abdeckung erreicht, die für einige Gebiete eine flächenhafte Vorhersage von Grundwasserständen realistisch macht (vgl. Kapitel 5). In anderen Gebieten dagegen ist die Dichte der verbleibenden Messstellen jedoch gering. Dies zeigt vor allem die Darstellung auf Ebene der Grundwasserkörper. Hier zeigen sich einige, in denen schon initial keine Messstellen vorhan-

den waren. Darüber hinaus hat sich aber vor allem in Gebieten, wo bereits zu Beginn die Abdeckung eher dünn war, diese zum Teil nochmals deutlich reduziert. Mehrheitlich dürfte dies an den hydrogeologischen Verhältnissen liegen, auf Grund derer sich die Grundwasserstandsganglinien lokal sehr unterschiedlich verhalten und daher eine Bildung sinnvoller Cluster erschweren (wie im Alpenvorland im Bereich des Molassebeckens), untergeordnet aber auch an mangelnder Datenqualität. Vor allem lückenhafte Daten und kurze Messzeiträume erschweren den Prozess des Clusterings und machen darüber hinaus auch eine spätere Vorhersage unmöglich.

Zur räumlichen Abschätzung einer potentiellen Eignung eines Gebiets für eine spätere Vorhersage wurden für verschiedene räumliche Aggregationen (Räume, Teilräume und Grundwasserkörper) die Durchschnitte der gewichteten Korrelationskoeffizienten ($\overline{R_w}$ -Werte) pro Gebiet berechnet und mit verbleibenden Messstellendichte multipliziert. Streng genommen hat der resultierende Wert damit die gleiche Dimension wie die Messstellendichte (Messstellen pro 100 km^2), sollte jedoch zu diesem Zeitpunkt mehr als relatives Maß zum Vergleich der räumlichen Einheiten untereinander verstanden werden. Abb. 2.10 zeigt die Darstellung für die Teilräume. Karten der Räume und Grundwasserkörper finden sich in Anhang F. Hier zeigt sich neben dem Einfluss der durchschnittlichen Korrelation, welche hauptsächlich von der Komplexität der hydrogeologischen Verhältnisse bestimmt wird, auch der Einfluss der Messstellendichte. Abgesehen davon, dass in Gebieten mit vielen Messstellen die Chance, Messstellen mit ähnlicher Dynamik und damit einer hohen Korrelation anzutreffen, höher ist, ist eine hohe Messstellendichte auch Voraussetzung dafür, später punktuelle Vorhersagen auch in die Fläche zu bringen (siehe Kapitel 5).

Zusammenfassend lässt sich sagen, dass die entwickelte Methode zur Referenzmessstellenauswahl sehr gut geeignet ist, Referenzmessstellen selbst sowie Messstellen, für die diese als Referenz gelten, zu definieren. Sie bietet gegenüber den bisher angewandten Methoden einige Vorteile. So liefert sie ein mathematisches Gütemaß des Zusammenhangs von Referenzmessstellen und Vorhersagemessstellen. Die Auswahl, die vorher oft subjektiv und auf Bundeslandebene unterschiedlich war, wird vereinheitlicht und mathematisch begründet objektiviert. Weiterhin eröffnet sie die Möglichkeit, in den Daten Zusammenhänge zu erkennen, wie sie vorher alleine auf Grund der schieren Menge der Daten nicht möglich gewesen wäre.

Dennoch ist weiterhin an vielen Stellen der Eingriff eines Bearbeiters mit hydrogeologischem Sachverstand erforderlich. Durch den notwendigen Kompromiss des Clusterings über beschreibende Features wird zunächst eine Vorauswahl dieser sowie ggf. für andere Gebiete eine zusätzliche Definition neuer Features nötig. Weiterhin müssen die gebildeten Cluster, deren Ähnlichkeit in den beschreibenden Features zum Teil nicht auf dem Verlauf der Ganglinien selbst, sondern auf z.B. der Gemeinsamkeit schlechter Datenqualität beruht, aussortiert und beurteilt werden. Die an Hand der höchsten Korrelation getroffene Wahl der Referenzmessstelle muss ebenfalls überprüft werden hinsichtlich weiterer Anforderungen zu deren Eignung. So ist die Methodik als eine gute Ergänzung und Unterstützung des Hydrogeologen zu sehen, die es erlaubt, die großen anfallenden Datenmengen einheitlich zu strukturieren und auszuwerten und so neue Erkenntnisse über die Grundwasserdynamik zu erlangen. Damit erscheint zukünftig auch eine bessere und genauere Vorhersage von Grundwasserständen in vielen Gebieten möglich.

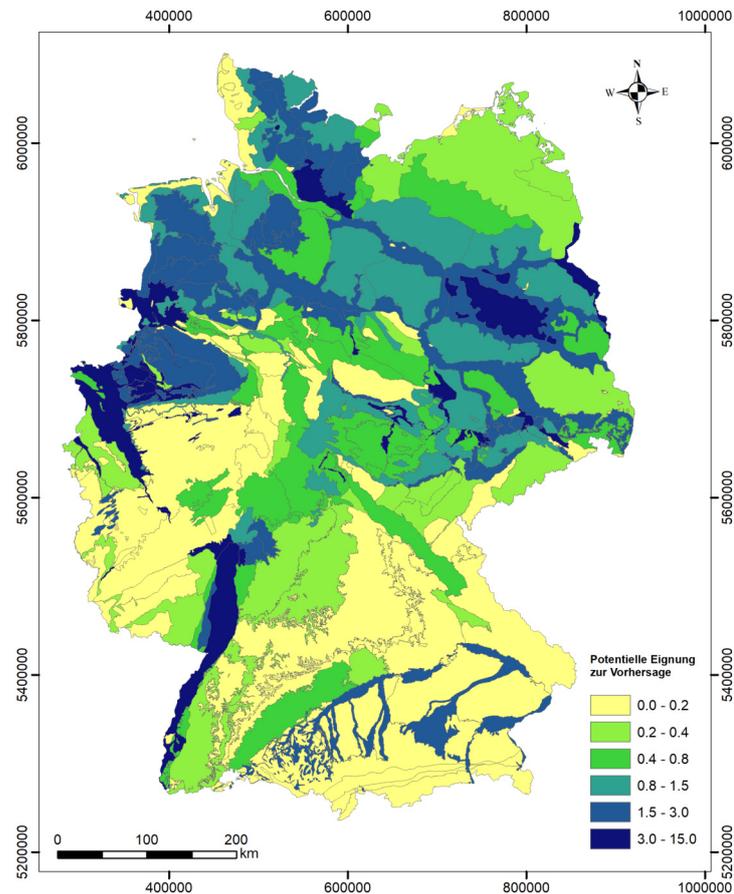


Abbildung 2.10: Potentielle Eignung der Teilräume für eine spätere Vorhersage, berechnet aus dem Durchschnitt der gewichteten Korrelationskoeffizienten ($\overline{R_w}$ -Werte) der verbleibenden Messstellen multipliziert mit der verbleibenden Messstellendichte eines Teilraums. Streng genommen beträgt die Einheit Messstellen pro 100 km^2 , sollte jedoch mehr als relatives Maß zum Vergleich der räumlichen Einheiten untereinander verstanden werden.

3 Grundwasserstandsvorhersage an Referenzmessstellen

3.1 Zielsetzung

Ziel dieses Arbeitsschrittes ist die Entwicklung und Anwendung einer übertragbaren, auf künstlichen neuronalen Netzen basierende Methodik zur Berechnung von Wochen-, Monats-, und Jahreszeitvorhersagen (3 Monate) an den 166 Referenzmessstellen. Datengrundlage hierfür waren zum einen die wöchentlichen Grundwasserstandszeitreihen der Referenzmessstellen, zum anderen beobachtete und modellierte Klimarasterdaten des Deutschen Wetterdienstes (DWD).

Künstliche neuronale Netze (KNN, Artificial Neural Networks ANN) sind eine Gruppe von etablierten Methoden des maschinellen Lernens, welche die Funktionsweise des menschlichen Gehirns nachahmen und so in der Lage sind, selbstständig komplexe Beziehungen zwischen Eingangs- und Zielgrößen zu erfassen. Dabei muss zwischen Input und Output keine direkte oder indirekte physikalische Beziehung existieren, da auch innerhalb des Modells keine physikalischen Gleichungen gelöst werden. Wichtig ist lediglich, dass zwischen beiden Größen ein mathematischer Zusammenhang existiert. Dies ist zwar vorwiegend für Größen der Fall, für die auch ein physikalischer Zusammenhang besteht, zwingend erforderlich ist dieser aber nicht. Verallgemeinert gesprochen bestehen KNN aus schichtweise angeordneten und gewichtet verbundenen Einheiten, sog. Neuronen. Die Anzahl der Input- und Output-Neuronen in der Input- bzw. Outputschicht wird durch die Anzahl der Input- und Outputvariablen bestimmt. Zwischen diesen Schichten befinden sich keine, eine oder mehrere verdeckte Schichten. Flache KNN (shallow ANN) besitzen maximal eine verdeckte Schicht, tiefe KNN (deep ANN) hingegen besitzen viele verdeckte Schichten. Die Auswahl eines entsprechenden KNN-Ansatzes richtet sich vor allem nach der verfügbaren Datenmenge. Erst bei sehr großen Datensätzen können Deep-Learning Methoden ihre Stärken ausspielen, flache und generell kleiner dimensionierte KNN sind für mittlere Datensatzgrößen der Ansatz der Wahl und liefern hierbei in der Regel sowohl bessere Ergebnisse als traditionelle Methoden des Maschinellen Lernens als auch Deep-Learning Ansätze.

Grundwasserstandsvorhersagen mit künstlichen neuronalen Netzen wurden bereits vielfach erfolgreich in der Literatur beschrieben (Adamowski und Chan, 2011; Alizamir et al. 2018; Barzegar et al. 2017; Chang et al. 2016; Chitsazan et al. 2015; Coppola et al. 2003; Coulibaly et al. 2001; Daliakopoulos et al. 2005; Gorgij et al. 2017; Guzman et al. 2017, 2019; Izady et al. 2013; Jha und Sahoo, 2015; Khalil et al. 2015; Kisi et al. 2017; Lallahem et al. 2005; Mohammadi, 2008; Nayak et al. 2006; Sahoo et al. 2017; Sreekanth et al. 2009; Taormina et al. 2012; Trichakis et al. 2011; Wunsch et al. 2018). Einen Überblick über die Anwendung von Methoden der künstlichen Intelligenz zur Grundwasserstandsvorhersage findet sich in Rajaei et al. (2019). Im vorliegenden Fall wurde der Ansatz

von Wunsch et al. (2018) verfolgt. Hierbei werden sogenannte nichtlineare autoregressive Netze mit exogenem Input (Nonlinear Autoregressive Networks with Exogenous Inputs NARX) genutzt, die basierend auf einfachen und flächendeckend verfügbaren, meteorologischen Inputvariablen (Niederschlag und Temperatur) Grundwasserstandsvorhersagen berechnen. Es handelt sich hierbei um einen sehr gut übertragbaren Ansatz, der zumindest für unbeeinflusste Messstellen nachweisbar gute Ergebnisse liefert. Unbeeinflusst meint in diesem Fall, dass die natürliche Grundwasserdynamik, die in der Regel v.a. von Grundwasserneubildung und damit von Niederschlag und Evapotranspiration abhängt, nicht von anderen Faktoren überprägt wird. Dies können beispielsweise anthropogene Eingriffe oder auch eine große Nähe zum Vorfluter sein. In diesen Fällen ist dann der angesprochene notwendige Zusammenhang zwischen Eingangs- (Niederschlag, Temperatur) und Zielgrößen (Grundwasserstand) so stark überprägt, dass er sich nicht mehr in den Daten widerspiegelt, da andere Faktoren die Grundwasserdynamik dominieren.

3.2 Datengrundlage

3.2.1 Grundwasserdaten

Die genutzten Grundwasserstandszeitreihen wurden freundlicherweise zu Projektbeginn von den Staatlichen Geologischen Diensten der Länder bereitgestellt. Für eine genaue Beschreibung des Datensatzes und des angewandten Preprocessings sei an dieser Stelle auf Kapitel 2.2 verwiesen. Aus diesem Gesamtdatensatz von wöchentlichen Grundwasserstandszeitreihen werden die der 166 betrachteten Referenzmessstellen zur Vorhersage genutzt. In der Regel wird die maximal verfügbare lückenlose Zeitreihenlänge verwendet, in einigen wenigen Fällen (9) wurden die Zeitreihen jedoch gekürzt, wenn sich die Grundwasserdynamik über die Zeit stark verändert hat. So wird verhindert, dass die KNN auf einen Zusammenhang trainiert werden, der heute so nicht mehr besteht. Abbildung 3.1 zeigt dies exemplarisch am Beispiel der Messstelle BW_157-260-8. Hier wurde für das Training der Bereich vor 1980 abgetrennt, da sich zwischen 1970 und 1980 eine deutliche Absenkung der Grundwasserstände zeigt, welcher sich ab den 1980er Jahren wieder erholt und stabilisiert.

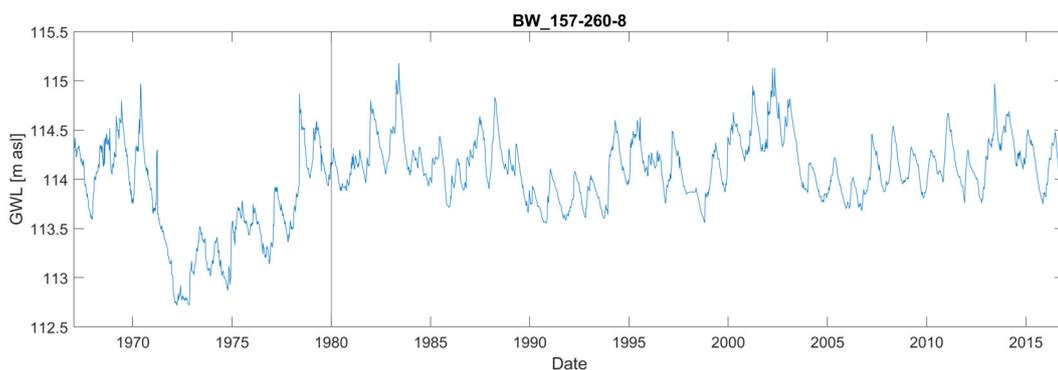


Abbildung 3.1: Zeitreihe der Messstelle BW_157-260-8, mit deutlicher Inhomogenität in den 1970er Jahren. Für das KNN-Training wurde die Zeitreihe gekürzt, mit neuem Beginn um 1980 (schwarze Linie).

Nicht für alle ausgewählte Referenzmessstellen liegt eine lückenlose Zeitreihe vor. Bei Bedarf wurden daher Datenlücken mit Hilfe des Verlaufes eng verwandter bzw. im Cluster (nicht räumlich, sondern mathematisch) benachbarter Messstellen gefüllt. Jede dieser sogenannten Nachbarschaftsinterpolationen beruht auf einer linearen Regression der fünf Werte, welche die Datenlücke auf jeder Seite umgeben. Diese Anzahl der Werte hat sich über trial-and-error als praktikabel erwiesen. Für gute Ergebnisse ist zudem eine hohe Gesamtkorrelation beider Zeitreihen nötig. Die Begrenzung der linearen Regression auf die nächste Umgebung der Datenlücke dient lediglich einer Erhöhung der Genauigkeit der Nachbarschaftsinterpolation im konkreten Zeitreihenabschnitt. Jede generierte Interpolation wurde händisch mit einer alternativen auf „Piecewise Cubic Hermite Interpolating Polynomial (PCHIP)“ basierenden Interpolation verglichen und die plausible ausgewählt. Im Falle von nicht plausibel interpolierbaren Datenlücken können Zeitreihen ganz allgemein bei diesem Ansatz entsprechend gekürzt werden. Dieser Fall trat für die 166 Referenzmessstellen jedoch nicht auf. Nur für 16 der betrachteten Referenzmessstellen war eine Interpolation nötig. Abbildung 3.2 illustriert die Nachbarschaftsinterpolation am Beispiel der Messstelle BB_32529245. Die Datenlücken wurden in diesem Fall mit dem zweiten Nachbarn im Cluster geschlossen, da der 1. Nachbar identische Datenlücken aufweist.

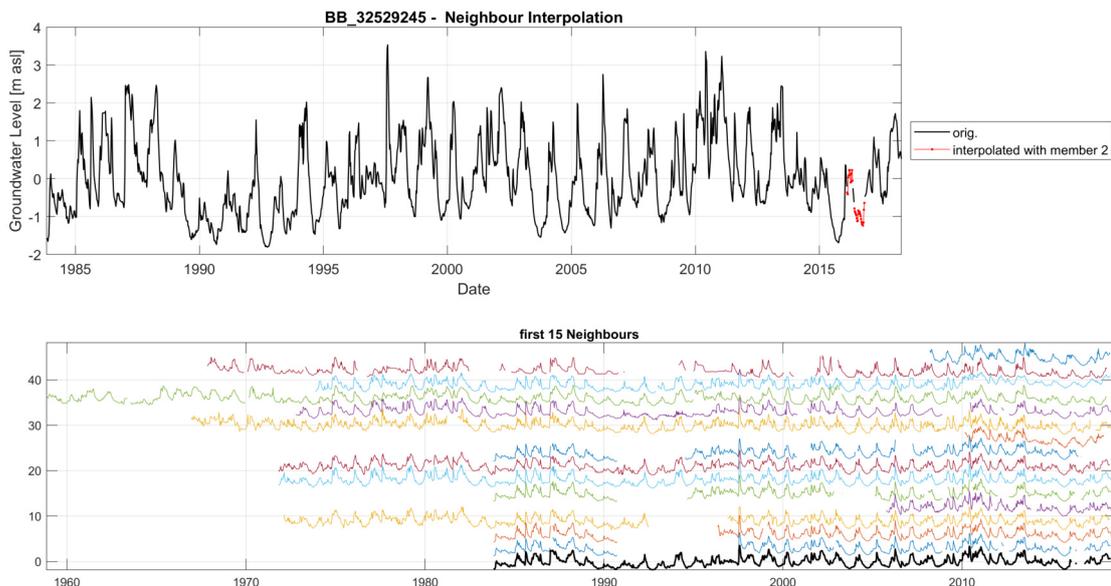


Abbildung 3.2: Illustration der Nachbarschaftsinterpolation für die Datenlücken der Referenzmessstelle BB_32529245 mit Hilfe des zweiten Nachbarn im Cluster

3.2.2 Klimadaten

Die für das KNN-Training und für einen operationellen Betrieb der Grundwasserstandsvorhersage notwendigen Inputdaten wurden freundlicherweise vom DWD zur Verfügung gestellt. Das Training der KNN erfolgt auf Basis von HYRAS (HYdrologische RASTERdatensätze), welche **beobachtete** Tagesdaten mit einer Auflösung von $12,5 \times 12,5 \text{ km}^2$ für Niederschlag (P), Temperatur (T) und relativer Luftfeuchte (rH) von 1951-2015 umfassen (Frick et al. 2014; Rauthe et al. 2013). Weiterhin standen Daten des TRY Datensatzes

(Windgeschwindigkeit und Solarstrahlung) zur Verfügung (DWD, 2017), welche sich jedoch im Verlaufe des Projektes als nicht geeignet herausgestellt haben und auf die aus diesem Grund nicht näher eingegangen werden soll.

Da für einen potenziellen operationellen Betrieb der Grundwasserstandsvorhersagen keine Beobachtungsdaten zur Verfügung stehen, wurde zudem der EPISODES Datensatz des DWD genutzt. Bei EPISODES handelt es sich eigentlich um eine Empirisch-Statistische Downscaling Methode (Kreienkamp et al. 2019), im vorliegenden Fall ist damit jedoch ein mit Hilfe von EPISODES erstellter Rasterdatensatz auf Basis saisonaler Vorhersagen des GCFS 2.0 (German Climate Forecasting System) (DWD, 2018, 2019) gemeint. Damit stehen **modellerte** Hindcasts von Tagesdaten mit einer Auflösung von $12,5 \times 12,5 \text{ km}^2$ für Niederschlag, Temperatur und Luftfeuchte von 1990-2017 zur Verfügung. Dabei wird monatlich ($12 \times \text{pro Jahr}$) ein Hindcast für die nächsten 6 Monate berechnet. Jeder Hindcast umfasst dabei 30 Ensemble Member, die unterschiedliche Startbedingungen/Initialisierungen repräsentieren. Für eine Verarbeitung wurden alle verwendeten Klimadaten von Tages- auf Wochendaten aggregiert. Neben den generell zu erwartenden Unsicherheiten, die bei einer Klimavorhersage auftreten, besitzt vorliegender Modelldatensatz zusätzlich je nach Startzeitpunkt eine variable Vorhersagegüte. Abbildung 3.3 zeigt die Güte der Ensembledittel-Klimavorhersage von Niederschlag und Temperatur als 3-Monatsmittel der Monate 1 bis 3 nach Modellstart, geordnet nach Startmonat (Quelle: DWD). Zur Evaluierung der Güte dient die direkte Korrelation (Pearson) mit dem langjährigen klimatologischen Mittel. Generell ist die Vorhersagegüte der EPISODES-Temperatur am besten im Sommer und Winter, am schlechtesten im Frühling und Herbst. Für Niederschlag ist das Bild nicht so eindeutig, da mehr räumliche Variabilität auftritt.

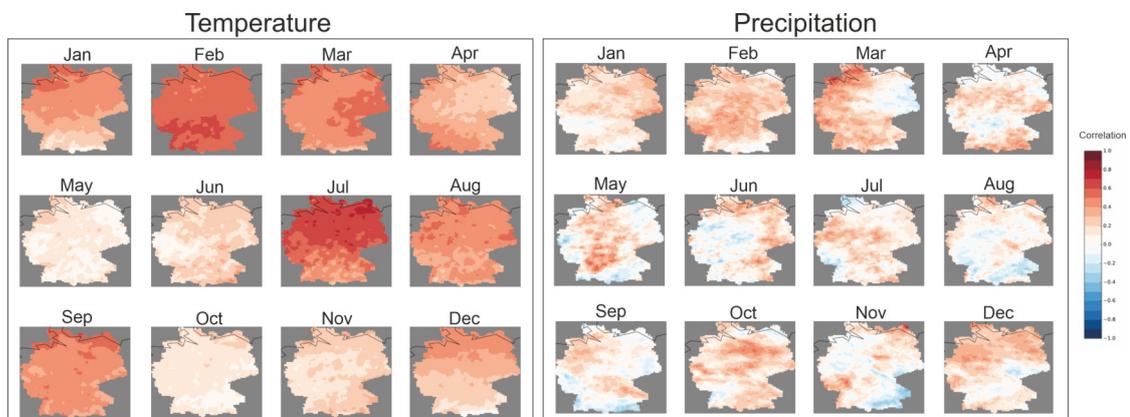


Abbildung 3.3: Güte der Ensembledittel-Klimavorhersage von Niederschlag und Temperatur als 3-Monatsmittel der Monate 1 bis 3 nach Modellstart, geordnet nach Startmonat (Quelle: DWD)

3.3 Methoden

3.3.1 Inputparameter

Die Wahl der Input-Parameter orientiert sich stark an dem von Wunsch et al. (2018) gewählten Ansatz. Hierbei werden lediglich die in der Regel flächendeckend verfügbaren meteorologischen Inputvariablen Niederschlag (P) und Temperatur (T) in Betracht gezogen. Der Niederschlag dient dabei als Surrogat für die Grundwasserneubildung, die Temperatur beinhaltet indirekt den Zusammenhang mit der Evapotranspiration und bietet dem Netz gleichzeitig Informationen zur Saisonalität durch den für gewöhnlich deutlichen Jahresgang. Im hier genutzten Ansatz wird zusätzlich die ebenfalls flächendeckend verfügbare relative Luftfeuchtigkeit (rH) genutzt, um diesen Zusammenhang noch genauer abbilden zu können. Als zusätzlicher synthetischer Inputparameter (Tsin) wird weiterhin ein Sinussignal an die Temperatur gefittet. Dieser bietet dem KNN somit rauschfreie Information über die Saisonalität, wodurch oft deutlich verbesserte Vorhersagen erzielt werden können. Der zweifellos wichtigste Inputparameter ist P, da die Grundwasserneubildung in der Regel den größten Einfluss auf die Grundwasserdynamik hat. Daher wird P immer als Inputparameter genutzt, die Eignung der restlichen Inputparameter wird für jede Messstelle einzeln überprüft und die optimale Inputkonfiguration über eine automatische Optimierung festgelegt (s.a. Abschnitt 3.3.2).

3.3.2 KNN-Modell

NARX - Nonlinear Autoregressive Networks with Exogenous Inputs

NARX gehören zur Gruppe der dynamischen, rekurrenten künstlichen neuronalen Netze (RNN) und besitzen eine Feedbackverbindung zwischen Output- und Inputschicht (Abb. 3.4a). Dadurch wird ein rückwärts gerichteter Informationsfluss ermöglicht sowie Berechnungen auch in Abhängigkeit vorheriger Zeitschritte durchgeführt. Dies bewirkt einen Gedächtniseffekt innerhalb des Modelles, weshalb sich NARX gut zur Modellierung nichtlinearer, dynamischer Zeitreihen eignen (Beale et al. 2016; Lin et al. 1996). Generell weisen RNN besonders im Vergleich zu KNN ohne Feedbackverbindungen (sog. Feed-forward Netze - FFN) eine erhöhte Trainingskomplexität auf, wodurch unter anderem eine stärkere Startwertabhängigkeit bei der Suche nach dem globalen Fehlerminimum entsteht. Die Betrachtung unterschiedlicher Netzinitialisierungen im Training ist daher für diese Suche von großer Bedeutung. NARX bieten zudem die Möglichkeit, die Feedbackverbindung während des Trainings ungenutzt zu lassen und das Feedback durch eine verzögerte Eingabe des Training-Targets (hier: beobachtete Grundwasserstände) zu ersetzen. Diese sogenannte open-loop-Konfiguration (Abb. 3.4b) verringert im Gegensatz zur Standardkonfiguration mit Feedbackverbindung (closed-loop) den Trainingsfehler für gewöhnlich deutlich, zudem wird die Trainingsdauer hierdurch oft stark verringert. Aufgrund des Bias-Variance Dilemmas, unter dem KNN ganz allgemein leiden (Geman et al. 1992), lässt sich zwar von einem kleinen Trainingsfehler nicht auf einen kleinen Vorhersagefehler schließen, jedoch wird in der Tat im open-loop der Startwert der Vorhersage genauer getroffen. Die Vorhersagen selbst werden dann jedoch immer im closed-loop Setup durchgeführt. Welche Trainingskonfiguration bessere Ergebnisse liefert, ist stark vom Einzelfall abhängig. Üblicherweise kann davon ausgegangen werden, dass ein Training im

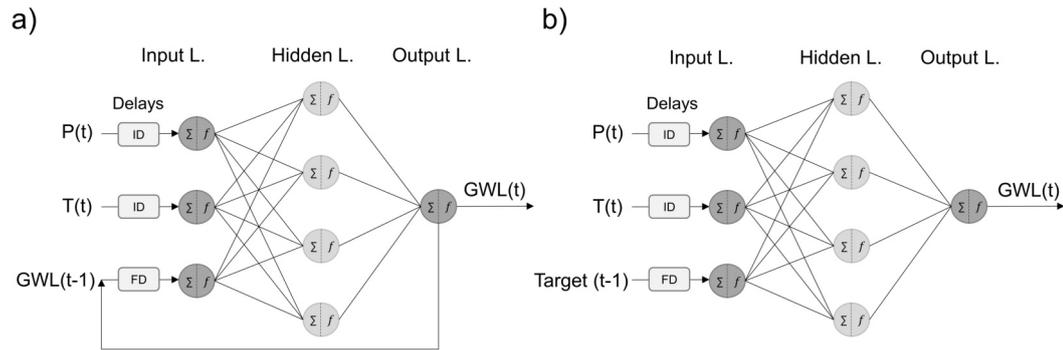


Abbildung 3.4: typische NARX Architekturen aus Wunsch et al. (2018) mit Inputs Niederschlag (P) und Temperatur (T), Output Grundwasserstand (GWL) sowie Inputdelays (ID) und Feedbackdelay (FD); a) closed-loop setup b) open-loop setup

open-loop bessere globale Fehlerwerte liefert. Nach closed-loop Trainings wird hingegen, bei gleichzeitig etwas höheren globalen Fehlerwerten, die Dynamik der Zeitreihe oft sehr viel besser erfasst, da das Modell wirklich auf die modellierten Outputwerte und nicht auf die beobachteten Outputs (Targets) trainiert wird. Auch die Länge des Vorhersagehorizontes spielt bei der Entscheidung für eine der beiden Trainingsarchitekturen eine wichtige Rolle. Je länger dieser ist, umso wichtiger wird eine gut angepasste Dynamik, und das rekurrente Training (closed-loop) ist dann in der Regel vorzuziehen. Vor allem für kurze Vorhersagehorizonte ist die Dynamik (je nach Zeitreihe) oft jedoch weniger wichtig als ein exakterer Startpunkt der Vorhersage. Ebendieser erweist sich vor allem bei suboptimaler Modellanpassung nach einem closed-loop Training oft als Problem und kann deutliche Fehler über die ersten Zeitschritte verursachen.

Sogenannte Delays (Inputdelays und Feedbackdelays) in der Inputschicht von NARX-Modellen übernehmen die Funktion eines Kurzzeitgedächtnisses (Barreto, 2007). Die Länge eines Delays legt fest, wie lange ein Wert als Input zur Verfügung steht, wodurch dem Netz immer Werte mehrerer Zeitschritte als Input präsentiert werden.

Um ein Übertraining (Overfitting) der Modelle zu verhindern, werden für NARX, wie auch für KNN im Allgemeinen, in der Regel Regularisierungstechniken eingesetzt. Im vorliegenden Fall handelt es sich dabei um Early-Stopping und Kreuzvalidierung (CV - Crossvalidation). Dafür wird der verfügbare Datensatz in drei Teile geteilt. Der Großteil der Daten wird zum Training des Netzes genutzt (Trainings-Set), zwei weitere kleinere Teile werden je für Early-Stopping (Stop-Set) und Kreuzvalidierung (Test-Set) genutzt. üblicherweise stehen die drei Teile in Verhältnissen der Größenordnung 70%-15%-15% oder 80%-10%-10%. Während des Trainings wird parallel auch immer der Trainingsfehler des Stop-Sets mitberechnet. Sobald der Trainingsfehler weiter fällt, der Stop-Set-Fehler jedoch ansteigt, wird das Training abgebrochen und so eine optimale Modellkomplexität bestimmt. Um die Generalisierungsfähigkeit des trainierten Netzes zu bestimmen, wird der Fehler im Test-Set bestimmt. Im genutzten Ansatz werden die ersten 85% der Zeitreihe bis zum Start der Vorhersage als Trainingsset, die letzten 15% als Stop-Set genutzt. Als Test-Set dienen die anschließend in Abhängigkeit des Vorhersagehorizonts

definierten 1, 4 oder 12 Werte (Wochen-, Monats-, Jahreszeitenvorhersage). Die Details werden in Abschnitt 3.3.4 erläutert.

Hyperparameteroptimierung

Als Hyperparameter (HP) werden üblicherweise Parameter bezeichnet, welche vor dem Training festgelegt werden müssen und einen Einfluss auf das Trainingsergebnis haben. Hierzu zählen im Falle von NARX beispielsweise die Festlegung der Input- und Feedbackdelays oder auch die Anzahl der verdeckten Neuronen. Eine Optimierung von HPs kann per Rastersuche erfolgen. Dabei werden alle Kombinationen im festgelegten HP-Wertebereich getestet und ein Optimum gesucht. Diese Suche ist extrem rechenintensiv, zeitaufwendig und wird schnell unpraktikabel, wenn die Anzahl der möglichen Kombinationen eine entsprechende Größe erreicht. Aus diesem Grund wird üblicherweise, so auch in vorliegendem Fall, eine bayesische Hyperparameteroptimierung durchgeführt. Hierbei handelt es sich um eine globale Optimierung, für die ein probabilistisches Modell zwischen HPs und Zielmetrik aufgebaut wird. Berücksichtigt werden sowohl Bereiche im Hyperparameteraum, über die wenig bekannt ist (Exploration), als auch Wissen aus gut bekannten Bereichen genutzt (Exploitation). Über eine iterative Optimierung wird somit eine HP-Kombination vorhergesagt, für die am wahrscheinlichsten ein optimales Trainingsergebnis erwartet wird. Als Abbruchkriterien für die Optimierung dienen entweder die Anzahl der Iterationen (hier: 200) oder die Zeit (hier: max. 1h Rechenzeit pro Modell). Die relative Verbesserung der Zielfunktion je Iteration ist als Abbruchkriterium ungeeignet, da der Optimierungsalgorithmus immer wieder unbekannte Bereiche im HP-Raum testet (Exploration) und somit keine stetige Verbesserung der Zielfunktion erfolgt. Die bayesische Hyperparameteroptimierung kann lediglich auf eine zu minimierende Zielmetrik angewandt werden. Oft wird als Zielfunktion der MSE (Mean Squared Error) verwendet, im vorliegenden Fall wurde eine Kombination aus modifizierter Nash-Sutcliffe Efficiency (NSE) und Bestimmtheitsmaß (R^2) verwendet (s.a. Abschnitt 3.3.3):

$$\text{Zielfunktion} = (-1 \cdot \text{NSE}) - R^2 \quad (3.1)$$

Um die in Abschnitt 3.3.2 angesprochenen Initialisierungseffekte einzubeziehen, wurde jede getestete HP-Kombination für fünf pseudozufällig initialisierte Modelle betrachtet. Für die Optimierung im Allgemeinen wurde jede Zeitreihe in 4 unterschiedlich große Abschnitte aufgeteilt. Die ersten 90% der Zeitreihe werden für Training und Optimierung verwendet. Dabei werden 70% dieses Abschnittes (63% ges.) zum Training des NARX-Modells verwendet. Die restlichen 30% (27% ges.) dienen zu gleichen Teilen als Stop- und Test-Set für das Training bzw. die bayesische Optimierung. Die zu minimierende Zielmetrik der Optimierung wird für genanntes Test-Set berechnet, welches deshalb nachfolgend nur noch als Optimierungs-Set (Opt) bezeichnet wird. Um die Generalisierungsfähigkeit der nach Abschluss der Optimierung gefundenen Hyperparameterkombination beurteilen zu können, werden die letzten 10% der Gesamtzeitreihe als unabhängiges Test-Set herangezogen. Diese zusätzliche Unterteilung ist notwendig, da das Optimierungs-Set zwar nicht für das Training der Modelle, aber eben in die iterative Optimierung einfließt. Man spricht dabei von einer sog. Daten-Leckage (data leakage). Abbildung 3.5 fasst die Aufteilung der Zeitreihen zusammen.

Neben der Auswahl der Inputparameter T, Tsin und rH, wurden im vorliegenden Fall folgende Hyperparameter optimiert: IDmax, FDmax und h. Erstere stellen die Anzahl

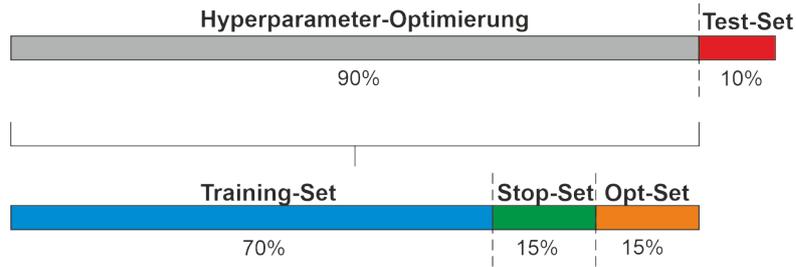


Abbildung 3.5: Aufteilung der Zeitreihe für die bayesische Hyperparameteroptimierung sowie die Beurteilung der resultierenden Vorhersagequalität.

der Zeitschritte, die als Input- und Feedbackdelay benutzt werden, dar und wurden für jeden Inputparameter getrennt im Bereich von jeweils 1 bis 52 optimiert. Der Maximalwert entspricht dabei also genau einem Jahr und wurde gewählt, um die Komplexität des KNN zu begrenzen. Ab diesem Wert werden in der Regel allein aufgrund der Autokorrelation der Zeitreihe wieder höhere Relevanz erwartet, weshalb dann bevorzugt hohe ID/FD Werte gewählt werden und so die Modellkomplexität ansteigt. Der Parameter h beschreibt die Anzahl Neuronen in der verdeckten Schicht des NARX und wurde auf 20 limitiert, ebenfalls um die Modellkomplexität nicht beliebig ansteigen zu lassen.

3.3.3 Fehlerkriterien zur Beurteilung der Vorhersagegüte

Zur Beurteilung der Vorhersagegüte kommen verschiedene Fehlerkriterien zum Einsatz: Nash-Sutcliffe Efficiency (NSE), Bestimmtheitsmaß (R^2), Root Mean Squared Error - absolut und relativ (RMSE, rRMSE), Bias - absolut und relativ (Bias, rBias) und Persistence Index (PI, PIop). Diese werden im Folgenden genauer erläutert. Jedes Fehlerkriterium hat Vor- und Nachteile und betrachtet andere Aspekte der Vorhersagegüte, keines lässt für sich allein genommen eine umfassende Beurteilung zu. Aus diesem Grund müssen immer mehrere Fehlerkriterien simultan berechnet und beurteilt werden (Gupta et al. 2009; Krause et al. 2005; Ritter und Muñoz-Carpena, 2013).

NSE nimmt Werte zwischen $-\infty$ und 1 an und beschreibt, ob eine Vorhersage besser ist als der Mittelwert ($NSE > 0$). Üblicherweise ist damit der Mittelwert der beobachteten Werte im Vorhersagezeitraum gemeint (Krause et al. 2005). Davon wird an dieser Stelle abgewichen und stattdessen der Mittelwert der Gesamtzeitreihe verwendet um (i) eine NSE Berechnung für einen Vorhersagehorizont von nur einem Zeitschritt (Wochenvorhersage) zu ermöglichen und (ii) dem operationellen Gedanken des NSE zu genügen (ist die Vorhersage besser als der zum Zeitpunkt des Vorhersagestarts bekannte Mittelwert, also besser als eine primitive Alternativvorhersage). Bei isolierter Betrachtung von einzelnen Vorhersagen mag der herkömmlich definierte NSE die Vorhersagequalität u.U. besser wiedergeben, gerade bei der gemischten Betrachtung vieler Einzelvorhersagen, um ein Gesamturteil über die Vorhersagequalität abzugeben, werden auf diese Weise jedoch schlechte NSE Werte übergewichtet, da der Wertebereich des NSE nach unten nicht begrenzt ist. Die im vorliegenden Fall genutzte Abweichung vom Mittelwert der Zeitreihe entspricht der (empirischen) Varianz der Zeitreihe (s^2).

$$NSE = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2}{\frac{1}{m} \sum_{i=1}^m (o_i - \bar{o})^2} = 1 - \frac{\frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2}{s^2} \quad (3.2)$$

o_i : beobachteter Wert, \bar{o} : Mittelwert beobachtet, p_i : vorhergesagter Wert, n : Anzahl vorhergesagter Werte, m : Anzahl Werte pro Zeitreihe, s : Standardabweichung der beobachteten Zeitreihe

Das Bestimmtheitsmaß kann als quadrierte Pearson-Korrelation (r) berechnet werden und nimmt Werte zwischen 0 (keine Korrelation) und 1 (optimale Korrelation) an. Es beschreibt den durch die Vorhersage beschriebenen Anteil an der Gesamtvarianz und berechnet sich wie folgt (Krause et al. 2005):

$$R^2 = \left(\frac{\sum_{i=1}^n (o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^n (o_i - \bar{o})^2} \sqrt{\sum_{i=1}^n (p_i - \bar{p})^2}} \right)^2 \quad (3.3)$$

\bar{p} : Mittelwert vorhergesagt

Der RMSE ist die Wurzel der mittleren quadratischen Abweichung. Eine Normierung (rRMSE) erfolgt mittels maximaler Spannweite der Zeitreihe und ermöglicht somit den Vergleich zwischen unterschiedlichen Zeitreihen. Der Wertebereich des RMSE liegt zwischen 0 (Optimum) und ∞ und trägt im vorliegenden Fall die Einheit [m] bzw. % (rRMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2} \quad (3.4)$$

$$rRMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{o_i - p_i}{o_{max} - o_{min}} \right)^2} \quad (3.5)$$

o_{max} : maximaler beobachteter Wert, o_{min} : minimal beobachteter Wert

Der Bias ist der mittlere Fehler einer Vorhersage, unterliegt keinem Wertebereich ($-\infty$ bis ∞ , Optimum = 0) und beschreibt systematische Abweichungen nach oben oder unten. Eine Normierung erfolgt analog zum rRMSE mittels Spannweite der Zeitreihe. Der Bias trägt im vorliegenden Fall die Einheit [m] bzw. % (rBias):

$$Bias = \frac{1}{n} \sum_{i=1}^n (o_i - p_i) \quad (3.6)$$

$$rBias = \frac{1}{n} \sum_{i=1}^n \left(\frac{o_i - p_i}{o_{max} - o_{min}} \right) \quad (3.7)$$

Der Persistency Index (Kitanidis und Bras, 1980; Kong A Siou et al. 2011) berechnet sich analog zu NSE, verwendet aber nicht den Mittelwert der Zeitreihe, sondern eine naive Vorhersage (auch: no-model Vorhersage) als Vergleich. Im Falle von PI ist damit der Wert gemeint, der f -Werte vor dem betrachteten Wert liegt, wobei f die Länge des Vorhersagehorizontes ist. Analog zum NSE nimmt der PI daher Werte zwischen $-\infty$ und 1 an, Werte größer 0 beschreiben Vorhersagen, die besser sind als der Vergleichswert. PIop ist erneut eine Modifikation, die dem operationellen Vorhersagegedanke Sorge trägt, und überprüft, ob die Vorhersage besser ist, als es der letzte beobachtete Wert gewesen wäre. PI ist in Abhängigkeit von f für Wochen-, Monats- und Jahreszeitenvorhersagen als PI1, PI4 und PI12 bezeichnet. Die Nomenklatur von PIop ist analog, auch wenn sich

rein formal die Berechnung für unterschiedliche f nicht unterscheidet. Für $f = 1$ gilt zudem $PI1 = PI1op$.

$$PI = 1 - \frac{\sum_{i=1}^n (o_i - p_i)^2}{\sum_{i=1}^n (o_i - o_{i-f})^2} \quad (3.8)$$

f : Vorhersagehorizont

$$PIop = 1 - \frac{\sum_{i=1}^n (o_i - p_i)^2}{\sum_{i=1}^n (o_i - o_{last})^2} \quad (3.9)$$

o_{last} : letzter beobachteter Wert vor Vorhersagestart

3.3.4 Vorhersageevaluierung

Generell erfolgt die Evaluierung der Vorhersagegüte in zwei Schritten. Zunächst werden Vorhersagen auf Basis von HYRAS-Daten berechnet, um so die maximal erreichbare Vorhersagegüte für optimale Inputdaten beurteilen zu können. Im zweiten Schritt werden als Vorhersageinputs EPISODES-Hindcasts genutzt und so die Unsicherheit der Inputdaten simuliert, wie sie im operationellen Betrieb auftauchen würde. Weiterhin erfolgt die Evaluierung der Vorhersagequalität in Abhängigkeit des Vorhersagehorizontes und nutzt die in Abschnitt 3.3.3 vorgestellten Fehlerkriterien.

Da operationelle Jahreszeitenvorhersagen auf EPISODES Daten basieren und für diese nur ein Modellstart pro Monat erfolgt, wird auch bei der Evaluierung der Vorhersagequalität auf Basis beobachteter Daten (HYRAS) von einer Vorhersage pro Monat ausgegangen (jeweils 3 Monate lang). Als Evaluierungszeitraum für die Jahreszeitenvorhersagen dienen die vollständigen Jahre 2006-2015. Es werden pro Vorhersagestart 10 unterschiedliche Modellinitialisierungen gerechnet, deren Median als Vorhersage des entsprechenden 3-Monatszeitraumes betrachtet wird. Die Fehlerkriterien berechnen sich als Median aller 3-Monatszeiträume. Für jede Zeitreihe werden mindestens 50% der Zeitreihenlänge für das Training des Modells genutzt. In seltenen Fällen (Zeitreihe beginnt nach 1996) verkürzt sich dadurch der Evaluierungszeitraum geringfügig. Werden die 10 Jahre voll ausgeschöpft, werden demnach pro Zeitreihe ca. 1200 Einzelmodelle trainiert und getestet.

Aufgrund des damit einhergehenden großen Rechen- und Zeitaufwandes wurde der Evaluierungszeitraum für Wochen- und Monatsvorhersagen dementsprechend verkleinert. Für beide stehen zukünftig mutmaßlich andere Vorhersageprodukte des DWD (z.B. MOS-MIX Wettervorhersagen für bis zu 10 Tage) zur Verfügung, die häufiger Vorhersagen liefern, als einmal monatlich. Daher wurden für den 4-Jahreszeitraum 2012-2015 wöchentlich Grundwasserstandsvorhersagen mit jeweils fünf pseudozufälligen Modellinitialisierungen berechnet. Damit ergibt sich eine Gesamtzahl der trainierten Modelle von ca. 1000 und liegt damit in einer ähnlichen Größenordnung wie für die 10-Jahresevaluierung der Jahreszeitenvorhersagen. Ein direkter Vergleich der Vorhersagegüten der Jahreszeitenvorhersagen mit den Wochen- und Monatsvorhersagen ist damit also nicht möglich, da sich diese auf verschiedene Zeiträume beziehen (wobei z.B. einzelne Jahre aufgrund der Witterung schwieriger vorherzusagen sind als andere). Tabelle 3.1 fasst die Eckpunkte der Evaluierung je Vorhersagezeitraum noch einmal zusammen.

Tabelle 3.1: Evaluierungsschemata für Wochen- (1 Step), Monats- (4 Steps) und Jahreszeitenvorhersagen (12 Steps)

	1 Step	4 Steps	12 Steps
Vorhersagehorizont	1 Woche	1 Monat	3 Monate
Eval.-Zeitraum	2012-2015	2012-2015	2006-2015
Startpunkte	wöchentlich	wöchentlich	monatlich
Initialisierungen pro Startpunkt	5	5	10
Gesamtzahl Modelle max.	1040	1015	1200

Für alle Startpunkte und Vorhersagehorizonte wurden sowohl open- als auch closed-loop trainierte Modelle erstellt und die im direkten Vergleich bessere Architektur ausgewählt. Ein identisches Schema wurde auf diese Auswahl dann auch für die Evaluierung der EPISODES-basierten Vorhersagen umgesetzt und im direkten Vergleich mit den HYRAS-basierten Vorhersagen der Performanceverlust ermittelt, welcher durch die Unsicherheiten der Klimavorhersagen entsteht.

3.4 Ergebnisse und Diskussion

Vor der Evaluierung der Vorhersageergebnisse wurde, wie in Abschnitt 3.3.2 beschrieben, eine messstellenspezifische Hyperparameteroptimierung durchgeführt (Anhang G). Tabelle 3.2 fasst einen Teil der Optimierungsergebnisse zusammen und zeigt auf, wie häufig welche Inputparameter bzw. deren Kombinationen gewählt wurden. Hierbei zeigt sich keine deutliche Systematik. Auch die räumliche Auswertung der Inputauswahl und deren Kombinationen ergab kein sichtbares Muster. Oft lagen viele Hyperparameterkombinationen bei der Optimierung in einem ähnlichen Gütebereich und die gewählten Hyperparameter unterliegen daher nicht dem Anspruch, die beste, sondern nur eine ausreichend gute Kombination darzustellen. Die Interpretierbarkeit der HP-Kombination im Einzelfall ist demnach begrenzt.

Tabelle 3.2: Häufigkeit der genutzten Inputparameter sowie deren Kombinationen als Ergebnisse der bayesischen Hyperparameteroptimierung.

Inputhäufigkeit der Einzelparam.		Inputkombinationen					
		1-2 Param.	3 Parameter	4 Parameter			
P	166 (alle)	(P)	1	(P, rH, T)	23	(P, rH, T, Tsin)	35
rH	95	(P, rH)	13	(P, rH, Tsin)	21		
T	108	(P, T)	20	(P, T, Tsin)	37		
Tsin	113	(P, Tsin)	16				

Dennoch sind gewisse Zusammenhänge erkennbar. Wie Abbildung 3.6 aufzeigt, scheint der Inputparameter relative Feuchte vorzugsweise für Messstellen mit niedrigen mittleren Grundwasserflurabständen wichtig zu sein. Dies stimmt mit dem intuitiven Verständnis der hydrogeologischen Prozesse überein, wonach davon ausgegangen werden kann, dass oberflächennahe Prozesse und Eigenschaften wie Evapotranspiration und Bodenfeuchte

mutmaßlich deutlich von der vergleichsweise variablen relativen Luftfeuchte abhängen können. Je größer der Flurabstand, desto weniger beeinflusst die relative Feuchte den generellen Prozess der Grundwasserneubildung. Dort spielt dann neben dem Grundwasserneubildungssignal vor allem die generelle (geglättete) Saisonalität bzw. auch eine übergeordnete, langjährige Periodizität eine Rolle (Liesch und Wunsch, 2019).

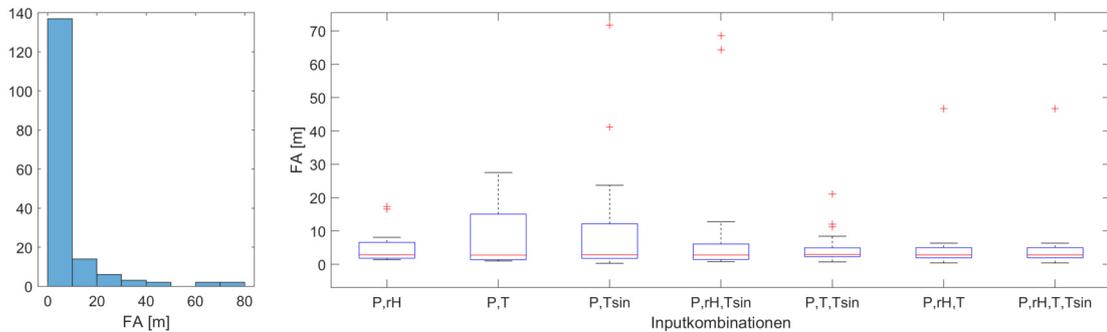


Abbildung 3.6: Zusammenhang zwischen mittlerem Flurabstand und den gewählten Inputkombinationen

3.4.1 Wochenvorhersagen

Gemäß dem in Abschnitt 3.3.4 beschriebenen Vorgehen wurden für die Evaluierung der Wochenvorhersagen die Jahre 2012-2015 betrachtet und wöchentlich mehrere pseudozufällig initialisierte Vorhersagen berechnet. Abbildung 3.7 gibt einen Überblick über relevante Fehlerkriterien aller Referenzmessstellen im entsprechenden Zeitraum, in Anhang H finden sind entsprechende Werte tabellarisch aufgeführt.

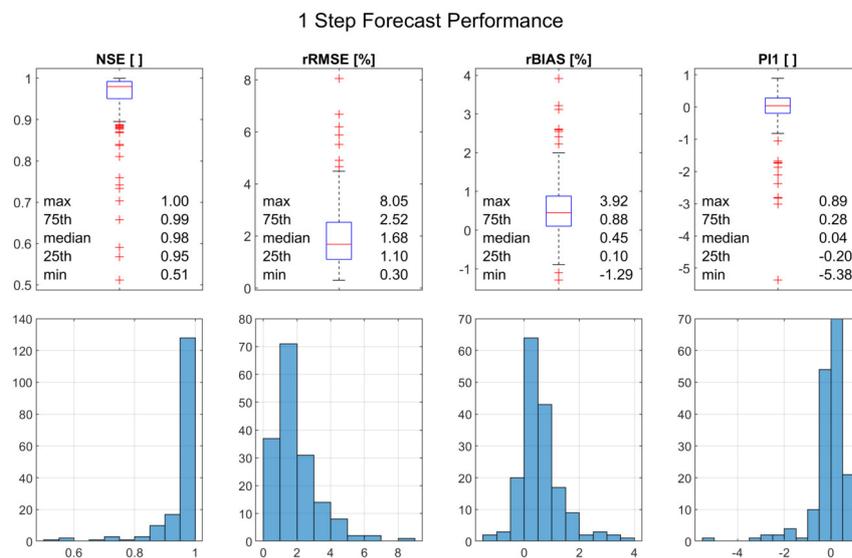


Abbildung 3.7: Fehlerwerte als Boxplots (oben) und Histogramme (unten) für Wochenvorhersagen aller Messstellen

Das Bestimmtheitsmaß lässt sich für Wochenvorhersagen (nur ein Zeitschritt) nicht be-

rechnen, absolute Werte von RMSE und Bias sind für einen Vergleich mehrerer Messstellen nicht geeignet und PI_{top} entspricht PI_1 . Diese Fehlerkriterien sind daher nicht abgebildet. Die Fehlerkriterien NSE, $rRMSE$ und $rBias$ sind zur Beurteilung von Wochenvorhersagen nahezu unbrauchbar, da für fast alle Messstellen eine sehr hohe Vorhersagegüte im Sinne dieser drei Kriterien erreicht wird. Dies erfolgt allein durch das regelmäßig wiederkehrende (wöchentliche) Training und Neustarten der Vorhersagen wodurch der Verlauf der Ganglinie zu einem großen Teil nachempfunden und eine Güte erreicht wird, die kaum in Zusammenhang mit der Güte der eigentlichen Vorhersagen steht. Dieses Problem tritt jeweils zunehmend abgeschwächt auch für Monats- und Jahreszeitevorschagen auf. Als aussagekräftigstes Fehlerkriterium erweist sich daher PI_1 (bzw. PI_{top}). Werte größer 0 bedeuten, dass der vorhergesagte Wert p_i näher am tatsächlichen Wert o_i liegt, als der zum Zeitpunkt der Vorhersage letzte verfügbare beobachtete Wert o_{i-1} (no-model forecast). Demzufolge sind im Mittel von 166 Referenzmessstellen für 90 Messstellen (54%) ausreichend gute ($PI > 0$) Wochenvorhersagen berechnet worden. Über 50 Messstellen weisen im Mittel PI_1 -Werte nur knapp < 0 auf, für diese ist im Einzelfall zu prüfen, ob ausreichend gute Vorhersagen für bestimmte Monate erzielt werden können (s.u.). Nur für wenige Messstellen muss daher davon ausgegangen werden, dass nicht einmal abschnittsweise sinnvolle Vorhersagen berechnet werden konnten ($PI_1 < < 0$).

Aufgrund der guten Aussagekraft von PI_1 basiert auch der direkte Performancevergleich zwischen open- und closed-loop Modellen und damit die Auswahl der besseren Trainingsarchitektur auf diesem Fehlerkriterium. Abbildung 3.8 verdeutlicht am Beispiel der Messstelle BW_2004-569-8 das im closed-loop häufig auftretende Startwertproblem (a), welches durch die Anwendung eines open-loop Trainings (b) oft gut behoben werden kann. Entsprechend verbessern sich auch die Fehlerkriterien der beiden Vorhersagen (von (a) nach (b)). Eine Evaluierung der Vorhersagegüte der jeweils besseren Trainingsarchitektur (vgl. Abb. 3.8b) ist für alle Messstellen im elektronischen Anhang beigefügt. Für Wochenvorhersagen wurden fast ausschließlich Modelle in der open-loop Konfiguration ausgewählt (164×open-loop, 2×closed-loop), vor allem wegen des im Vergleich oft exakteren Vorhersagestartwerts (vgl. Abschnitt 3.3.2).

Bei den betrachteten Fehlerwerten in Abbildung 3.7 handelt es sich um Gesamtjahresmittel mehrerer Jahre, es ist jedoch davon auszugehen, dass die Vorhersagegenauigkeit im Jahresverlauf variabel ist. In Abbildung 3.9 sind daher die Fehlerwerte nach Startmonat der Vorhersage aufgesplittet. Wie zu erwarten sind im Jahresverlauf Performanceunterschiede erkennbar. PI_1 zufolge sind in den Monaten Dez./Jan. und Apr.-Aug. für den Betrachtungszeitraum im Mittel bessere Vorhersagen erzielt worden. Dies stimmt teilweise auch mit den Aussagen der restlichen drei Fehlerkriterien überein. Mit gewissem Spielraum deckungsgleich für alle vier Fehlerkriterien ist zudem die generelle Tendenz zu besseren Vorhersagen im Hochsommer. Trotz der begrenzten Aussagekraft von NSE, $rRMSE$ und $rBias$ für die Gesamtbeurteilung der Vorhersagequalität von Wochenvorhersagen über alle Messstellen sind diese doch gut geeignet, um die Variabilität der Vorhersagegenauigkeit über die einzelnen Startmonate zu illustrieren. Generell sind die Fehlerkriterien NSE, $rRMSE$ und $rBIAS$ recht gut untereinander korreliert, die Korrelation zu PI_1 ist jedoch eher gering (Abbildung 3.10). Vor allem für $rBIAS$ und $rRMSE$ existieren keine Grenzwerte, die zur pauschalen Beurteilung der Güte herangezogen werden könnten. Sie eignen sich vor allem dazu, die Eigenschaften der generierten Vorhersagen genauer einordnen zu können.

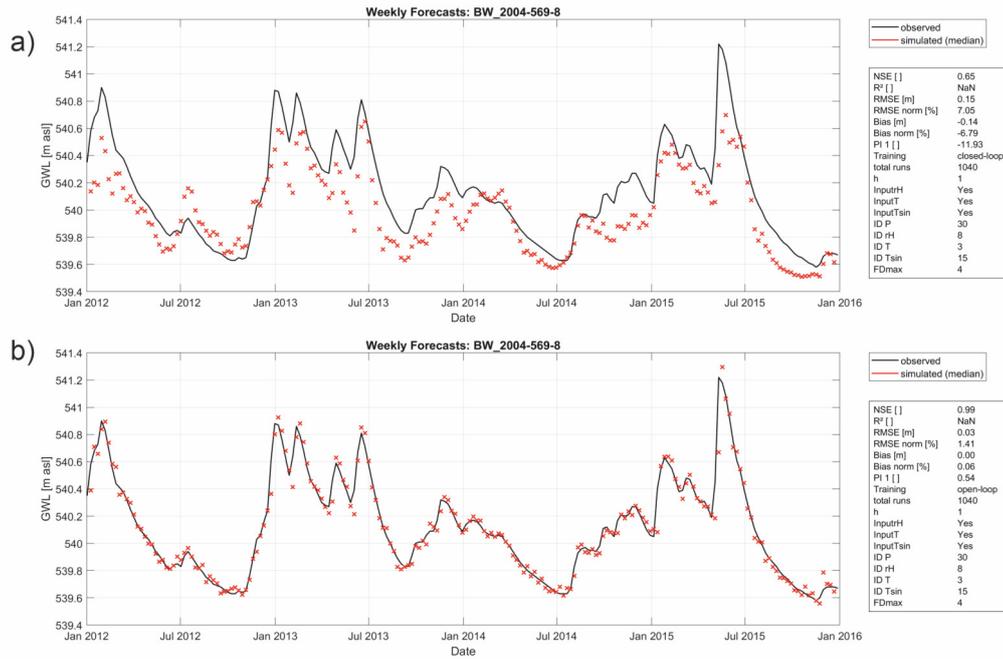


Abbildung 3.8: Vergleich closed-loop (a) und open-loop (b) für Wochenvorhersagen am Beispiel der Messstelle BW_2004-569-8

Durch den Vergleich der Fehlerkriterien wird deutlich, weshalb prinzipiell mehrere Fehlerkriterien parallel zu betrachten sind (s.a. Abs. 3.3.3) und warum aber gleichzeitig eine allgemeine Interpretation der Ergebnisse zumindest bei Kombination der Kriterien gewisse Schwierigkeiten verursacht. Im Zweifelsfall ist immer eine Einzelbetrachtung und Beurteilung von Messstellen vorzuziehen. Dennoch gibt vor allem der PI1-Boxplot in Abb. 3.7 einen guten Eindruck über die Vorhersagequalität der Wochenvorhersagen an allen Referenzmessstellen.

Die Abbildungen 3.11 und 3.12 geben einen räumlichen Überblick über die Vorhersagegüte der Wochenvorhersagen basierend auf PI1. In Abb. 3.11 entspricht die Größe der Kreise der Anzahl an Clustermessstellen, die den jeweiligen Referenzmessstellen zugeordnet sind, die Farbe gibt die Vorhersagegüte wieder. In Abb. 3.12 wird die Vorhersagegüte auf alle zugehörigen Clustermessstellen übertragen und gibt einen guten Eindruck über die zu erwartenden regionalen Unterschiede bezüglich der Vorhersagegüte sowie die räumliche Abdeckung des Gesamtgebiets. Beiden Abbildungen zufolge konzentrieren sich ungenügende Vorhersagegüten auf die Mitte und den Osten Deutschlands sowie auf einen Ballungsbereich im Ruhrgebiet. Gute und sehr gute Güten finden sich vor allem im Nordwesten, dem Oberrheingraben und dem äußeren Süden Deutschlands. Zur optischen Beurteilung der Güte werden für beide Kartendarstellungen alle PI1-Werte in drei Klassen eingeordnet: <0 , ≥ 0 und $<0,5$ sowie $\geq 0,5$. Für die erste Klasse (<0) ist im Mittel kein Mehrwert im Vergleich zu einer no-model Vorhersage zu erwarten, für die beiden anderen Klassen jedoch schon. Eine Unterteilung wurde lediglich zur optischen Abgrenzung besonders guter Wochenvorhersagen ($>0,5$) durchgeführt. Für die Klasse $PI < 0$ ist im Einzelfall der Jahresgang zu unterscheiden, für viele Messstellen ist eine gute Vorhersagegüte für einige Monate denkbar.

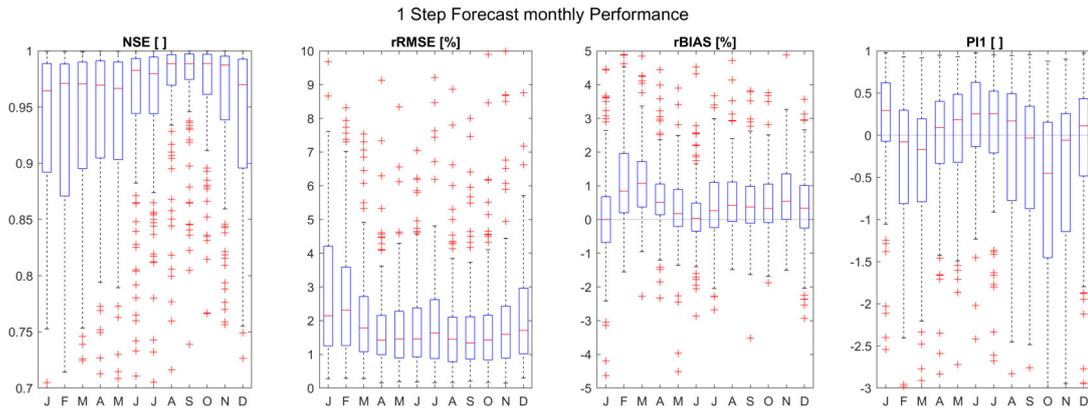


Abbildung 3.9: Gesamtüberblick der Fehler für Wochenvorhersagen je Startmonat der Vorhersage.

Dies wird deutlich, wenn man sich beispielhaft den Verlauf der Vorhersagegüte (PI1) für die Messstelle BB_34426014 genauer betrachtet (Abb. 3.13). Der Median aller Monate liegt unter 0, dennoch ist für mehrere Monate im Schnitt eine brauchbare Wochenvorhersage zu erwarten (v.a. Apr., Sep., Okt., Dez.) (Abb. 3.13b). In den Übersichtskarten bzw. der Gesamtstatistik wird diese Messstelle jedoch trotzdem negativ eingestuft. Um diese Erkenntnisse über die monatliche Vorhersagegüte zu untermauern, wäre eine größere Stichprobengröße wünschenswert, da jeder Boxplot nur noch etwa ($<$)16 Werte repräsentiert und die Abhängigkeit von Einzelereignissen damit recht hoch sein dürfte.

Abbildung 3.14 überträgt diese Erkenntnis auf alle betrachteten Messstellen und zeigt die mittlere auf PI1-basierende Vorhersagegüte für den Monat Juni. Es wird deutlich, wie sich im direkten Vergleich zu Abb. 3.11 die Anteile der Klassen verschieben und sowohl deutlich mehr sehr gute Vorhersagen als auch weniger ungenügende Vorhersagen auftreten. Dies zeigt deutlich, welche enge Grenzen einer Gesamtbetrachtung über alle Messstellen und Zeiträume gesetzt sind. Analoge Kartendarstellungen zur Beurteilung der Vorhersagegüte von Wochenvorhersagen sind für ausgewählte Monate dem Anhang I beigefügt.

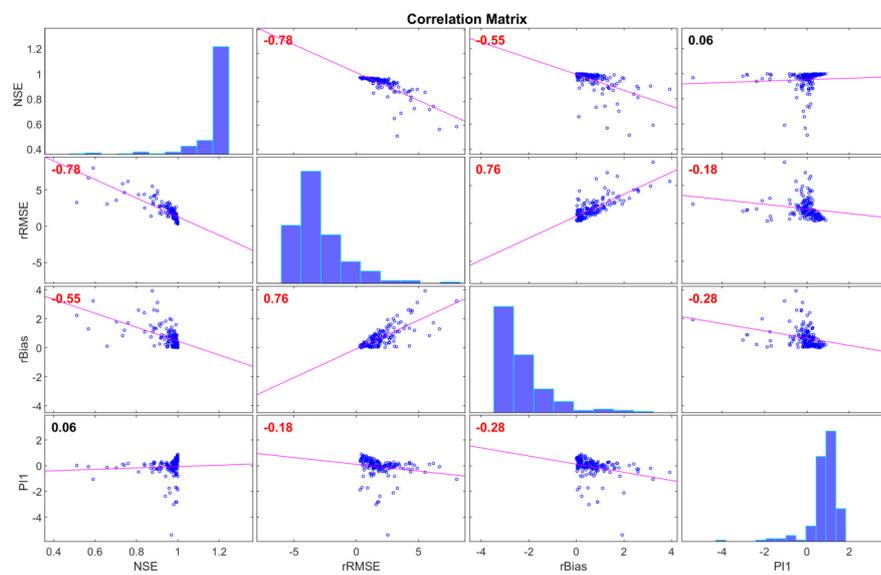


Abbildung 3.10: Übersicht über die Pearson-Korrelation verschiedener Fehlerkriterien für Wochenvorhersagen (rot: signifikant, schwarz: n.sign.). Die Analyse wurde für den Betrag von rBias anstelle der absoluten Werte durchgeführt.

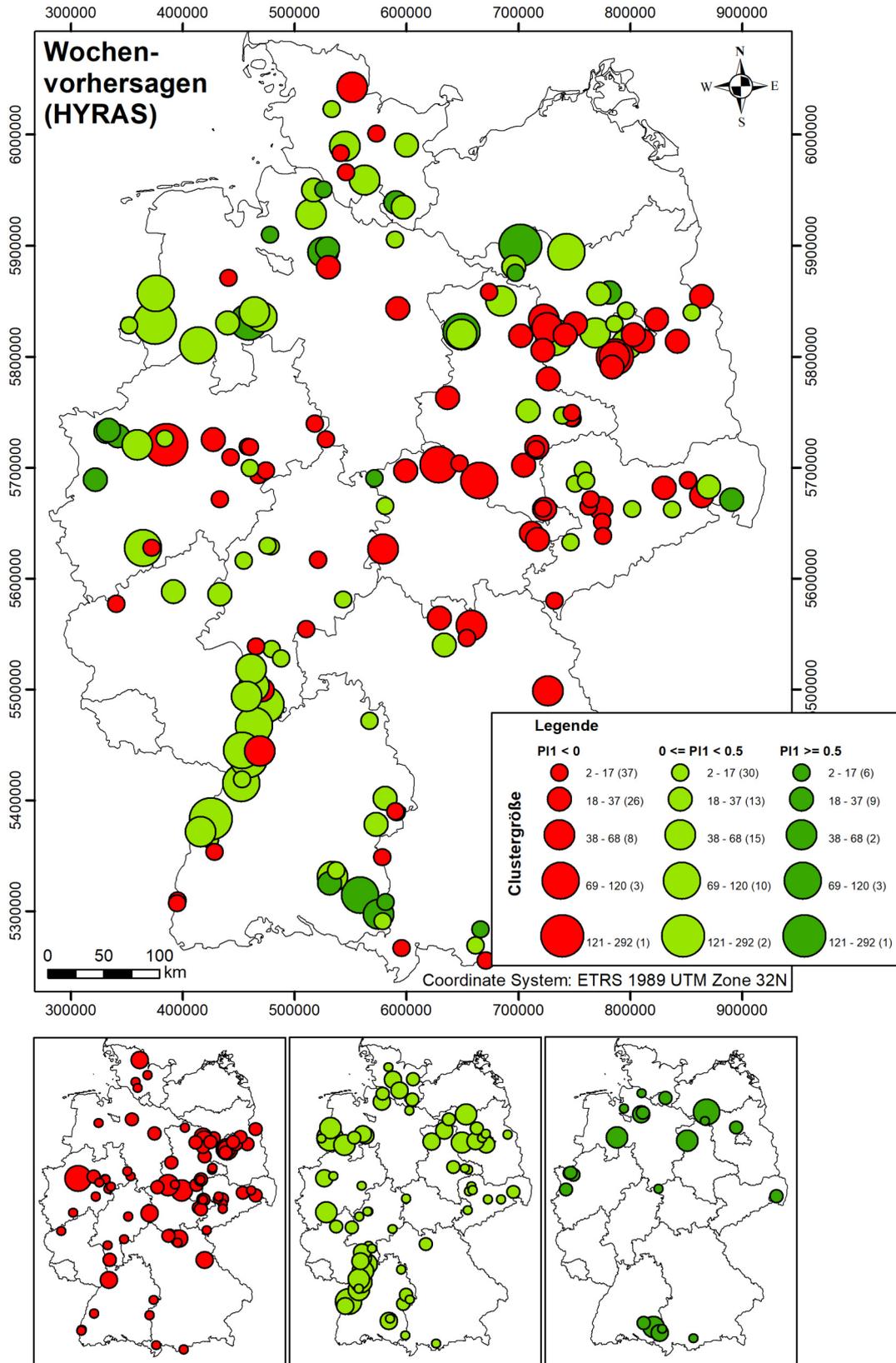


Abbildung 3.11: Übersichtskarte der Vorhersagegüte von Wochenvorhersagen anhand P11 (Farbe) im Evaluierungszeitraum (2012-2015) sowie die mit den RM verknüpfte Messstellenanzahl (Clustergröße, Kreisgröße)

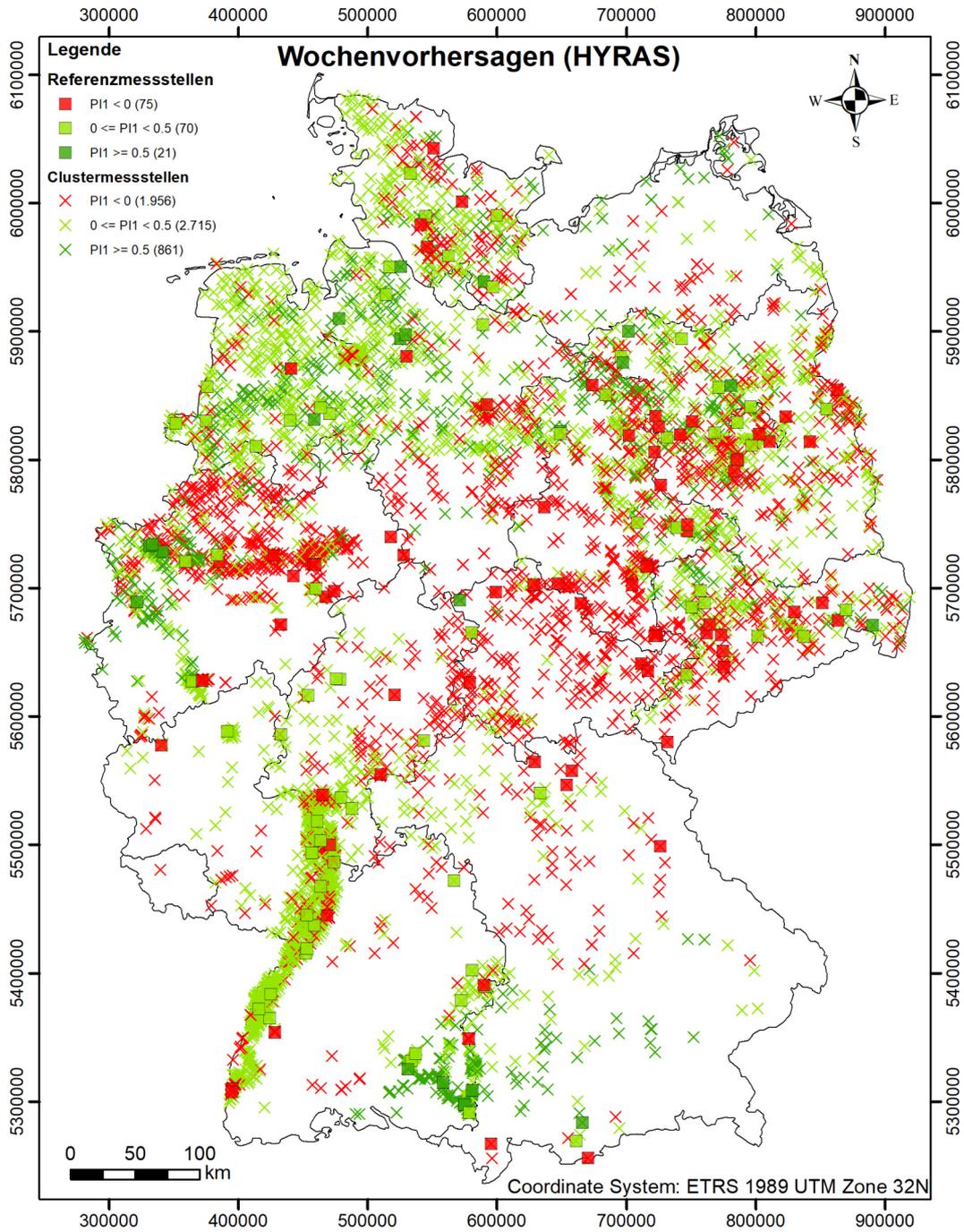


Abbildung 3.12: Übersichtskarte der Vorhersagegüte von Wochenvorhersagen anhand PI1 (Farbe) im Evaluierungszeitraum (2012-2015) sowie alle mit den RM (□) verknüpften Clustermessstellen (×)

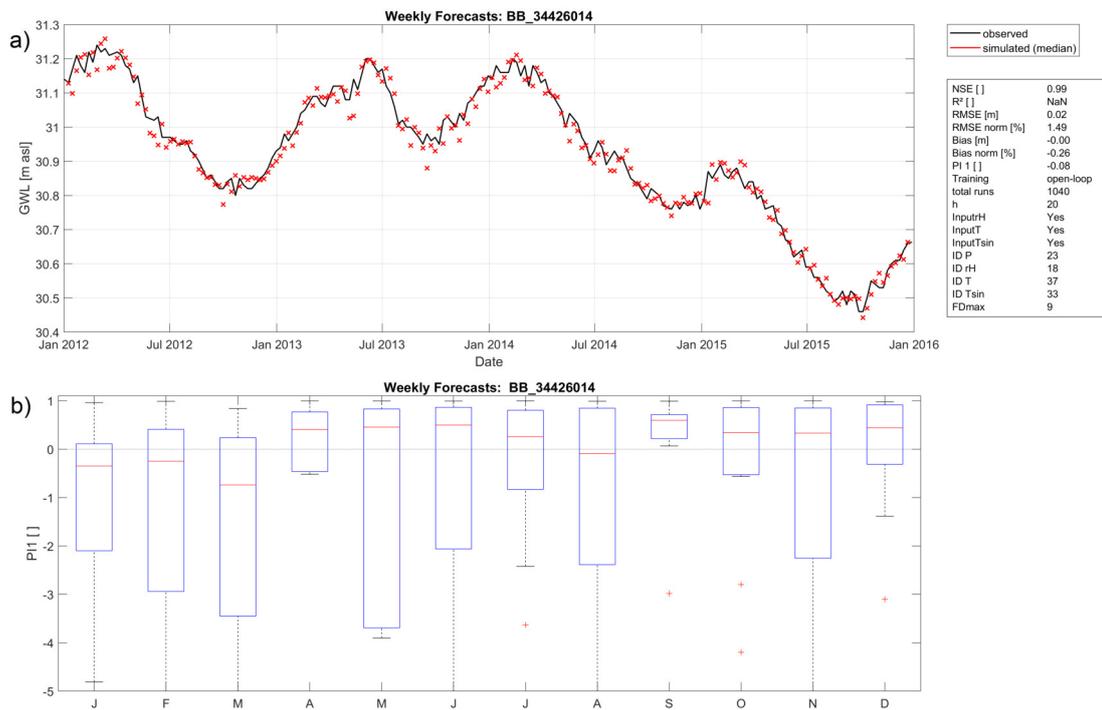


Abbildung 3.13: Wochenvorhersagen von Messstelle BB_34426014 (a) und PI1 nach Startmonat gesplittet (b)

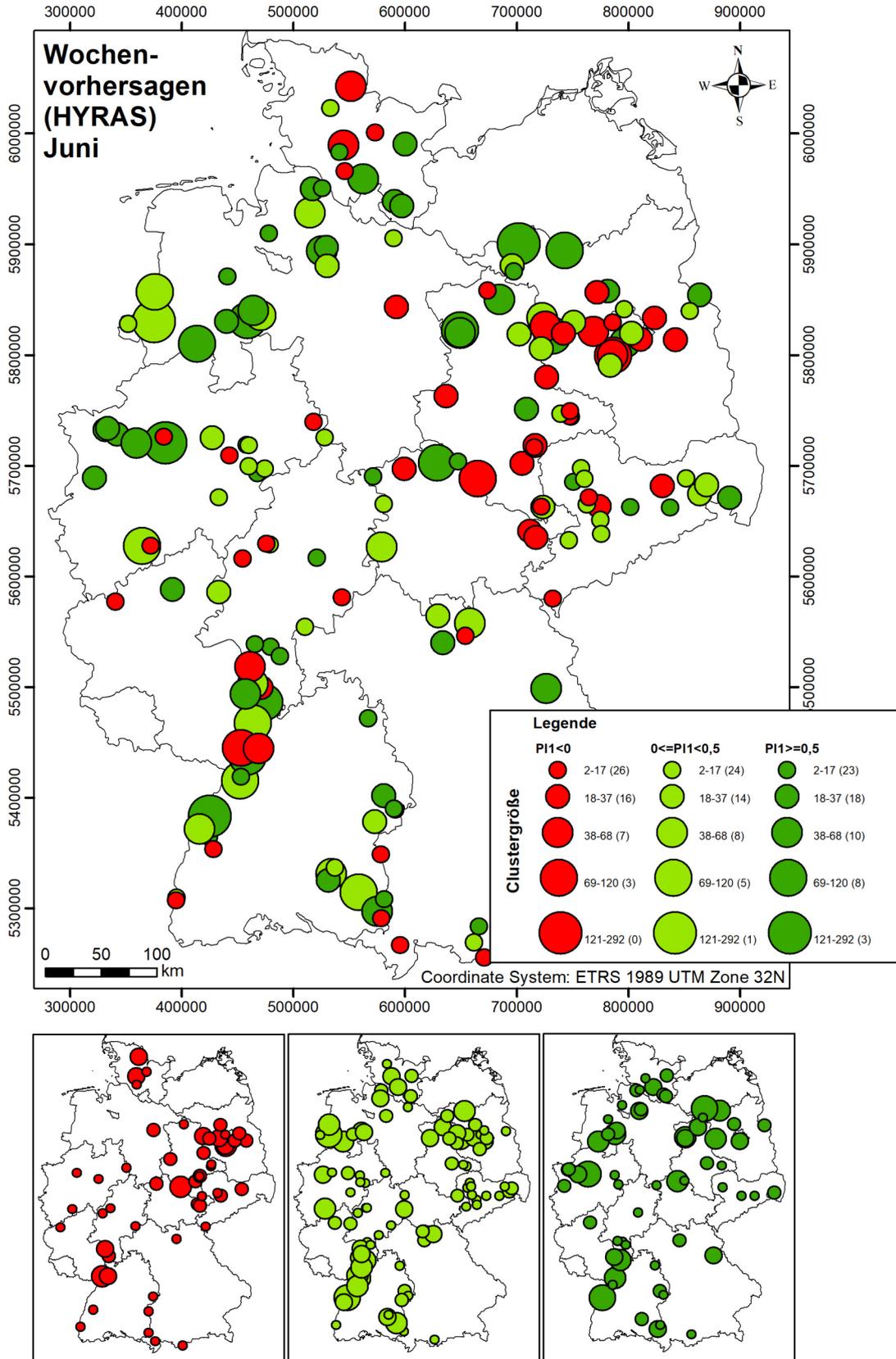


Abbildung 3.14: Übersichtskarte der mittleren Vorhersagegüte von Wochenvorhersagen im Juni anhand PI1 (Farbe) im Evaluierungszeitraum (2012-2015) sowie die mit den RM verknüpfte Messstellenanzahl (Clustergröße, Kreisgröße)

3.4.2 Monatsvorhersagen

Die Evaluierung der Monatsvorhersagen erfolgte analog der Evaluierung von Wochen-
vorhersagen. Betrachtet werden demnach die Jahre 2012-2015 sowie wöchentlich gest-
artete Monatsvorhersagen der Grundwasserstände (je 5 Initialisierungen). Jede Vorher-
sage überlappt daher mit den darauf folgenden Monatsvorhersagen um in der Regel
3 Zeitschritte. Abbildung 3.15 gibt einen Überblick über relevante Fehlerkriterien. Im
Vergleich zu Wochenvorhersagen sind nun auch das Bestimmtheitsmaß R^2 sowie eine
Unterscheidung in PI4 und PI4op dargestellt. RMSE und BIAS sind auch hier wieder
nur in relativer Form (rRMSE, rBias) dargestellt, da sich die absoluten Werte nicht für
einen Vergleich unterschiedlicher Messstellen eignen. Anhang H fasst alle Fehlerwerte
tabellarisch zusammen.

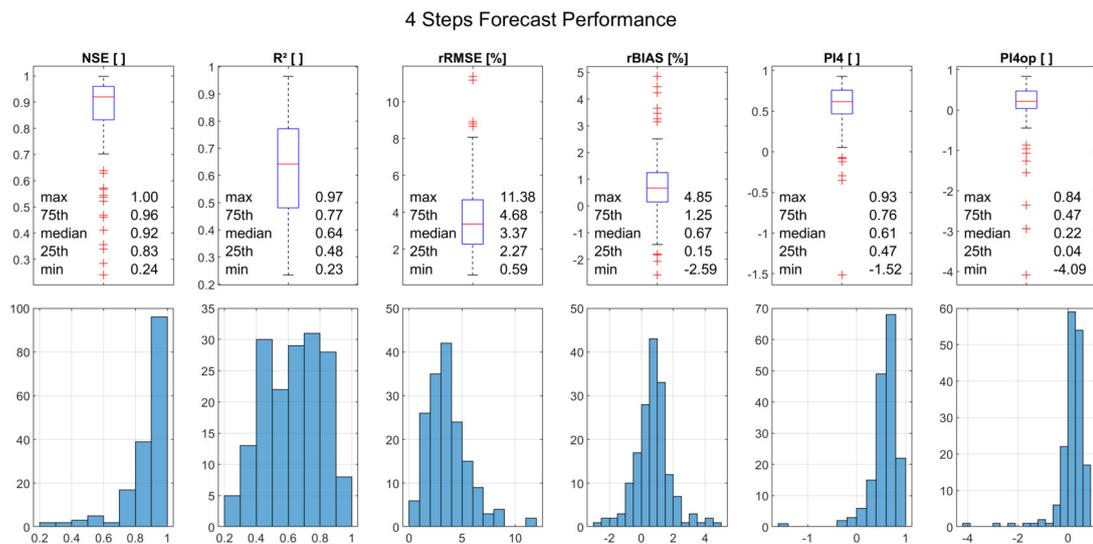


Abbildung 3.15: Fehlerwerte als Boxplots (oben) und Histogramme (unten) für Monats-
vorhersagen aller Messstellen

Auch für Monatsvorhersagen erweist sich PI4op als aussagekräftigster Fehlerwert, wes-
halb analog zum Vorgehen bei Wochenvorhersagen auf Basis von PI4op der direkte Ver-
gleich von open- und closed-loop Modellen durchgeführt wurde. Der Anteil von open-loop
Modellen ist nun etwas geringer (136×open-loop, 30×closed-loop).

Generell liefern Monatsvorhersagen bessere PI4op Werte als Wochenvorhersagen, dies
war jedoch vorab zu erwarten, da es mit größer werdendem Abstand zum letzten beob-
achteten Wert leichter wird, eine no-model Vorhersage zu schlagen ($PI4op > 0$). Konkret
weisen mehr als 75% aller Messstellen (>125) einen PI4op Wert größer Null auf. Die Ge-
samtinterpretation der Fehlerkriterien verursacht weiterhin gewisse Schwierigkeiten, da
die Fehlerwerte noch immer nicht genau mit dem visuellen Eindruck bei der Beurteilung
der Vorhersagegüte übereinstimmen. Die NSE-Werte verbleiben (zu) hoch, vor allem
verursacht durch den angesprochenen wöchentlichen Neustart der Vorhersagen (s.a. Ab-
schnitt 3.4.1). Dies bewirkt in der Regel eine Vergleichsweise hohe Güte im Vergleich zum
Gesamtmittelwert der Zeitreihe. Ein generelles Absinken der NSE-Werte im Vergleich
zu Wochenvorhersagen zeigt jedoch, dass nun nach und nach mit größer werdendem
Vorhersagehorizont ein realistischeres Bild (im Vergleich zur visuellen Beurteilung) der

Vorhersagegüte gezeichnet wird. Das für Monatsvorhersagen nun ebenfalls betrachtete Bestimmtheitsmaß ergänzt das Set an verfügbaren Fehlerkriterien sinnvoll, besitzt jedoch aufgrund des recht kurzen Vorhersagehorizontes noch eine eher begrenzte Aussagekraft. Dennoch weisen fast 75% der betrachteten Messstellen im Mittel ein R^2 von 0,5 oder größer auf. $rRMSE$ und $rBias$ (im Betrag) zeigen im Vergleich zu Wochenvorhersagen etwas größere Werte, $rRMSE$ dabei ein wenig deutlicher als $rBias$. Beide Werte bewegen sich jedoch noch immer in ähnlichen Wertebereichen.

Abbildung 3.16 illustriert die Korrelationen der Fehlerwerte untereinander. $PI4$ und $PI4op$ sind sehr stark korreliert, $PI4op$ ist jedoch aufgrund seines klareren Bezugspunktes leichter zu interpretieren. Generell sind viele der Fehlerwerte über mittlere Korrelationswerte verbunden, die Lage der Punkte im Scatterplot verdeutlicht jedoch die Schwierigkeit, welche bei einer kombinierten Betrachtung auftritt. So sind zum Beispiel sehr hohe NSE und sehr hohe R^2 -Werte gut korreliert, vor allem niedrige Werte des Einen bedingen jedoch keine niedrigen Werte des Anderen. Dies erschwert die Definition von kombinierten Grenzwerten für zufriedenstellende Vorhersagen über mehrere Kriterien hinweg enorm, da rein visuell in einigen Fällen ein einzelner niedriger Wert für eins von beiden Fehlerkriterien als akzeptabel eingestuft werden kann. Dieses Problem multipliziert sich mit der Anzahl der simultan einzubeziehenden Kriterien.

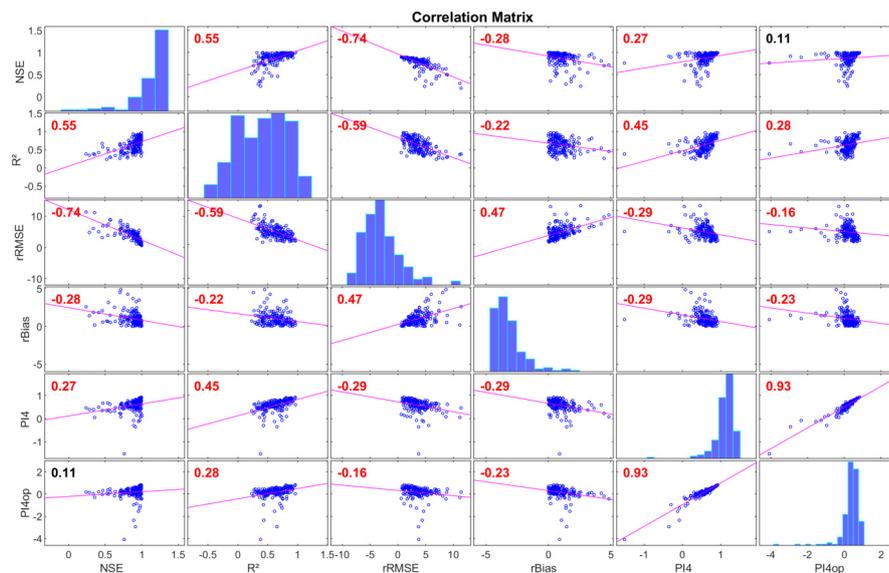


Abbildung 3.16: Übersicht über die Pearson-Korrelation verschiedener Fehlerkriterien für Monatsvorhersagen (rot: signifikant, schwarz: n.sign.). Die Analyse wurde für den Betrag von $rBias$ anstelle der absoluten Werte durchgeführt.

Wie für Wochenvorhersagen ist auch für Monatsvorhersagen eine variable Vorhersagegüte im Jahresverlauf zu erwarten. In Abbildung 3.17 sind daher die Fehlerwerte nach Startmonat der Vorhersage aufgesplittet. Es gilt zu beachten, dass Startmonat und Ausgagemonat (also der Monat für den die Vorhersage gemacht wird) für die meisten Vorhersageschritte nicht übereinstimmen sondern meist eine Überlappung mit dem Folgemonat existiert. Das sich ergebende Bild stimmt dennoch sehr gut mit dem der Wochenvorher-

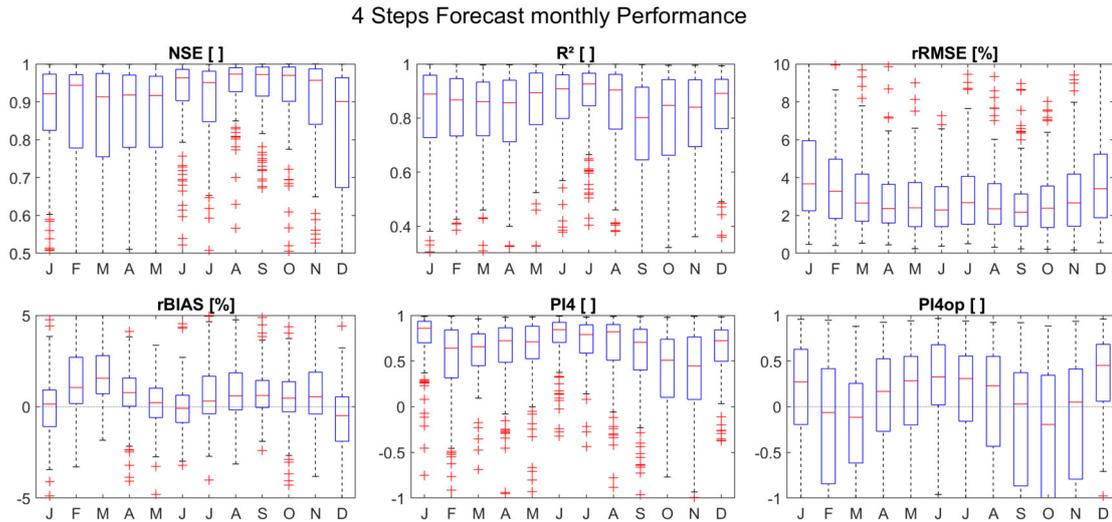


Abbildung 3.17: Gesamtüberblick der Fehler für Monatsvorhersagen je Startmonat der Vorhersage.

sagen überein. Mit gewissem Spielraum zwischen den Kriterien existiert weiterhin eine Tendenz zu besseren Vorhersagen in den Sommermonaten und zum Jahreswechsel.

Die Abhängigkeit der Vorhersagegüte vom Startmonat wirkt sich daher für Einzelmessstellen in ähnlicher Weise aus wie im Fall von Wochenvorhersagen, sodass trotz ungenügender Güte im Mittel abschnittsweise durchaus gute Ergebnisse möglich sind. Abbildung 3.18 zeigt erneut für Messstelle BB_34426014 die Monatsvorhersage selbst und zugehörige globale Fehlermetriken (a) sowie den Verlauf der Vorhersagegüte (anhand PI4op) gesplittet nach Startmonat (b). Wieder liegt zwar der Median von PI4op unter Null, vor allem für die Vorhersagen im April bis Juli sowie September sind jedoch durchaus hohe Güten zu erwarten. Eine Einzelfallbetrachtung der jeweiligen Messstellen ist daher sinnvoll und notwendig.

Die generell bessere Vorhersagegüte von PI4op im Vergleich zu PI1 (Wochenvorhersagen) zeigt sich auch optisch deutlich in der Übersichtskarte der Vorhersagegüte (Abb. 3.19). Analog zu Abb. 3.12 überträgt auch Abb. 3.20 die Güte der RM-Vorhersagen auf die zugehörigen Clustermessstellen. Es zeigt sich eine recht deutliche Konzentration ungenügender Vorhersagegüten auf den Osten und die Mitte Deutschlands. Bei der Betrachtung von Einzelmonaten ist auch hier eine deutliche Umverteilung zwischen den Klassen zu erwarten (vgl. Abb. 3.14). Entsprechende Karten sind für ausgewählte Monate Anhang I beigefügt.

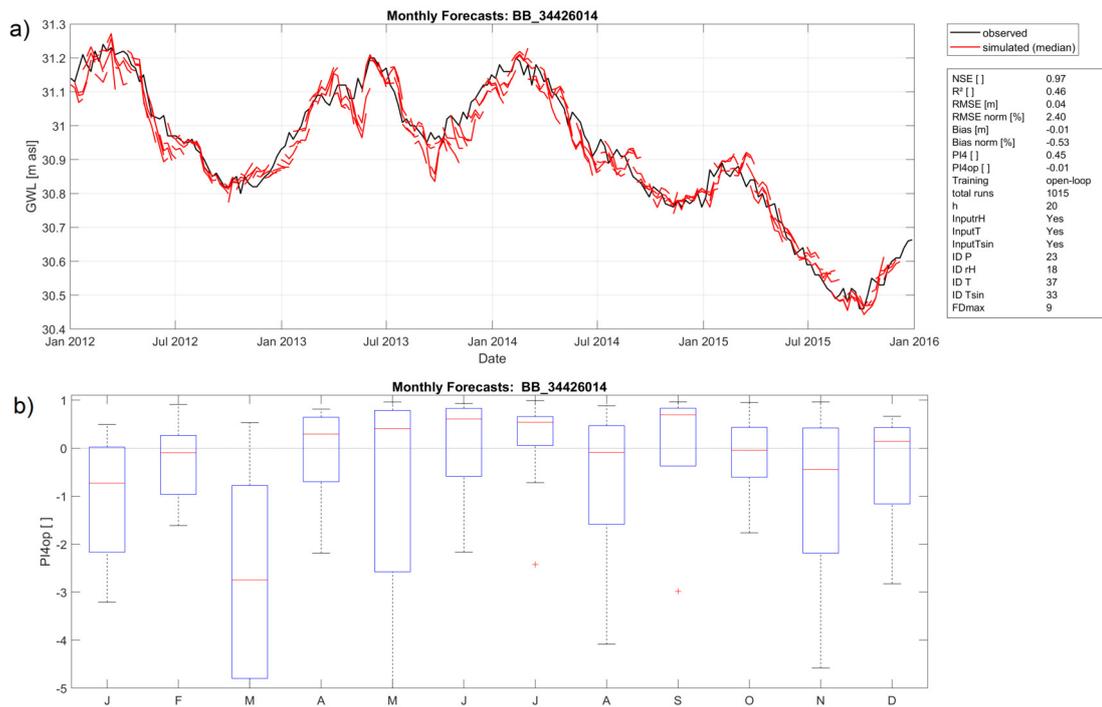


Abbildung 3.18: Monatsvorhersagen von Messstelle BB_34426014 (a) und PI4op nach Startmonat gesplittet (b)

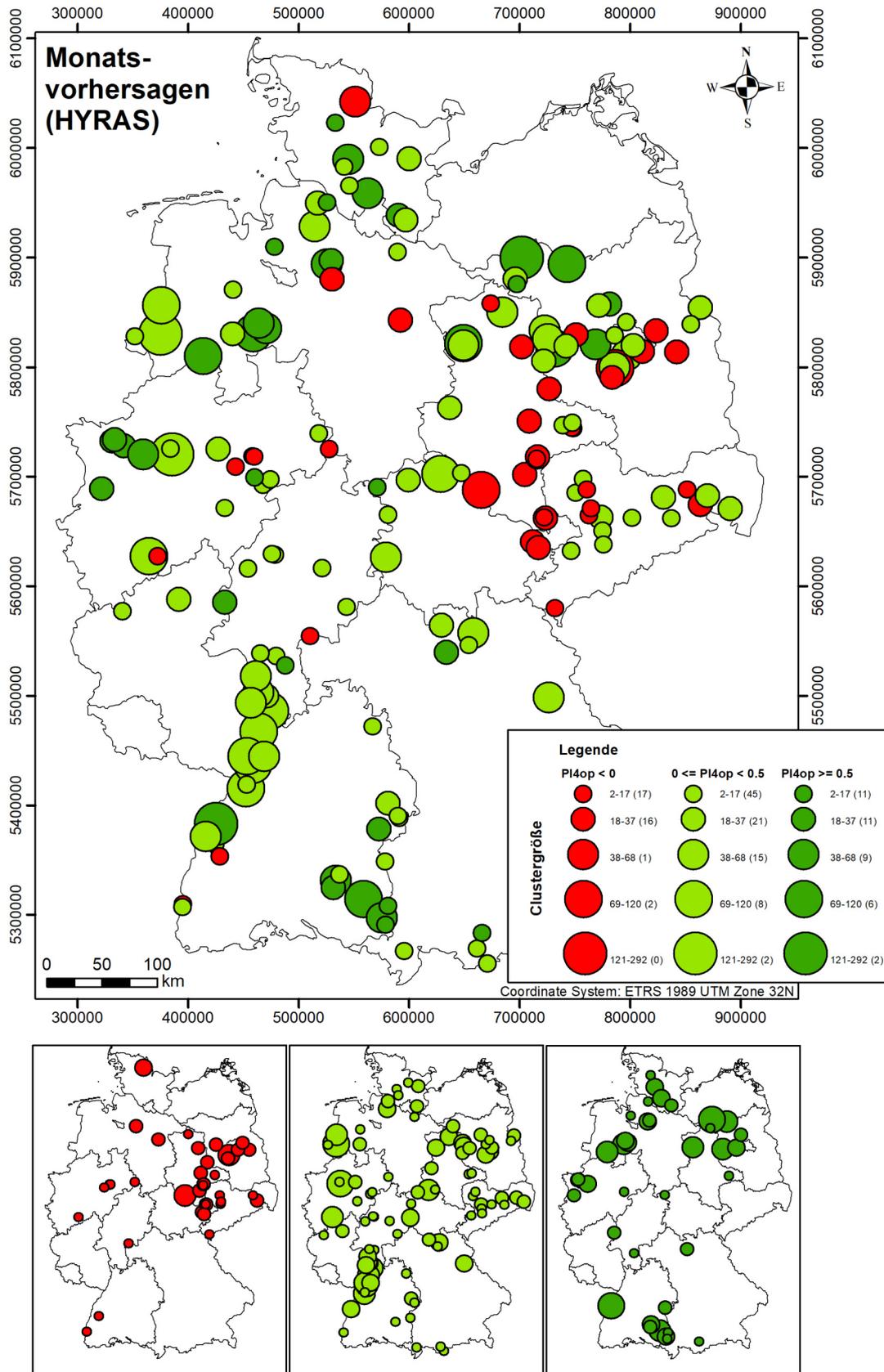


Abbildung 3.19: Übersichtskarte der Vorhersagegüte von Monatsvorhersagen anhand PI4op (Farbe) im Evaluierungszeitraum (2012-2015) sowie die mit den RM verknüpfte Messstellenanzahl (Clustergröße, Kreisgröße)

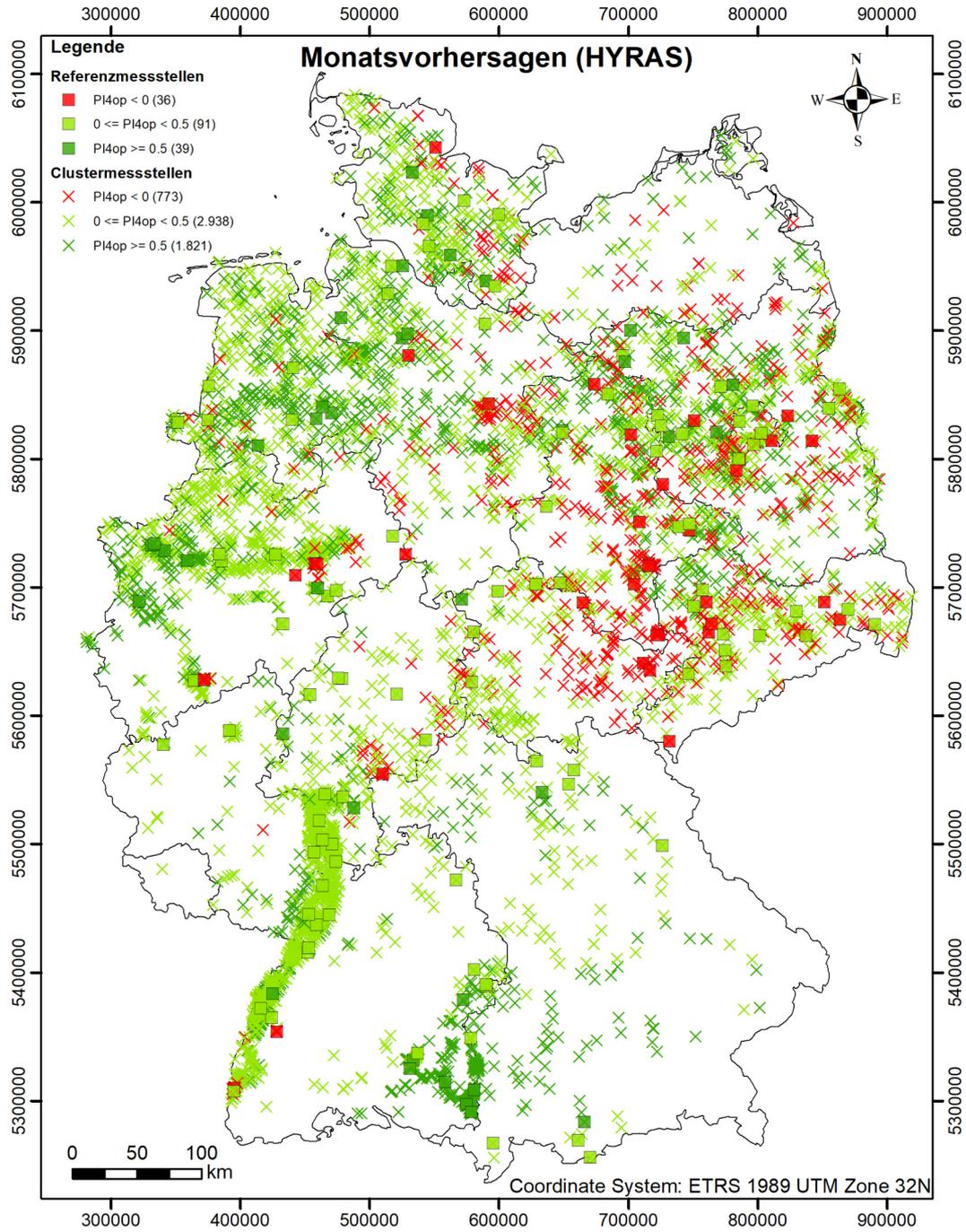


Abbildung 3.20: Übersichtskarte der Vorhersagegüte von Monatsvorhersagen anhand PI4op (Farbe) im Evaluierungszeitraum (2012-2015) sowie alle mit den RM(□) verknüpften Clustermessstellen (×)

3.4.3 Jahreszeitenvorhersagen

Der Evaluierungsansatz für Jahreszeitenvorhersagen unterscheidet sich deutlich von dem für Wochen- und Monatsvorhersagen. Da auch später operationell nur einmal im Monat neue Jahreszeitenvorhersagen der Input-Parameter zur Verfügung stehen werden, wurde die Güte von 12 Vorhersagen jährlich (monatlich, je 10 Initialisierungen) über 10 Jahre untersucht (2006-2015). Zur Beurteilung kommen die gleichen Fehlerkriterien wie für Monatsvorhersagen (außer PI12/PI12op statt PI4/PI4op) zum Einsatz (Abbildung 3.21, Anhang B). Mit weiter vergrößertem Vorhersagehorizont erfolgt eine erneute Annäherung zwischen visueller Beurteilung und den durch die Fehlerkriterien ausgedrückten Werten. Sowohl NSE als auch R^2 geben nun eine solide Vorstellung über die allgemeine Vorhersagequalität, jedoch erfasst PI12op erneut vor allem im Vergleich zum NSE die Güte am besten. Es ist hervorzuheben, dass sich die Fehler für den überwiegenden Teil der Messstellen bei allen Kriterien im guten bis sehr guten Bereich befinden. So weisen beispielsweise 75% aller untersuchten Messstellen NSE-Werte $> 0,72$, R^2 -Werte $> 0,52$ und PI12op-Werte $> 0,51$ auf. Für letzteres Kriterium gilt erneut, dass im Vergleich höhere Werte als für die bisherigen Vorhersagehorizonte zu erwarten waren, da es mit zunehmendem Abstand vom letzten beobachteten Wert zusehends leichter wird, diesen bezüglich des Fehlers zu unterbieten. Für die Beurteilung von PI12op wurde daher 0,5 als neuer Grenzwert für zufriedenstellende Vorhersagen definiert, welcher sich an gebräuchlichen Beurteilungen des eng verwandten NSE orientiert (Moriassi et al. 2007; Moriassi et al. 2015). Aus diesem Grund wird auch in den folgenden Kartendarstellungen der Farbcode für Werte zwischen 0 und 0,5 angepasst und von hellgrün auf gelb geändert.

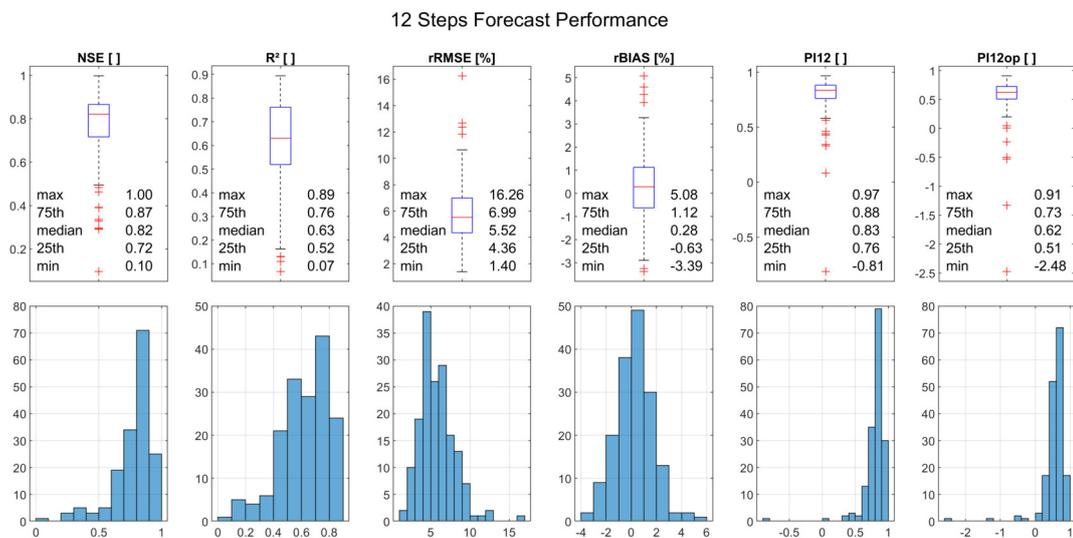


Abbildung 3.21: Gesamtüberblick Fehler Jahreszeitenvorhersagen

Wird auch hier die Korrelationsmatrix der Fehlerkriterien betrachtet (Abb. 3.22), wandelt sich das bisherige Bild ein wenig. Weiterhin in recht engem Zusammenhang stehen je NSE, R^2 und rRMSE, ebenso wie PI(12) und PI(12)op. Letztere unterscheiden sich wieder hauptsächlich in der Interpretierbarkeit ihrer Ergebnisse, PI12op ist daher erneut vorzuziehen. Für die übrigen Zusammenhänge schwächen sich viele Korrelationen ab, einige verlieren zudem ihre Signifikanz. Dies hat zur Folge, dass die Interpretierbarkeit

der Vorhersageergebnisse auch für Jahreszeitenvorhersagen herausfordernd bleibt. Der direkte Vergleich von open- und closed-loop Ergebnissen erfolgte weiterhin auf Basis von PI12op, jedoch mit leicht veränderten Randbedingungen. Aufgrund der besseren Dynamikerfassung von closed-loop Modellen wurde grundsätzlich ein closed-loop Training bevorzugt, lediglich wenn dabei PI12op-Werte $< 0,5$ auftraten, wurde ein zweites open-loop Modell aufgebaut und die Güte verglichen. Entsprechend kehrt sich der Anteil von closed- und open-loop Modellen nahezu um (49 open-loop, 117 closed-loop). Für Modelle mit schlechter closed-loop Performance war es zuweilen möglich, durch open-loop Training drastische Güteverbesserungen zu erzielen. Trotz i.d.R. schlechterer Dynamikerfassung stellen diese also für einen Teil der Messstellen eine sehr gute Alternative dar.

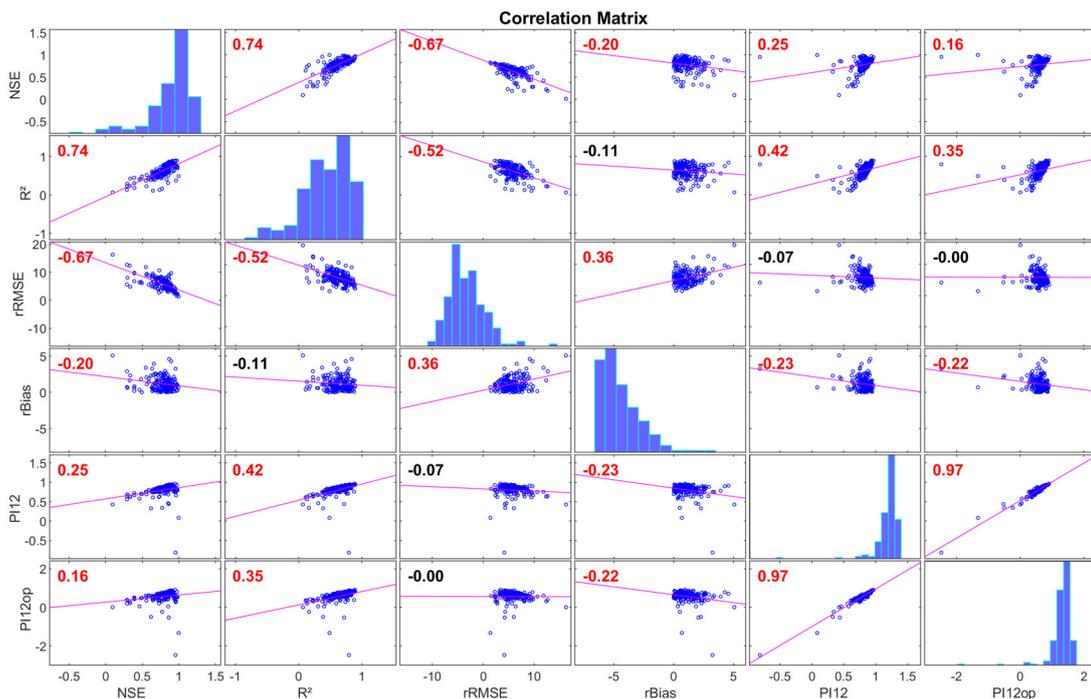


Abbildung 3.22: Übersicht über die Pearson-Korrelation verschiedener Fehlerkriterien für Jahreszeitenvorhersagen (rot: signifikant, schwarz: n.sign.). Die Analyse wurde für den Betrag von rBias anstelle der absoluten Werte durchgeführt.

Für die Untersuchung der saisonalen Schwankung der Vorhersagegüte wird wie bisher nach Startmonat der Vorhersage unterschieden (Abb. 3.23), wobei für die Interpretation der Ergebnisse die Überlappung bzw. die Länge von drei Monaten der Vorhersagen im Hinterkopf behalten werden muss. Entsprechend ist teilweise eine Verlagerung der bisher beobachteten Güte-Saisonalität erkennbar. Beispielsweise verschiebt sich die maximale Güte im Falle von PI12op um ca. 2 Monate nach vorne von vormals Juni/Juli auf April/Mai, ähnlich verlagert sich die Güte im Bereich Dez./Jan auf Nov./Dez.. Während PI12op-Werte im Vergleich insgesamt deutlich erhöht sind, so sind NSE und R²-Werte für Jahreszeitenvorhersagen nun deutlich niedriger, letztere zeigen zudem eine erheblich stärkere saisonale Schwankung als bisher. Für NSE bleibt der Zeitpunkt der maximalen Güte ähnlich, ist jedoch nicht mehr so deutlich ausgeprägt. Auch rRMSE nimmt

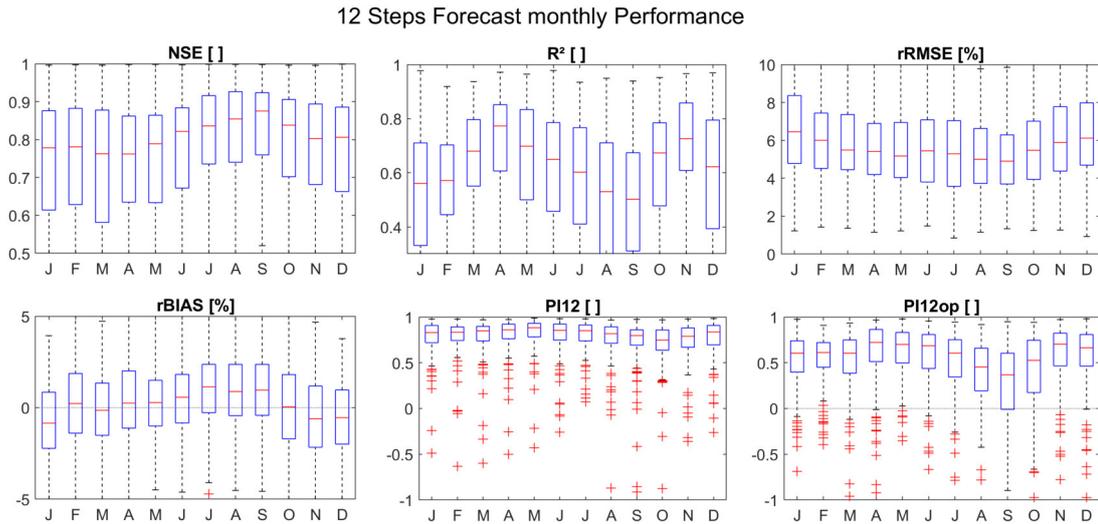


Abbildung 3.23: Gesamtüberblick der Fehler für Jahreszeitemvorhersagen je Startmonat der Vorhersage.

insgesamt und wie zu erwarten war für längere Vorhersagen zu. Dennoch ist keineswegs pauschal mit einer Abnahme der Vorhersagegüte mit zunehmendem Abstand vom Vorhersagestart zu rechnen, Vorhersagen in Monat 3 der Jahreszeitemvorhersage können durchaus bessere oder gleiche Güte aufweisen wie beispielsweise Monat 1 der Jahreszeitemvorhersage. Sollten für einen operationellen Betrieb häufiger initialisierte Inputvorhersagen verfügbar sein, ist es daher unter Umständen lohnenswert, die Güte der einzelnen Monate der Jahreszeitemvorhersage im Detail mit der der Monatsvorhersage zu vergleichen. Für einige Messstellen und einzelne Monate kann so durchaus noch eine deutliche Verbesserung der Güte erreicht werden, indem einzelne Monatsvorhersagen durch Ausschnitte der Jahreszeitemvorhersagen ersetzt werden.

Der Vollständigkeit halber zeigt Abbildung 3.24 die Jahreszeitemvorhersagen der bereits bekannten Messstelle BB_34426014 (a) sowie die saisonale Schwankung der Vorhersagegüte anhand von PI12op (b). Für saisonale Vorhersagen wird für diese Messstelle im Gegensatz zu Wochen- und Monatsvorhersagen auch im Mittel eine zufriedenstellende Güte ($>0,5$) von PI12op = 0,63 erreicht, die deutliche Saisonalität der Güte bleibt weiterhin bestehen.

Abbildung 3.25 zeigt in bekannter Manier die räumliche Verteilung der Vorhersagegüte anhand von PI12op in Kombination der mit den einzelnen Referenzmessstellen verknüpften Messstellenanzahl. Die Farbwahl zur Einstufung der Güte wurde modifiziert, da Werte zwischen 0 und 0,5 nicht mehr zwingend als ausreichend beurteilt werden können (gelb statt hellgrün). Für Messstellen mit diesen Werten sollte bei Bedarf eine Einzelbewertung durchgeführt werden, die sich auch nach Einsatzzweck der Vorhersagen richtet. Es sind Szenarien denkbar, für die diese Güte ausreichen kann. Im Vergleich zu den Abb. 3.19 und 3.11 ist die Verbesserung der Vorhersagequalität deutlich sichtbar. Die Lage der Messstellen mit ungenügender Güte konzentriert sich nun noch stärker auf einen Streifen von Thüringen über Sachsen-Anhalt bis Brandenburg. Für die Jahreszeitemvorhersagen mit PI12op-Werten zwischen 0 und 0,5 fällt ein Schwerpunkt im nördlichen Oberrheingraben ins Auge. Dieser Bereich zeichnet sich durch intensive Grundwasserbe-

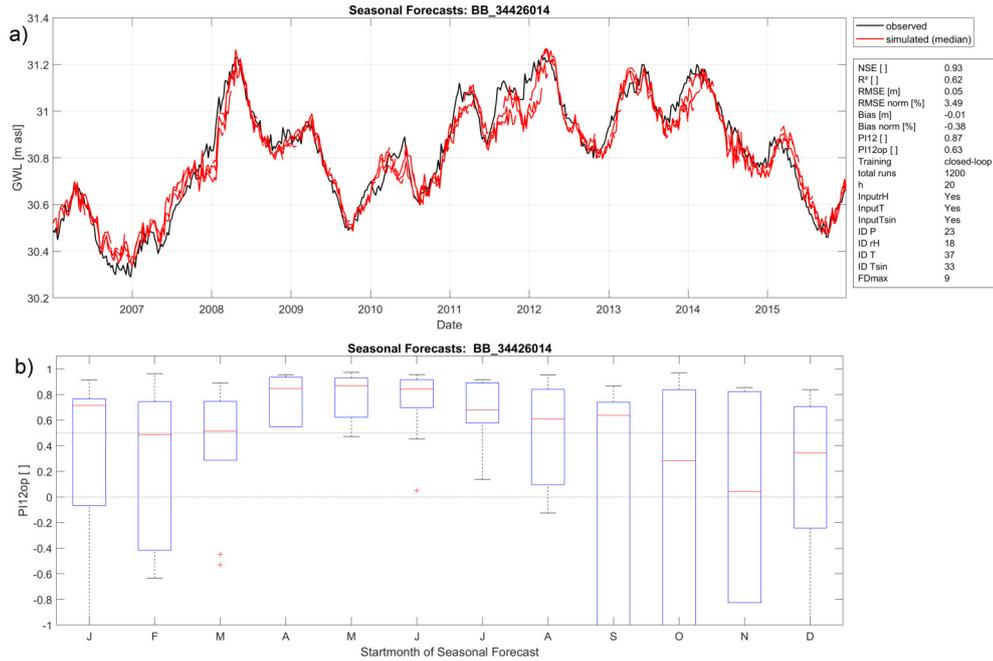


Abbildung 3.24: Jahreszeitenvorhersagen von Messstelle BB_34426014 (a) und PI12op nach Startmonat gesplittet (b)

wirtschaftung aus und ist daher plausibel erschwert über mehrere Monate vorhersagbar. Diese räumlichen Muster zeigen sich auch in Abb. 3.26, die lineare N-S-Struktur relativiert sich jedoch etwas. Die Konzentration auf einen Bereich zwischen Brandenburg und Thüringen bleibt weiterhin bestehen.

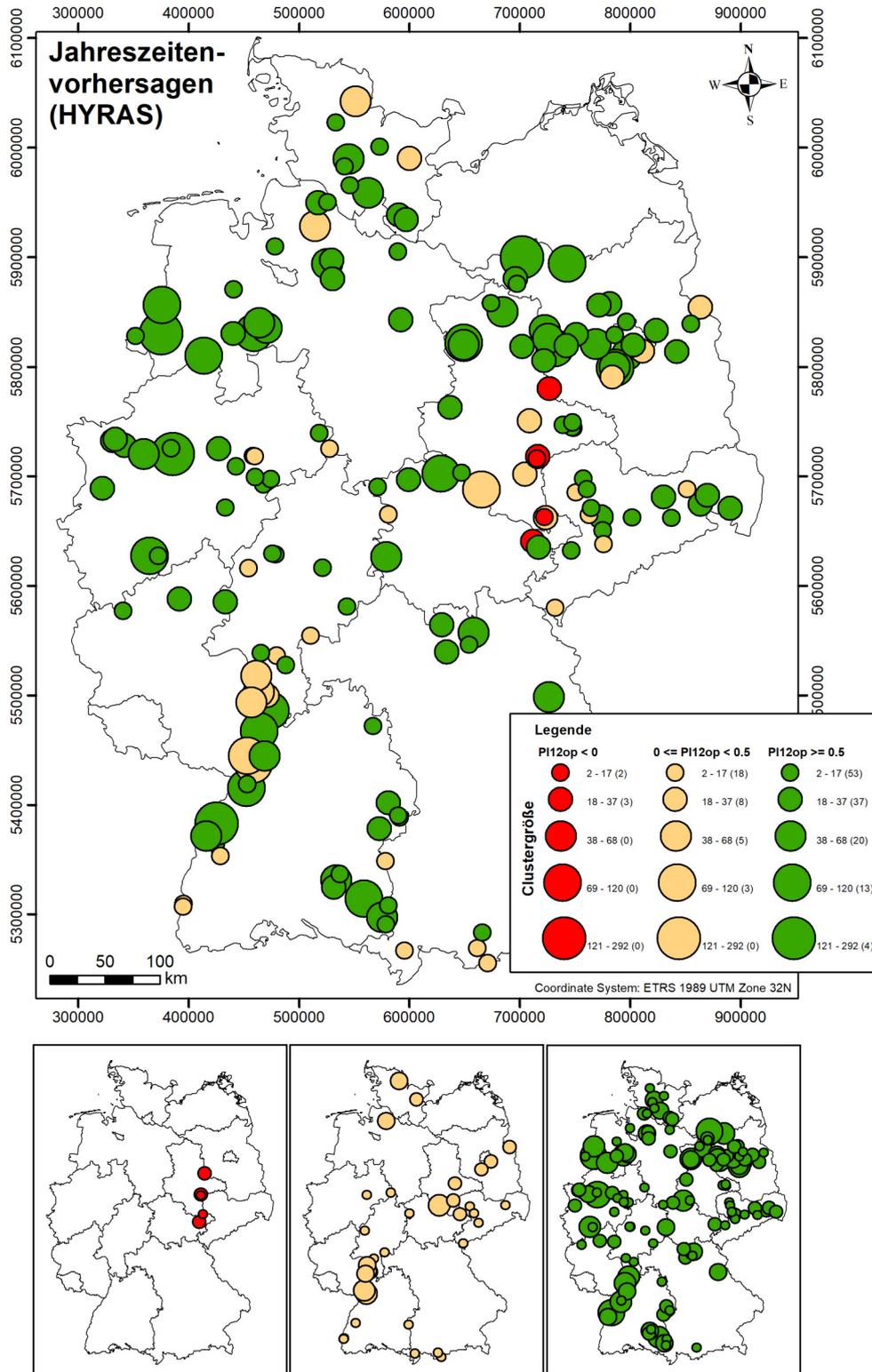


Abbildung 3.25: Übersichtskarte der Vorhersagegüte von Jahreszeitenvorhersagen anhand PI12op (Farbe) im Evaluierungszeitraum (2006-2015) sowie die mit den RM verknüpfte Messstellenanzahl (Clustergröße, Kreisgröße).

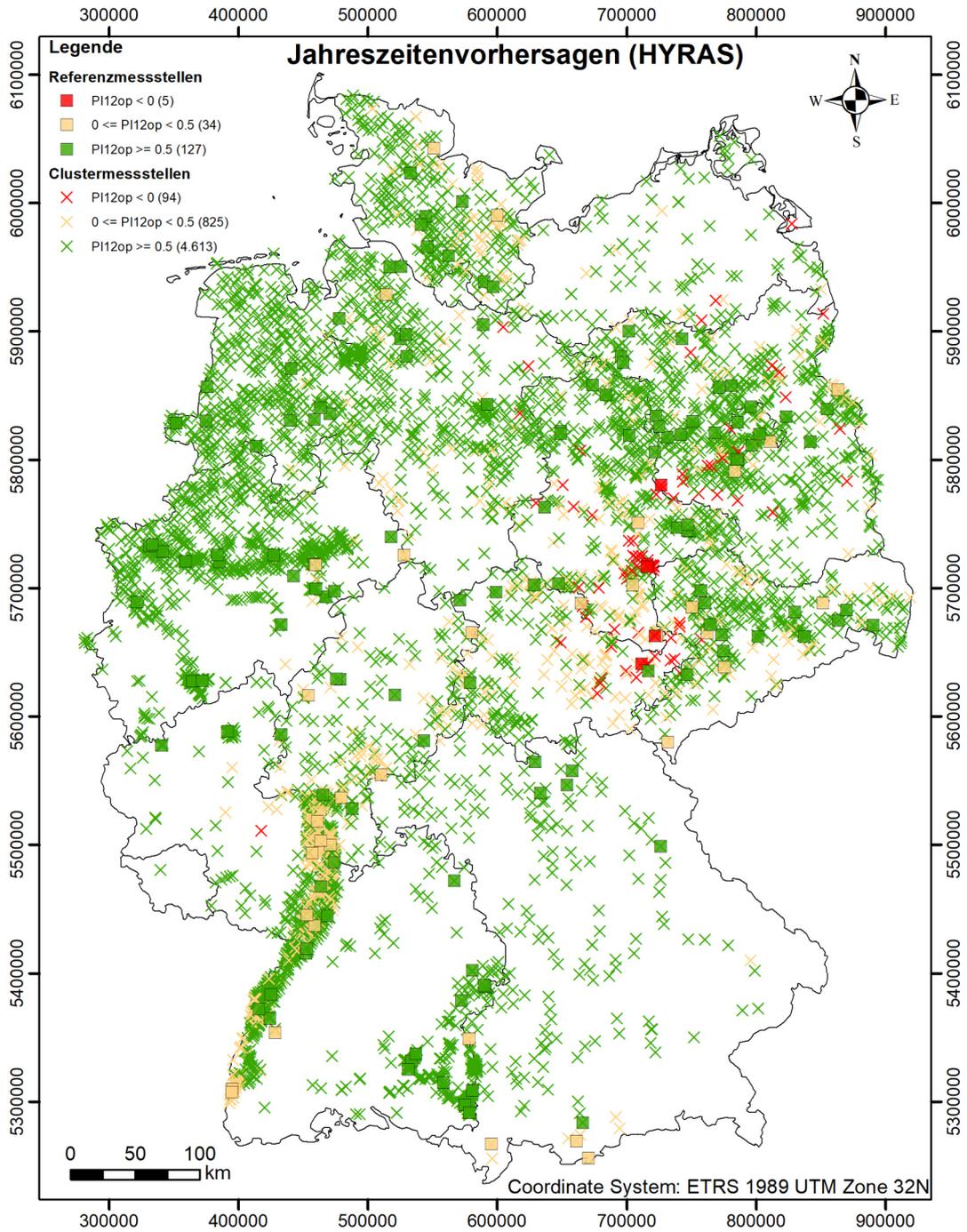


Abbildung 3.26: Übersichtskarte der Vorhersagegüte von Jahreszeitenvorhersagen anhand PI12op (Farbe) im Evaluierungszeitraum (2006-2015) sowie alle mit den RM verknüpften Clustermessstellen

3.4.4 Diskussion der Modellanpassungen und mögliche Fehlerquellen

Für alle Grundwasserstandsvorhersagen wurde, wie bereits in Abschnitt 3.3.1 beschrieben, ein auf leicht zu messenden und flächendeckend verfügbaren meteorologischen Messgrößen basierender Ansatz gewählt. Grundgedanke ist, dass für Messstellen mit überwiegend natürlicher Grundwasserdynamik der Zusammenhang zwischen Grundwasserstand und den wichtigen Prozessen der Grundwasserneubildung und Evapotranspiration über die meteorologischen Parameter Niederschlag, Temperatur und relative Feuchte abzubilden ist. Gerade für Messstellen, deren Dynamik wesentlich von weiteren Faktoren beeinflusst wird, gilt dies jedoch in der Regel nur eingeschränkt. Hierdurch erklärt sich auch die ungleiche Vorhersagegüte aller Messstellen über die verschiedenen Vorhersagehorizonte hinweg.

Faktoren, welche die natürliche Beziehung zwischen treibenden meteorologischen Kräften und Grundwasserstand überprägen können, sind beispielsweise anthropogene Eingriffe wie Trink- und Brauchwasserentnahmen (Landwirtschaft und Industrie) oder auch Grundwasserhaltungs- und Grundwassermanagementmaßnahmen im Allgemeinen. Daneben spielen auch Oberflächengewässer eine wichtige Rolle, da diese bei entsprechender räumlicher Nähe über Ex- und Infiltration aus und ins Grundwasser starken Einfluss auf die Grundwasserdynamik haben können.

Des Weiteren haben auch weniger dynamische Rahmenbedingungen einen Einfluss auf den Zusammenhang, der zwischen Inputparametern und Grundwasserstand hergestellt werden kann. So können hohe mittlere Flurabstände bei entsprechender Geologie für eine schwache Reaktion des Grundwassers auf meteorologische Ereignisse sorgen, da die entsprechenden Signale nur verzögert und deutlich gedämpft im Grundwasser ankommen (Corona et al. 2018). Je länger diese Verzögerung, umso schwerer ist diese für NARX abzubilden, da die Delays nur bis zur Länge eines Jahres optimiert wurden und rekurrente KNN darüber hinaus meist nur zehn (oder weniger) weitere Zeitschritte direkt erinnern können (Bengio et al. 1994). Dieser Sachverhalt steht in Zusammenhang mit dem sog. „vanishing/exploding gradient problem“.

Neben den genannten Faktoren spielt auch die Landnutzung eine Rolle, da beispielsweise durch Flächenversiegelung der beschriebene Zusammenhang abgeschwächt werden kann. Weiterhin führen auch verschiedene Landnutzungsformen zu unterschiedlich stark ausgeprägten Grundwasserneubildungs- und Evapotranspirationsraten.

Ebenfalls denkbar als Ursache für eine verminderte Vorhersagegüte der Grundwasserstandsvorhersagen ist die Datenqualität des Trainingsdatensatzes. Unzulänglichkeiten der Grundwasserstandszeitreihen beeinflussen die Qualität der Trainingsanpassung ebenso, wie auch ungenaue meteorologische Inputdaten. Ungenauigkeiten letzterer können theoretisch bei der Erstellung des HYRAS Datensatzes ihren Ursprung haben, da die Rasterdaten aus räumlich diskreten Punktdaten mit unterschiedlicher räumlicher Verteilung erzeugt werden. Dies dürfte jedoch nach Einschätzung des DWD aufgrund der hohen Stationsdichte in ganz Deutschland im konkreten Fall keine Rolle spielen. Denkbar sind auch regional variable Ausprägungen des Klimas. Dazu zählen beispielsweise lokale/regionale Häufungen instabiler Wetterlagen, überregionale Unterschiede wie eher kontinental oder eher maritim geprägtes Klima oder auch topographische Begebenheiten (Mittelgebirge etc.), die innerhalb einzelner HYRAS Rasterzellen für signifikant heterogene Bedingungen sorgen können. Weiterhin ist es auch denkbar, dass eine Messstelle

aufgrund eines abweichenden Neubildungsgebietes besser auf die klimatischen Signale von benachbarten Rasterzellen reagieren würde. Diese aufwändige räumliche Analyse war im Rahmen dieses Projektes jedoch nicht zu leisten.

Ein gewisser Einfluss ist auch von der vorherrschenden Geologie an jeder Messstelle zu erwarten. Dies lässt sich vor allem unter dem Gesichtspunkt der Art des Grundwasserleiters gesammelt betrachten. Für Abbildung 3.27 wurde jeder Referenzmessstelle ein Grundwasserleitertyp zugeordnet. Dies basiert teils direkt auf verfügbaren Stammdaten zum Grundwasserleiter, teils wurde diese Information aus verfügbaren Stammdaten in Kombination mit der Lage in hydrogeologischen Teilräumen abgeleitet. Der überwiegende Teil der Messstellen (127) ist in Porengrundwasserleitern verfiltert. Alle anderen (39) in Kluft-, Karst- und Kluft-/Karst-, Karst-/Kluft oder Poren-/Kluftaquiferen wurden zur Vereinfachung und wegen der geringeren Anzahl zusammengefasst. Neben der bereits diskutierten absoluten Zunahme von PI_{top} über die Vorhersagehorizonte weg sind im Mittel für Wochen- und Monatsvorhersagen in Porengrundwasserleitern leicht bessere Vorhersagen zu erwarten. Im Fall von Jahreszeitenvorhersagen gilt dies nicht mehr, jedoch zeigen auch Kluft-/Karstaquifer-Vorhersagen keine signifikant bessere Vorhersagegüte. Die Güte der saisonalen Vorhersagen scheint also nicht in erster Linie vom Grundwasserleiter abzuhängen, gute und schlechte Modellanpassungen sind in allen Grundwasserleitertypen möglich.

Für den Zusammenhang zwischen Vorhersagegüte und hydrogeologischen Großräumen konnte keine klare Systematik ausgemacht werden. Großräume mit hohem Porenaquiferanteil zeigen zwar im Mittel leicht bessere Güte, jedoch existieren auch Großräume die überwiegend als Kluft- und oder Karstaquifere charakterisiert werden, deren Messstellen aber dann hauptsächlich in (fluviatilen) quartären Sedimenten verfiltert sind. Die Interpretation dieser sowieso schwach ausgeprägten Zusammenhänge ist demnach problematisch. Eine Systematik zwischen hydrogeologischen Räumen und Vorhersagegüte ergibt sich schon bereits aufgrund der geringen Anzahl pro Raum nicht, da keine sinnvolle Statistik möglich ist, wenn keine, nur einzelne oder sehr wenige Messstellen in einem Raum existieren.

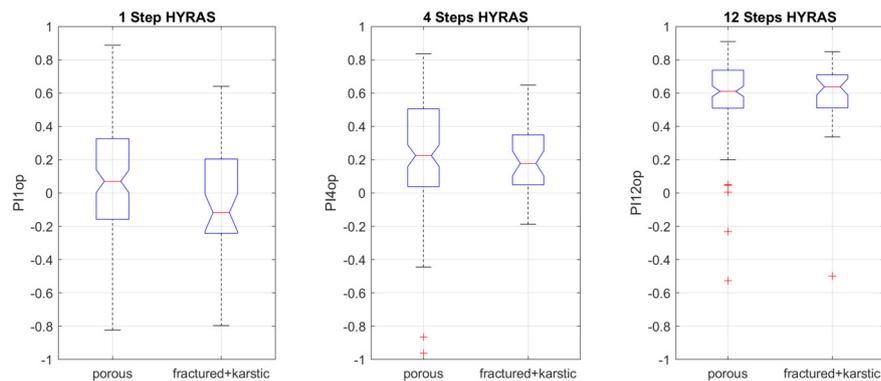


Abbildung 3.27: Übersicht Vorhersagegüte aller Vorhersagehorizonte anhand PI_{op} aufgeteilt nach Typ der Grundwasserleiters.

3.4.5 Operationelle Vorhersagegüte

Die Vorhersagegüte der Grundwasservorhersagen ist in hohem Maß von der Güte der Inputdaten abhängig. Die bisherigen Betrachtungen erfolgten auf Basis von Beobachtungsdaten (HYRAS) und stellen damit bestmögliche Vorhersagen dar, die auf nahezu optimalen Inputdaten beruhen. Für die operationelle Vorhersage stehen jedoch keine Beobachtungsdaten zur Verfügung. Um die Güte der operationellen Vorhersagen abschätzen zu können, wurde daher eine Evaluierung der trainierten Modelle anhand modellierter Input-Zeitreihen (EPISODES Hindcasts) vorgenommen. Diese beinhalten naturgemäß teils erhebliche Unsicherheiten. Im vorliegenden Fall betrifft dies neben der für Deutschland generellen räumlich und zeitlich stark variablen Vorhersagegüte für meteorologische Parameter vor allem natürlich die Vorhersage von Starkniederschlägen. Diese können je nach Messstelle erheblichen Einfluss auf den Grundwasserstand haben, welcher u.U. schon mit beobachteten Daten nur erschwert abgebildet werden kann. Da die zur Verfügung stehenden EPISODES Vorhersagen für Jahreszeitenvorhersagen verwendet werden sollen, erfolgt die Evaluierung des zu erwartenden Güteverlustes an dieser Stelle ebenfalls nur anhand des 3-Monats-Vorhersagehorizontes. Für kürzere Vorhersagehorizonte (vgl. Abschnitte 3.4.1, 3.4.2), können im operativen Betrieb voraussichtlich andere und häufiger als monatlich zur Verfügung stehende Vorhersageprodukte des DWD genutzt werden, die jedoch aufgrund des hohen Speicherbedarfes nicht für Vergleiche in den gewählten Evaluierungszeiträumen (2012-2015) zur Verfügung stehen.

Als Grundlage für die Evaluierung dienten die selben Netze, die bereits für die HYRAS-Evaluierung trainiert wurden. Diese wurden lediglich neu eingeladen und entsprechend die Testinputs ausgetauscht. Es existiert also keine Unsicherheit bezüglich unterschiedlicher Trainingsanpassungen. Abb. 3.28 fasst die Fehlerkriterien für die EPISODES-basierten Jahreszeitenvorhersagen zusammen, Abb. 3.29 vergleicht diese direkt mit den HYRAS-basierten Vorhersagen.

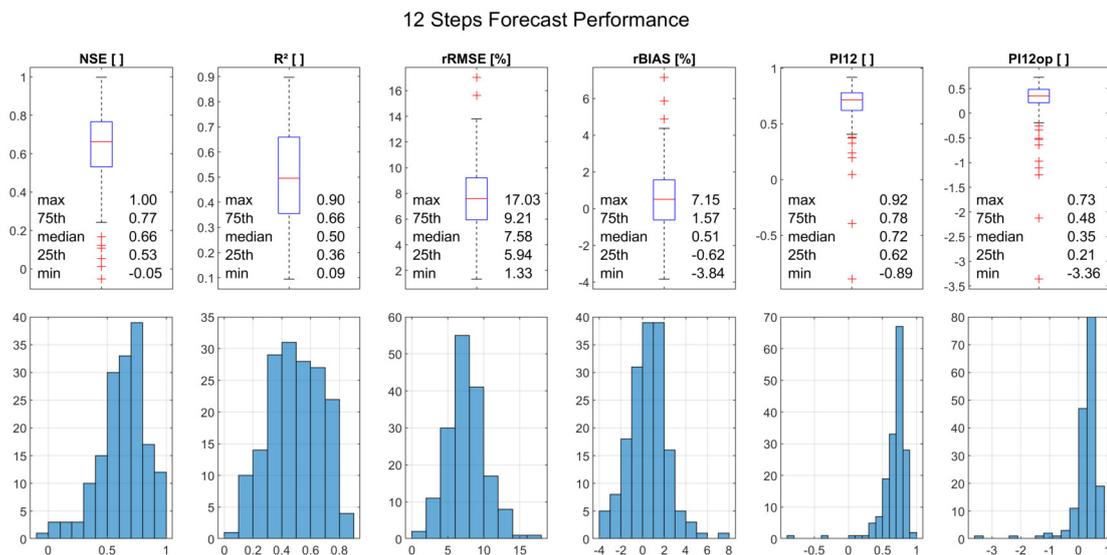


Abbildung 3.28: Gesamtüberblick der Fehler EPISODES-basierter Jahreszeitenvorhersagen

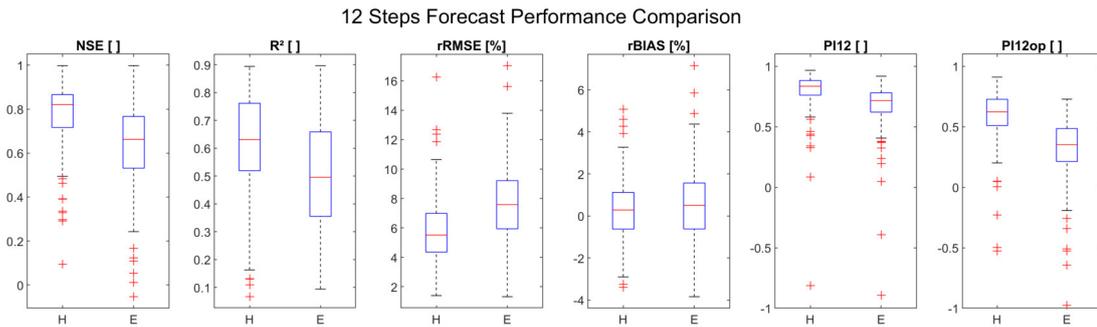


Abbildung 3.29: Vergleich der Güte von Jahreszeitemvorhersagen auf HYRAS-Basis (H) und EPISODES-Basis (E)

Die Unsicherheiten aus den Inputzeitreihen pflanzen sich sichtlich in die Grundwasserstandsvorhersage fort, die Gesamtgüte nimmt für alle Fehlerwerte ab. Die neuen Jahreszeitemvorhersagen weisen nur noch zu <25% PI12op-Werte größer 0,5 auf. Dies entspricht einem recht dramatischen Verlust von ca. 50% aller Messstellen, die dem zuvor definierten Grenzwert ausreichend guter Vorhersagen genügen. Für NSE und R^2 fällt der Güteverlust etwas geringer aus, ist jedoch ebenfalls deutlich. Abbildung 3.30 zeigt die räumliche Verteilung der Vorhersagegüte von EPISODES-basierten Vorhersagen in Zusammenhang mit der Anzahl der zugeordneten Clustermessstellen. Es ist eine Konzentration der Messstellen mit PI12op-Werten größer 0.5 überwiegend auf die größeren norddeutschen Porenquiferbereiche zu erkennen. Die Anzahl der Referenzmessstellen mit Werten kleiner 0 steigt im Vergleich zu Abbildung 3.25 von 5 auf 20. Ins Gewicht fällt hier vor allem, dass nun auch größere Cluster der Kategorien bis 120 Messstellen in diese Gruppe fallen. Zuvor (HYRAS) betraf dies nur Referenzmessstellen, die Cluster bis max. 37 Messstellen repräsentierten.

Bei detaillierter Einzelbetrachtung der Vorhersagen wird deutlich, dass für 26 der betrachteten 166 Messstellen Modellinstabilitäten eines der Hauptprobleme der EPISODES-basierten Jahreszeitemvorhersagen sind. Diese treten für 16 der 26 Messstellen deutlich und häufig, für weitere 10 seltener und nicht ganz eindeutig auf (Anhang J). Abbildung 3.31a zeigt die Jahreszeitemvorhersage der Messstelle BB_34426014, deren Güte im Vergleich zur HYRAS-basierten Vorhersage (vgl. Abb. 3.24) sichtbar abnimmt, jedoch in einem Bereich liegt, der aufgrund der Unsicherheiten der Inputzeitreihen zu erwarten wäre. NSE und R^2 sinken geringfügig um 0,03 bzw. 0,12, PI12op jedoch schon deutlicher um 0,23 auf 0,4. Dennoch ist rein visuell gut ersichtlich, dass die EPISODES Vorhersagen über viele Bereiche durchaus brauchbar und von akzeptabler Güte sind. Hier gilt es, durch detaillierte Einzelbetrachtung und je nach Vorhersagezweck abzuwägen, ob und wie diese Vorhersagen eingesetzt werden können. Abbildung 3.31b hingegen zeigt nun am Beispiel von Messstelle BB_28410218 deutlich, welche Rolle Vorhersageinstabilitäten spielen können. Das Bild ist geprägt von Vorhersagen wechselnder Güte, die sich jedoch grob innerhalb der üblichen Variabilität der Zeitreihe bewegen, sowie extremen Ausreißern nach unten und nach oben, meist sogar innerhalb einer Vorhersageperiode. Jedoch zeigen nicht alle betroffenen Zeitreihen eine so hohe Häufigkeit an Modellinstabilitäten. Die Ursache dieser Instabilitäten ist nur schwierig einzugrenzen. Das Training selbst, welches durchaus in Frage kommen würde, kann weitestgehend ausgeschlossen werden, da sich mit identischen Netzen bei HYRAS-basierten Vorhersagen solche Auffälligkeiten

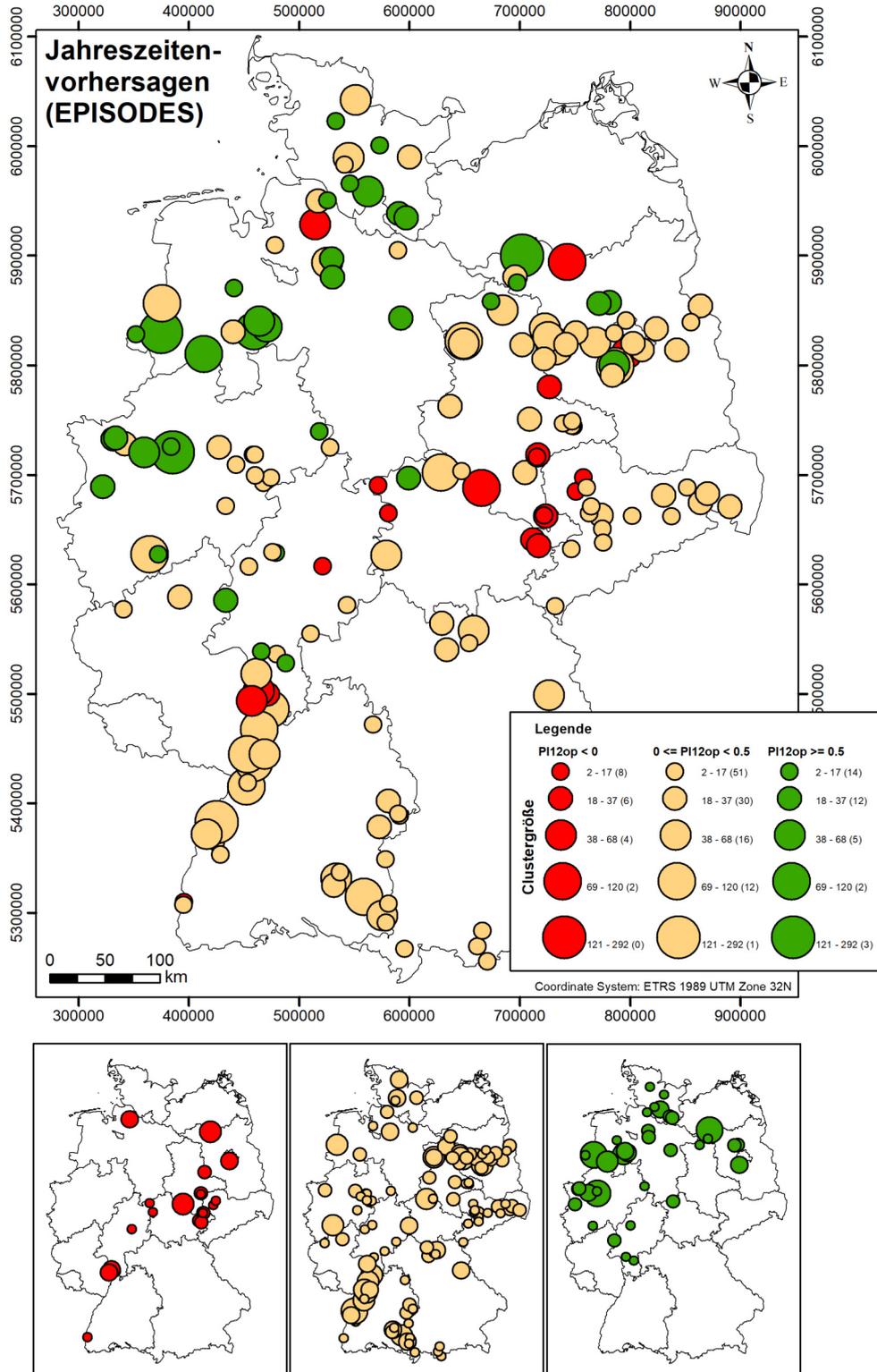


Abbildung 3.30: Übersichtskarte der Vorhersagegüte von EPISODES-Jahreszeiten-
vorhersagen anhand PI12op (Farbe) im Evaluierungszeitraum (2006-
2015) sowie die mit den RM verknüpfte Messstellenanzahl (Clustergrö-
ße, Kreisgröße).

nicht zeigten. Auch ein direkter Vergleich der HYRAS und EPISODES Inputdaten zeigt keine Auffälligkeiten, die solch starke Abweichungen erklären würden. Auch hinsichtlich weiterer Parameter (z.B. genutzte Inputs, Trainingskonfiguration, etc.) lässt sich keine Systematik feststellen. Üblicherweise können einzelne und kleinere Instabilitäten durch Anwendung anderer Netzinitialisierung durchaus behoben werden und mitteln sich bei einer Medianbetrachtung des Vorhersageensembles heraus. Die dargestellten Vorhersagen sind jedoch bereits das Produkt aus meist um die 36.000 Einzelvorhersagen (30 Inputensemblemembers von EPISODES \times 10 KNN-Initialisierungen \times 12 Startmonate \times 10 Jahre Evaluierungszeitraum), wodurch deutlich wird, dass diese Art von stark fehlerbehafteten Vorhersagen beim überwiegenden Teil der Einzelvorhersagen auftreten.

Um den ohne Modellinstabilitäten zu erwartenden Güterverlust der Grundwasserstandsvorhersagen, welcher aus der Unsicherheit der Inputdaten herrührt, beurteilen zu können vergleicht Abbildung 3.32 analog zu Abb.3.29 die Güte von HYRAS- und EPISODES-basierten Vorhersagen, diesmal zusätzlich ohne Messstellen, die Modellinstabilitäten aufweisen. Die Abnahme der Performance von HYRAS zu EPISODES-basierten Vorhersagen ähnelt stark der Abnahme, die bereits zuvor festgestellt wurde. Zudem ist der Unterschied der Fehlerwerte mit und ohne Messstellen mit Instabilitäten nur gering, weshalb trotz signifikanter Auswirkungen je Einzelmessstelle auf eine untergeordnete Relevanz für die Gesamtbetrachtung der Güte ausgegangen werden kann.

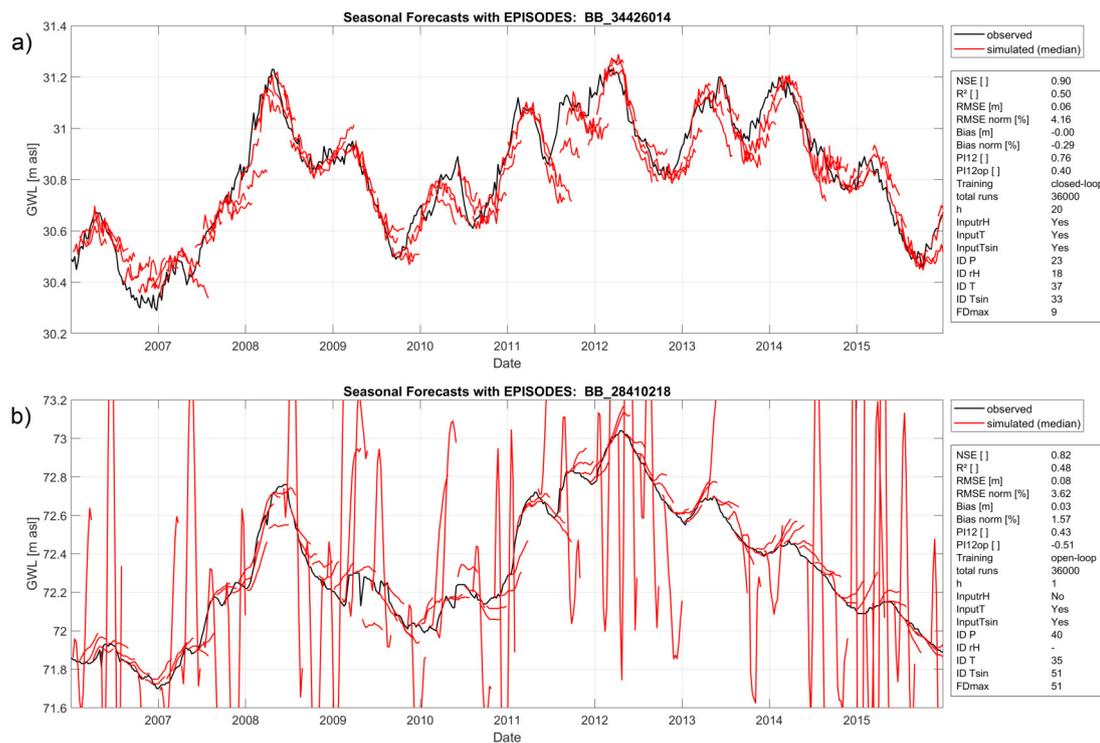


Abbildung 3.31: Vergleich der EPISODES-basierten Jahreszeitenvorhersagen der Messstellen BB_34426014 (a) wie erwartet mit etwas verringerter Vorhersagegüte im Vergleich zur HYRAS-basierten Vorhersage (vgl. Abb. 3.24a) und für Messstelle BB_28410218 (b) mit unerwarteten Instabilitäten.

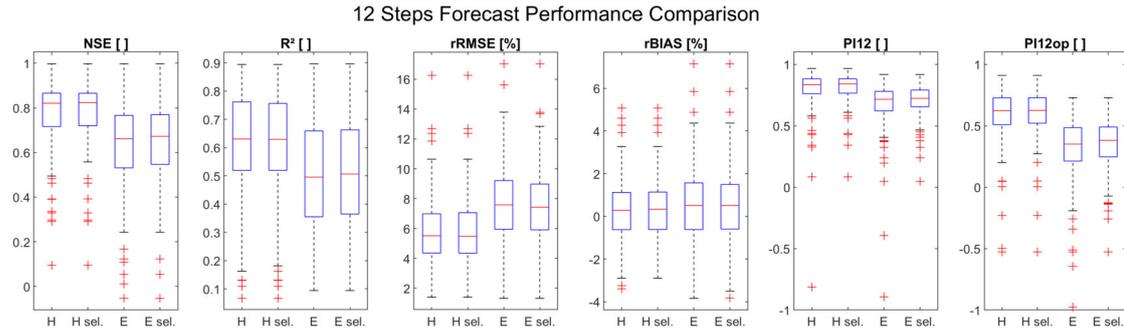


Abbildung 3.32: Vergleich der Güte von Jahreszeitenvorhersagen auf HYRAS-Basis (H) und EPISODES-Basis (E) sowie jeweils ohne Messstellen mit Instabilitäten (sel.)

Der Saisonalität der Vorhersagegüte kommt bei der Betrachtung der operationellen Vorhersagegüte eine besondere Bedeutung zu. Zunächst ist zu erwarten, dass sich die Saisonalität, wie sie auch schon für die Berechnungen auf HYRAS-Basis auftreten, so in die operationellen Jahreszeitenvorhersagen fortpflanzen. Weiterhin besitzen aber auch die EPISODES-Daten selbst eine variable Güte je nach Startmonat und sind zudem räumlich teils sehr variabel (vgl. Abbildung 3.3). Aus diesem Grund unterscheidet sich die Saisonalität der EPISODES-basierten Vorhersagegüten (siehe Abb. 3.33 links) von der der HYRAS-basierten Vorhersagen. Beide saisonalen Verläufe sind leicht versetzt, die besten EPISODES-basierten Vorhersagen sind etwas später, nämlich für die Startmonate Mai - Juli zu erwarten. Im Startmonat September werden jeweils für beide Ansätze die schlechtesten Vorhersagegüten erzielt. Dies zeigt sich auch in Abb. 3.33 rechts, dort sind die paarweise berechneten Differenzen zwischen PI12op(HYRAS) und PI12op(EPISODES) dargestellt. Die geringsten Differenzen sind für Sommer und Winter zu erwarten, wo beide Ansätze gute Ergebnisse erzielen, die höchsten Differenzen im Herbst. Auffällig sind erhöhte Werte um den Startmonat März herum, da dieser Güteabfall für HYRAS-basierte Vorhersagen dort nicht so stark ausfällt, wie für EPISODES-basierte Vorhersagen.

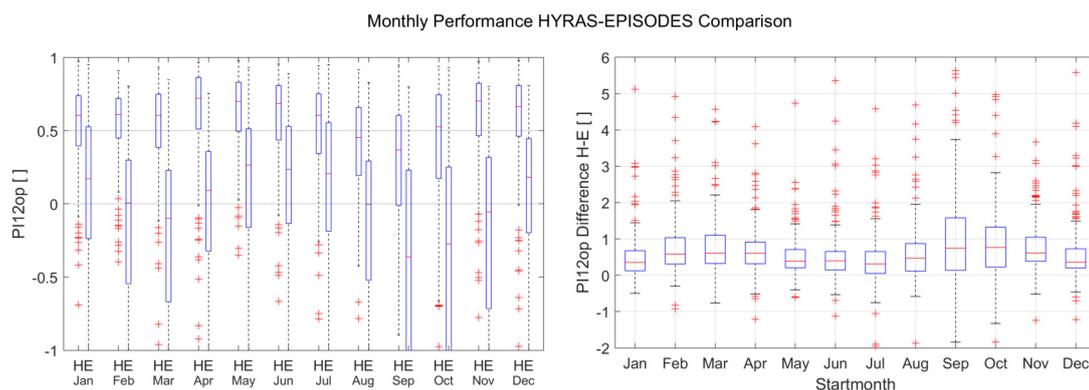


Abbildung 3.33: Vergleich der Vorhersagegüte von Jahreszeitenvorhersagen auf HYRAS-Basis (H) und EPISODES-Basis (E) nach Startmonat (links) sowie die Differenz der Vorhersagen (rechts)

Bei den bisher durchgeführten Betrachtungen lag der Fokus auf dem Vergleich zwischen Hindcasts und beobachteten Werten. Wie bereits angesprochen liegen den betrachteten Medianen jedoch je Messstelle insgesamt ca. 36.000 Einzelvorhersagen zu Grunde. Diese unterscheiden sich sowohl aufgrund der Variation in den Inputdaten (30 Member) selbst, als auch wegen der Betrachtung unterschiedlicher Netzinitialisierungen (10). Jede saisonale Vorhersage basiert daher auf 300 Einzelrealisierungen einer Grundwasserstandsvorhersage. Abbildung 3.34 illustriert anhand der Vorhersagen der Messstelle BB_34426014 für das Jahr 2007 die Variabilität der saisonalen Vorhersagen (vgl. auch Gesamtverlauf in Abb. 3.31). Es sind die Bereiche, in denen 95% und 50% der Einzelvorhersagen liegen sowie der Verlauf des Medians aller Vorhersagen eingezeichnet. Wichtig für die Beurteilung dieser Darstellung ist, dass keine Informationen über die Güte des Modells in der Vergangenheit und damit über die zu erwartende Güte mit einfließen. Es handelt sich lediglich um die Darstellung der Bandbreite der Einzelrealisierungen aufgrund Inputvariabilität sowie Modellinitialisierungen und damit in gewisser Weise um die Konfidenz der Medianvorhersage. Im dargestellten Zeitabschnitt ist die Bandbreite der Konfidenzintervalle meist ähnlich, und die Beobachtungswerte befinden sich zudem über weite Abschnitte im 50% Intervall, nahe dem Median der Vorhersagen. Für einzelne Abschnitte (z.B. Nov 2007 - Jan 2008) tritt eine deutliche Unterschätzung des tatsächlichen Verlaufes auf und die beobachteten Werte befinden sich fast vollständig außerhalb des 95% Konfidenzintervalls. Der Vergleich mit Abbildung 3.24b macht jedoch deutlich, dass auch für optimale Inputdaten (HYRAS) für den Startmonat November generell die schlechteste Vorhersagegüte zu erwarten ist. Das Problem für Startmonat November ist demnach eine generell schlechtere bzw. ungenügende Modellanpassung und nicht auf die Unsicherheit der Vorhersage der Inputdaten zurückzuführen.

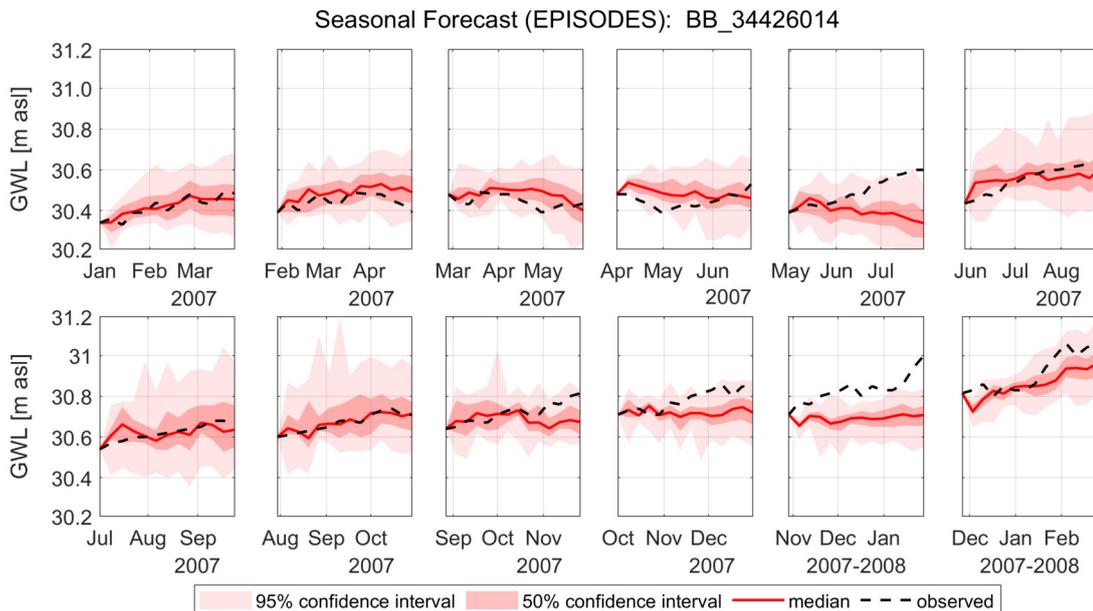


Abbildung 3.34: Konfidenzbereiche der Grundwasserstandsvorhersagen an Messstelle BB_34426014 für die Vorhersagen des Jahres 2007, resultierend aus Inputvariabilität und Modellinitialisierungen.

Abbildung 3.35 visualisiert, erneut am Beispiel der Messstelle BB_34426014, zu welchen zeitlichen Anteilen sich die Konfidenzbereiche im gesamten Evaluierungszeitraum mit den tatsächlich beobachteten Grundwasserständen überschneiden. Hierbei schließt das 95%-Intervall das 50%-Intervall mit ein und es wird zudem nach Aussagemonat (Abb. 3.35a) und Startmonat (Abb. 3.35b) unterschieden. Für den Aussagemonat ist der Startpunkt der saisonalen Vorhersage irrelevant, für einen Startmonat werden wiederum jeweils die zwei Folgemonate mitbetrachtet. Auch die Ergebnisse dieser Abbildung decken sich mit den bisherigen Erkenntnissen (vgl. Abbildung 3.24b). So ist beispielsweise der Anteil des 50% Konf.-Intervalls der Startmonate März bis August generell höher als im restlichen Jahr, oder auch die allgemeine Tendenz zu besseren Vorhersagen in den Sommermonaten (3.35a) deutlich erkennbar. Auffällig ist der Startmonat Februar, für den nach Abb. 3.24b eine eher verringerte Güte zu erwarten ist. Dies scheint jedoch weniger auf ungenügender Modellanpassung zu basieren, da sich sehr viele Vorhersagen im 50% Intervall bewegen, sondern zu großem Teil an den schwierigen Rahmenbedingungen im Frühjahr (variable Niederschläge, Schnee, Schneeschmelze, etc.). Den Abbildungen 3.34 und 3.35 analoge Darstellungen für alle Messstellen sind dem elektronischen Anhang beigelegt.

3.5 Zusammenfassung der Methodenentwicklung zur Grundwasserstandsvorsage

Die zweite Projektphase umfasste die eigentliche Entwicklung und Anwendung von Algorithmen zur Vorhersage von Grundwasserständen auf Basis der Methode Künstlicher Neuronaler Netze. Gewählt wurde ein leicht übertragbarer Vorhersageansatz, basierend auf flächendeckend verfügbaren klimatischen Eingangsdaten (Niederschlag, Temperatur und relative Luftfeuchte) und rekurrenten KNN, genauer sog. NARX-Netzen. Für jede der in Kapitel 2 ausgewählten Referenzmessstellen wurde gesondert ein KNN trainiert. Die optimale Hyperparameter-Konfiguration wurde hierbei mittels bayesischer Optimierung ausgewählt. Entsprechend den Projektzielen erfolgte die Evaluierung der Vorhersagegüte für drei Vorhersagehorizonte: 1 Woche, 1 Monat und 3 Monate. Die Evaluierungszeiträume für die verschiedenen Horizonte waren 4 Jahre (2012-2015) für Wochen- und Monatsvorhersagen sowie 10 Jahre (2006-2015) im Falle von Jahreszeitenvorhersagen (3-Monatsvorhersagen). Die maximal erreichbare Vorhersagegüte für Grundwasserstände rein aufgrund der Modellanpassung wurde dabei auf Basis von beobachteten klimatischen Eingangsdaten (HYRAS-Daten) berechnet und für die Jahreszeitenvorhersagen der Güte von Vorhersagen auf Basis modellierter klimatischer Eingangsdaten (EPISODES Hindcasts) gegenübergestellt. Letztere erlauben die Beurteilung einer im operationellen Betrieb zu erwartenden Güte.

Insgesamt können mit dem gewählten Ansatz für den überwiegenden Teil der Messstellen zufriedenstellende bis sehr gute Grundwasserstandsvorhersagen über alle Vorhersagehorizonte berechnet werden. Es existiert dabei eine Tendenz zu besseren Vorhersagen bei längeren Vorhersagehorizonten, wenn man als Basis für die Beurteilung eine naive no-model Vorhersage heranzieht. Alle Vorhersagen weisen im Jahresgang eine variable Güte mit saisonalem Verlauf auf, sodass im Zweifel für jede Messstelle und je nach Verwendungszweck der Vorhersage eine monatliche Evaluierung der Vorhersagegüte zu empfehlen ist. Auffällig ist zudem das wiederkehrende räumliche Muster von im Vergleich schlechteren

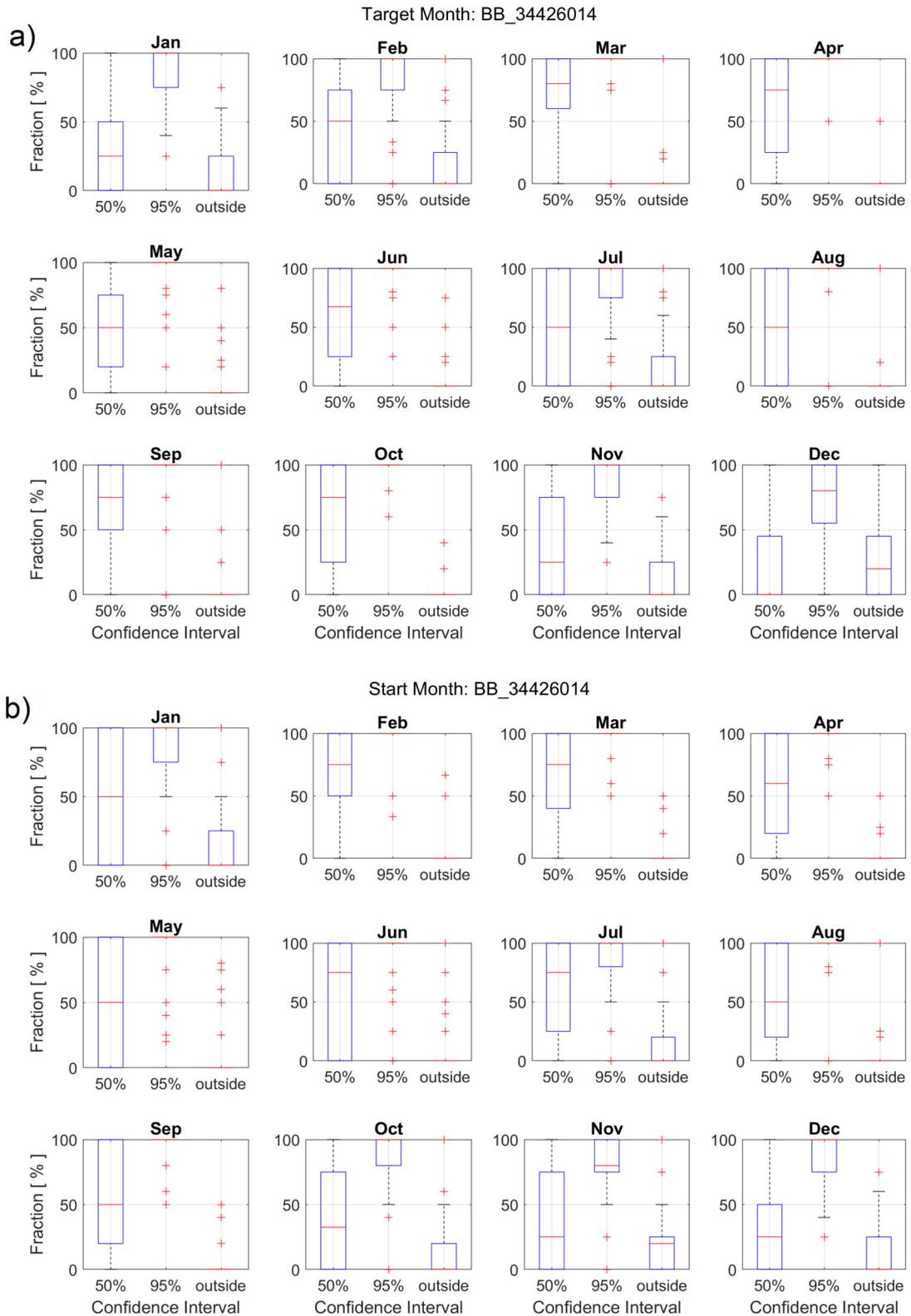


Abbildung 3.35: Zeitlicher Anteil der Lage der beobachteten Werte im Konfidenzbereich der Vorhersagen für BB_34426014 nach Start- und Aussagemonat.

Vorhersagen im Bereich von Ostdeutschland. Eine Hauptursache hierfür lässt sich nur schwer fassen. Im Falle von operationellen Vorhersagen ist ein deutlicher, jedoch auch häufig (je nach Verwendungszweck) verkräftbarer Güterverlust der Grundwasserstandsvorhersagen zu erwarten. Für einen kleinen Teil der Modelle treten zudem in diesem Fall Modellinstabilitäten ungeklärter Ursache auf, die jedoch in der Gesamtbetrachtung über alle Messstellen kaum ins Gewicht fallen.

Für die Messstellen, für die keine zufriedenstellende Vorhersagegüte erzielt werden konnte, liegt dies maßgeblich daran, dass die einbezogenen Input-Parameter die Dynamik der Messstellen nicht ausreichend erklären können. Als Ursachen hierfür kommen vor allem anthropogene Beeinflussungen, Interaktionen mit Oberflächengewässern und allgemein Landnutzungsformen in Frage, die eine natürliche Reaktion des Grundwasserspiegels auf meteorologische Input-Parameter weitgehend überprägen. Aus diesem Grund erscheint auch die Auswahl anderer Messstellen des jeweiligen Clusters (z.B. die zweite oder dritte Messstelle relativ zum Clusterzentrum) als alternative Kandidaten zur Vorhersage als wenig vielversprechend, da die genannten Ursachen die Dynamik des gesamten Clusters beeinflussen. Gegebenenfalls könnte jedoch eine bessere Korrelation zu den lokalen meteorologischen Input-Parametern zu leichten Verbesserungen in der Vorhersagegüte führen.

Ausblickend lässt sich sagen, dass mit den entwickelten datenbasierten Modellen eine operationelle Vorhersage von Grundwasserständen an Referenzmessstellen mit einem Vorhersagehorizont von bis zu 3 Monaten auf überregionalem Maßstab auf jeden Fall sinnvoll erscheint. Gerade bei den Monats- und Jahreszeitenvorhersagen bieten die entwickelten Vorhersage-Modelle einen erheblichen Mehrwert gegenüber einer naiven no-model Vorhersage. Bei Wochenvorsagen, die schon für Beobachtungsdaten als Input für viele Messstellen keine oder nur eine geringe Verbesserung gegenüber einer no-model Vorhersage erzielen, sollte der Aufwand einer operationellen Vorhersage überprüft bzw. in Relation zum Nutzen gestellt werden. Mittelfristig erscheint darüber hinaus eine Prüfung einer möglichen Einbeziehung weiterer Input-Parameter (z.B. Koppelung mit Abflussvorhersagemodellen bei Messstellen, die von Oberflächengewässern beeinflusst sind) sinnvoll. So könnte auch für Messstellen, die bisher eine nicht zufriedenstellende Vorhersagegüte zeigen, eine deutliche Verbesserung erzielt werden.

4 Vergleich mit numerischen Grundwasserströmungsmodellen

4.1 Zielsetzung

Im folgenden Kapitel werden die Ergebnisse eines Vergleiches der entwickelten Vorhersagemethodik mit den Daten existierender numerischer Grundwasserströmungsmodelle vorgestellt. Hierfür wurden drei numerische Modelle ausgewählt und die Zeitreihen von insgesamt 16 Messstellen vergleichend betrachtet. Nach Möglichkeit wurden dabei bevorzugt die in Kapitel 2 ausgewählten und in Kapitel 3 vorhergesagten Referenzmessstellen einbezogen. Die Auswahl umfasst sowohl große überregionale, als auch kleinräumig spezialisierte Modelle. Numerische Grundwassermodelle sind, im Gegensatz zum in diesem Projekt entwickelten KNN-Ansatz, physikalisch basierte Modelle und stellen ein etabliertes und vielfach verifiziertes Werkzeug zur Modellierung von Grundwasserströmung und weiteren Prozessen (z.B. Stoff- und Wärmetransport) dar. Sie werden oft zur Untersuchung von Fragestellungen eingesetzt, die sich nicht in einem Realexperiment untersuchen lassen. Die Komplexität eines Modells hängt in starkem Maße von den realen Gegebenheiten ab. So ist ein Porengrundwasserleiter oft relativ homogen und isotrop und lässt sich im Gegensatz zu einem stark heterogenen Karstgrundwasserleiter relativ leicht modellieren. Zur Erstellung eines numerischen Grundwasserströmungsmodells sind detaillierte Informationen über den Aufbau des Untergrundes und dessen Eigenschaften im Modellgebiet notwendig. Eine entsprechende Datenerhebung ist für gewöhnlich langwierig, aufwändig und damit teuer. Man unterscheidet gemeinhin zwei Hauptarten, Finite-Elemente und Finite-Differenzen Modelle, die nach den jeweils genutzten Ansätzen zur Lösung der betrachteten Differentialgleichungen benannt sind. Im Gegensatz zu KNN, die nur auf einem Input-Output Ansatz beruhen, lassen sich in numerischen Modellen spezifische Strukturen, Randbedingungen und lokale oder regionale Einflussfaktoren (z.B. Oberflächengewässer, GW-Entnahmen) gezielt einbinden.

4.2 Material und Methoden

4.2.1 Künstliche Neuronale Netze

Ziel in Kapitel 3 war es, eine übertragbare, auf KNN basierende Methodik zur Vorhersage von Grundwasserständen zu entwickeln und diese zur Berechnung von Wochen-, Monats-, und Jahreszeitenvorhersagen (3 Monate) an den 166 Referenzmessstellen (siehe 2) anzuwenden. Datengrundlage hierfür waren zum einen die wöchentlichen Grundwasserstandszeitreihen der Messstellen aus 2, zum anderen beobachtete und modellierte Klima-Rasterdaten des DWD. Die im Zuge dessen erstellten Modelle und erzeugten Vorhersageergebnisse auf Basis von beobachteten Daten, wurden nach Möglichkeit zum

Vergleich für diesen Bericht herangezogen. Weitere Messstellen, die sich in den Modellgebieten der ausgewählten numerischen Grundwassermodelle befinden und nicht als Referenzmessstellen definiert wurden, wurden dem gleichen Ansatz folgend bearbeitet und vorhergesagt. Demnach wurden zusätzlich NARX-Modelle (Nonlinear Autoregressive Networks with Exogenous Inputs), wie in Kapitel 3 beschrieben, trainiert und die Güte der Wochen-, Monats- und Jahreszeitevorhersagen evaluiert. Diese dienen als Grundlage für den Vergleich mit den numerischen Grundwassermodellen.

4.2.2 Grundwassermodelle des Oberrheingrabens Süd und Nord

Bei den numerischen Grundwassermodellen Oberrheingraben Süd und Oberrheingraben Nord handelt es sich um große, überregionale Modelle die u.A. von der Landesanstalt für Umwelt Baden-Württemberg (LUBW) genutzt und von dieser gepflegt sowie stetig fortgeschrieben werden. Das Modell Süd umfasst grob das Gebiet des Oberrheingrabens von Basel bis südlich von Karlsruhe, das Modellgebiet Nord schließt nördlich direkt an und endet auf der Höhe von Worms (Abb. 4.1). Es handelt sich um elaborierte Finite-Differenzen Modelle auf der Basis von MODFLOW, welche jeweils eine Gitterzellengröße von 100m x 100m sowie 10 bzw. 12 vertikale Schichten aufweisen. Die zeitliche Diskretisierung der Modelle sind je 1 Monat und die Kalibrierzeiträume starten in den Jahren 1970 (Nord) bzw. 1986 (Süd) und enden aktuell im Jahr 2018. Für eine detaillierte Beschreibung des Modells Oberrheingraben Nord sei auf LUBW (2018) verwiesen. Das Modell Oberrheingraben Süd ist im Rahmen des LOGAR-Projektes (Länderübergreifende Organisation für Grundwasserschutz am Rhein) entstanden (LUBW, 2020; Région Alsace, 2012). Vergleichsdaten für beide Modelle wurden freundlicherweise von der LUBW bereitgestellt. Tabelle 4.1 gib einen Überblick über alle Vergleichsmessstellen, deren Lage in Abb. 4.1 vermerkt ist.

Tabelle 4.1: Liste der Vergleichsmessstellen in den Modellgebieten der Modelle Oberrheingraben Süd und Oberrheingraben Nord

ProjID	ID	Name	X¹⁾	Y¹⁾
BW_104-114-5	104/114-5	1807 Renchen 1	425124	5383055
BW_104-307-0	104/307-0	GWM SK 22 - 3021A, Reilingen	468049	5458156
BW_1-115-8	1/115-8	1115 Schutterwald	416098	5369591
BW_126-020-2	126/020-2	3534 Hartheim 8	395316	5309566
BW_138-019-0	138/019-0	FLB 5 Breisach- Hochstetten	396354	5319363
BW_140-020-6	140/020-6	GWM 3543 Bremgarten 9, Hartheim am Rhein	394990	5307000
BW_152-308-7	152/308-7	GWM 1 Flach, Forst	468727	5444693
BW_157-260-8	157/260-8	1555 Bruchhausen	453043	5418954
BW_181-304-4	181/304-4	Heddesheim	473907	5486083
BW_1851-305-1	1851/305-1	GWM 1 Tief, Dossen- heim	473896	5477008
BW_236-115-9	236/115-9	Hohnhurst	415933	5371701
BW_259-259-7	259/259-7	GWM 19-1 T BWV, Blankenloch	458746	5436966
BW_710-256-3	710/256-3	Altlussheim M	463216	5461626
BW_765-306-3	765/306-3	Ketsch Schleiwiese T	463711	5467331
BW_781-304-2	781/304-2	GWM Tief, Heddes- heim	473915	5486080

¹⁾ ETRS 1989 UTM 32N

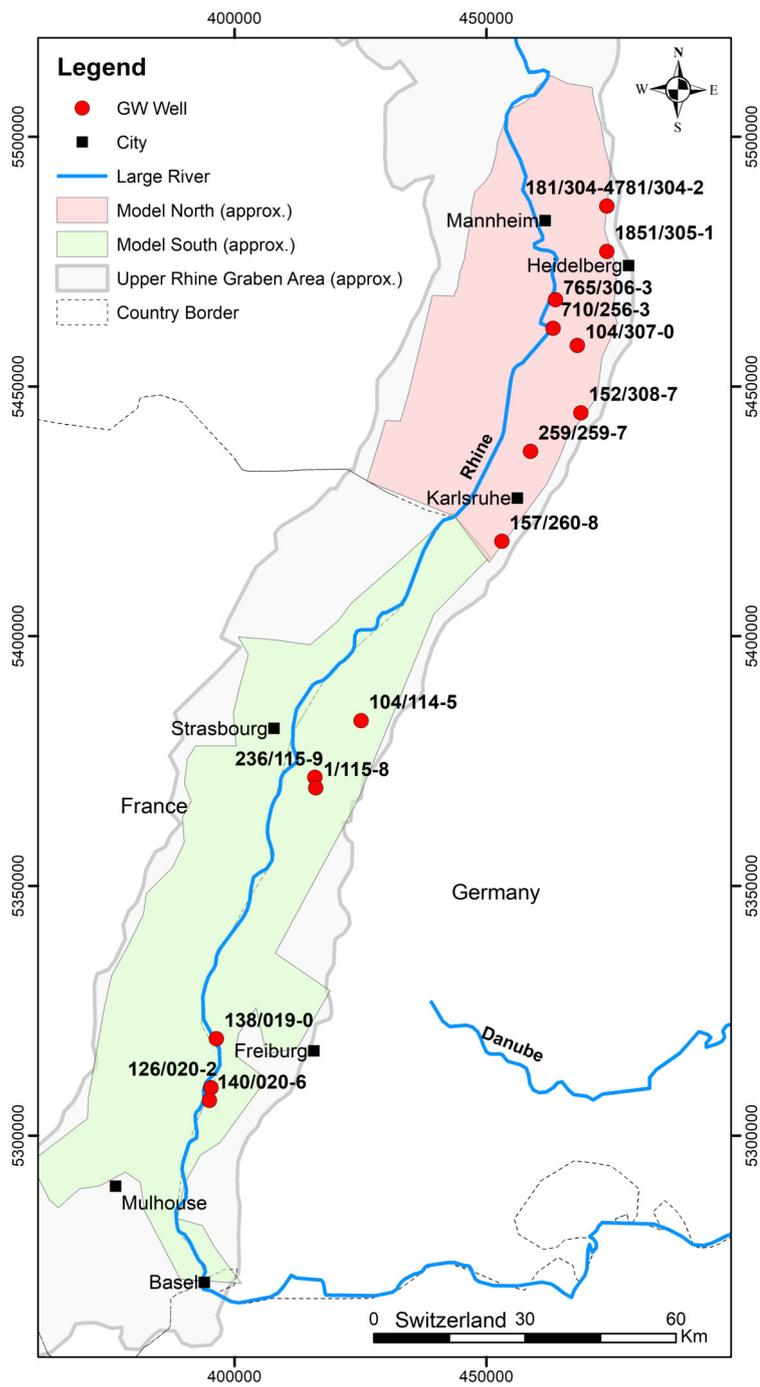


Abbildung 4.1: Lage der Vergleichsmessstellen in den Modellgebieten der Modelle Oberrhein Graben Süd und Oberrhein Graben Nord

4.2.3 Grundwassermodell Lenzen an der Elbe

In Folge einer großen Deichrückverlegungsmaßnahme in Lenzen an der Elbe 2009 wurde ein ca. 4 km² großer Retentionsbereich geschaffen. Um die Auswirkungen der veränderten hydraulischen und ökologischen Verhältnisse durch die Schaffung von Flutmulden und Änderungen am bestehenden System aus Bewässerungsgräben zu quantifizieren und zu beurteilen, wurde ein numerisches Grundwassermodell aufgebaut (Merk, 2015). Es handelt sich um ein besonders im Vergleich zu den Oberrheingrabenmodellen kleinräumiges Finite-Elemente Grundwassermodell auf Basis von Feflow. Die gekoppelte Modellierung von Grund und Oberflächenwasser erfolgte über IfmLake. Das Modell nutzt drei vertikale Modellschichten und arbeitet mit einer variablen zeitlichen Auflösung < 1 Tag. Das Modell befindet sich in Händen der Bundesanstalt für Wasserbau (BAW). Die Daten des Modells wurden freundlicherweise mit Zustimmung der BAW von Markus Merk zur Verfügung gestellt.

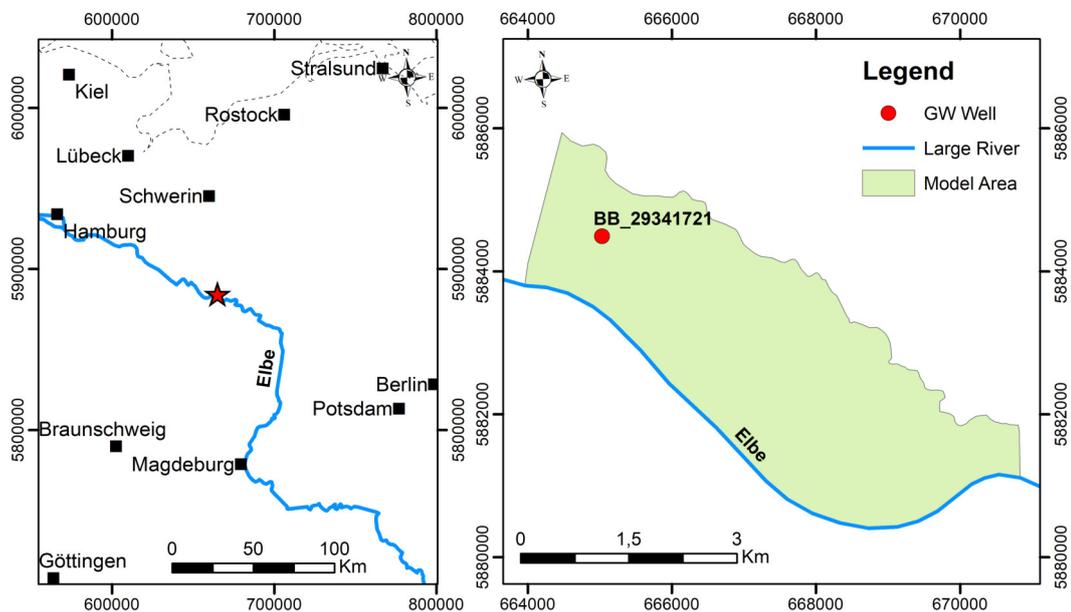


Abbildung 4.2: Lage des Modellgebietes und der Vergleichsmessstelle nahe Lenzen an der Elbe

4.2.4 Vergleichsstrategie

Zur Beurteilung der Vorhersagegüte und zum Vergleich der verschiedenen Modelle werden die bereits in Kapitel 3 erläuterten Fehlermaße genutzt. Hierbei wird trotz der guten Aussagekraft in Bezug auf die Vorhersagegüte auf die Verwendung des (operationellen) Persistency Index verzichtet, da dieser für die numerischen Modelle nicht sinnvoll berechnet werden konnte.

Vergleich mit Modellen Oberrheingraben Süd und Nord

Ein direkter Vergleich von Grundwasserstandsvorhersagen auf unterschiedlichen Zeitskalen zwischen NARX und numerischen Modellen des Oberrheingrabens ist aufgrund einiger Eigenschaften der numerischen Modelle nicht ohne Weiteres möglich. Beide numerischen Modelle arbeiten mit einer Stressperiodenlänge von einem Monat, instationäre Zeitreihen auf Wochenbasis zu erzeugen, ist also nicht möglich. Weiterhin handelt es sich um solch aufwändige Modelle, dass keine Anpassung von Kalibrier- und Validierungszeitraum möglich war. Aus diesem Grund müssen im Folgenden monatliche Kalibrierzeitreihen an 15 Messstellen zum Vergleich mit den Grundwasserstandsvorhersagen auf NARX-Basis herangezogen werden. Damit ist klar, dass von vornherein die numerischen Modelle einen gewissen Vorteil besitzen, da den Modellen im Kalibrierzeitraum die wahren Grundwasserstände bekannt sind und es Ziel der Kalibrierung ist, eine möglichst gute globale Anpassung zu erzeugen. Für den konkreten Vergleich wurden exakt die gleichen Zeiträume wie in Kapitel 3 gewählt. Wochen- und Monatsvorhersagen werden demnach im Zeitraum 2012 bis einschließlich 2015, saisonale Vorhersagen über 10 Jahre von 2005 bis einschließlich 2015 mit den monatlichen Kalibrierergebnissen der numerischen Modelle verglichen. Um eine verbesserte Vergleichbarkeit zu erreichen, wurde zusätzlich eine Aggregierung der NARX-Vorhersagen (NARX agg.) auf Monatsmittel durchgeführt. Für Monatsvorhersagen wurden dabei lediglich eine Vorhersage pro Monat genutzt, die restlichen Vorhersagen wurden für die Aggregierung verworfen, da sonst eine Zuordnung zum jeweiligen Monat nicht eindeutig ist.

Vergleich mit Modell Lenzen a.d. Elbe

Für einen Vergleich mit NARX-basierten Vorhersagen kommt nur eine Messstelle im Modellgebiet in Frage (Abb. 4.2). Dabei handelt es sich um die Messstelle „Lenzen, Weg zur Fähre OP“ (ProjID: BB_29341721), im Modell trägt diese den Namen GWM07. Weitere Beobachtungspunkte des Modells/Grundwassermessstellen im Modellgebiet sind nicht in dem in Kapitel 2 zusammengestellten Datensatz vorhanden. Es liegen zudem keine ausreichenden Datenreihen von Seiten der BAW vor, um für die existierenden Messstellen dennoch NARX-Modelle aufzubauen. Auch die für BB_29341721 vorhandenen historischen Grundwasserstände reichen lediglich bis 2004 zurück. Als Vergleichszeitraum dient der Validierungszeitraum des Modells (12.08.2013 - 31.03.2014). Aufgrund der hohen zeitlichen Auflösung der Modelldaten ist ein direkter Vergleich von wochenbasierten Zeitreihen möglich. Der Vorhersagezeitraum der NARX-Modelle wurde an den Validierungszeitraum des numerischen Modells angepasst.

4.3 Ergebnisse und Diskussion

4.3.1 Vergleich mit den Modellen Oberrheingraben Süd und Nord

Der Gütevergleich der zwei Modellierungsansätze zeigt für die Modelle im Oberrheingraben ein recht eindeutiges Bild. Wie in Abbildung 4.3 klar zu erkennen ist, sind die NARX-basierten Vorhersagen den numerischen Kalibrierergebnissen meist deutlich überlegen. Für diesen Überblick wurde auf die Darstellung von RMSE und Bias (absolut)

verzichtet, da diese sich nicht zum Vergleich über mehrere Messstellen hinweg eignen. Es wurden zudem für die Wochenvorhersagen der NARX-Modelle (1 step ahead) kein Bestimmtheitsmaß R^2 berechnet, das hierbei keine kontinuierlichen Vorhersagewerte sondern Einzelwerte erzeugt werden und so kein Zusammenhang zwischen den Werten besteht, der die Berechnung einer Korrelation rechtfertigen würde. Davon wurde im Falle der aggregierten Monatsmittel abgewichen, da hier sowieso mehrere Vorhersagewerte gemeinsam betrachtet werden. Einzig für R^2 zeigen die Ergebnisse der numerischen Modelle im Falle von Monats- und Jahreszeitenvorhersagen bessere Werte als die zum Vergleich herangezogenen NARX-Ergebnisse. Dies gilt zumindest für die auf Wochendaten basierenden Vorhersagen, die aggregierten NARX-Vorhersagen sind weiterhin überlegen. Dies liegt den vorliegenden Daten zufolge vermutlich hauptsächlich daran, dass numeri-

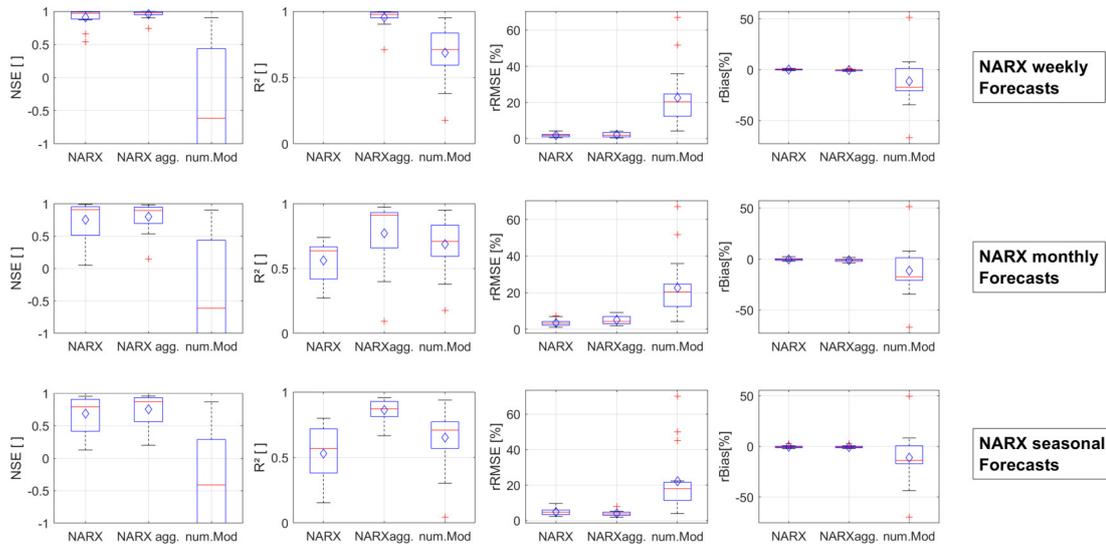


Abbildung 4.3: Überblick über die Vorhersagegüte von NARX-Modellen und die Güte der numerischen Kalibrierungen an den ausgewählten Vergleichsmessstellen im Oberrheingraben

sche Modelle während der Kalibrierung eine globale Optimierung der Fehler im gesamten Modellgebiet durchführen müssen. Durch die hier gute Integration wesentlicher, die Grundwasserdynamik beeinflussender Prozesse, kann der Verlauf der Ganglinien meist gut abgebildet werden, es entstehen jedoch häufig systematische Abweichungen (s.a. rBias). NARX Modelle hingegen werden messstellenspezifisch optimiert und können diesen Vorteil in kleinere Abweichungen umsetzen. Der beschriebene Sachverhalt zeigt sich beispielsweise in Abbildung 4.4. Diese fasst die Ergebnisse aller durchgeführten Vergleiche an Messstelle BW_104-114-5 zusammen. Teilabbildungen a,c und e nutzen für den Vergleich die aggregierten NARX-Vorhersagen, b,d und f die wochenbasierten Vorhersagen. Die Abbildungen e und f erweitern zudem wie beschrieben den Evaluierungszeitraum auf 10 Jahre. Für alle Teilabbildungen gilt, dass sich die blaue Kurve der numerischen Kalibrierung sowie die aggregierten NARX-Vorhersagen auf die schwarz-gestrichelte Kurve der beobachteten Monatsmittel beziehen. Die wochenbasierten Vorhersagen hingegen beziehen sich entsprechend auf die durchgängig schwarze Ganglinie der beobachteten Grundwasserstände. In allen Abbildungen ist deutlich die gute Korrelation der numerischen Kalibrierung an die beobachteten Werte zu erkennen, gleichzeitig existiert jedoch

ein im Vergleich starker, negativer Bias. Absolut gesehen ist dieser mit etwas mehr als einem halben Meter zu beziffern, je nach Einsatzzweck kann dies ausreichend genau sein. Für weiterführende Interpretationen wäre jedoch die Güte in einem Validierungszeitraum gesondert zu betrachten. Aus genannten Gründen erzielen die NARX deutlich bessere Ergebnisse, trotz der Tatsache, dass hier echte Vorhersagen anstelle von Kalibrierdaten bzw. Trainingsanpassungen betrachtet werden.

Abbildung 4.5 fasst die Vergleichsergebnisse für BW_710-256-3 analog zur Abbildung 4.4 zusammen. Hierbei handelt es sich um die einzige der verglichenen Messstellen, für die das entsprechende numerische Modell den NARX-Vorhersagen meist überlegene Werte liefert. Das NARX-Modell erzeugt lediglich im Falle von Wochenvorhersagen noch genauere Werte als das numerische Modell. Beide zeigen insgesamt jedoch eine sehr hohe Vorhersagegüte, sodass beide als sehr gut betrachtet werden können.

Sowohl das numerische Modell, als auch die KNN-Modelle scheitern an der Aufgabe einer guten Anpassung, wenn Ganglinien stark durch nicht implementierte oder in den Input-Daten abgebildeten Einflüsse geprägt sind, wie z.B. für Messstelle BW_138-019-0 (FLB 5 Breisach-Hochstetten) (Abbildung 4.6). Diese spezielle Messstelle ist durch das Möhlinwehr beeinflusst, deutlich sichtbar am mutmaßlichen Abstauen des Wehres im Jahr 2010. Dies ist nicht im numerischen Modell implementiert. Während die NARX-Modelle im Falle von Wochenvorhersagen (Abb. 4.6b) noch scheinbar viele passende Vorhersagewerte liefern und auch die Gütekriterien akzeptable Werte zeigen, wird vor allem an den Unterabbildungen d und f deutlich, dass der Verlauf der Monats und Jahreszeitemvorhersagen nur selten der beobachteten Dynamik folgt. Während des Abstauvorgangs im Jahr 2010 streben alle Vorhersagen steil nach oben in den gewohnten Wertebereich. Die Werte des numerischen Modells zeigen abschnittsweise bessere und schlechtere Anpassungen, in Summe jedoch keine zufriedenstellende Güte.

Abbildungen analog zu den hier eingefügten sind für alle weiteren Messstellen dem Anhang beigelegt.

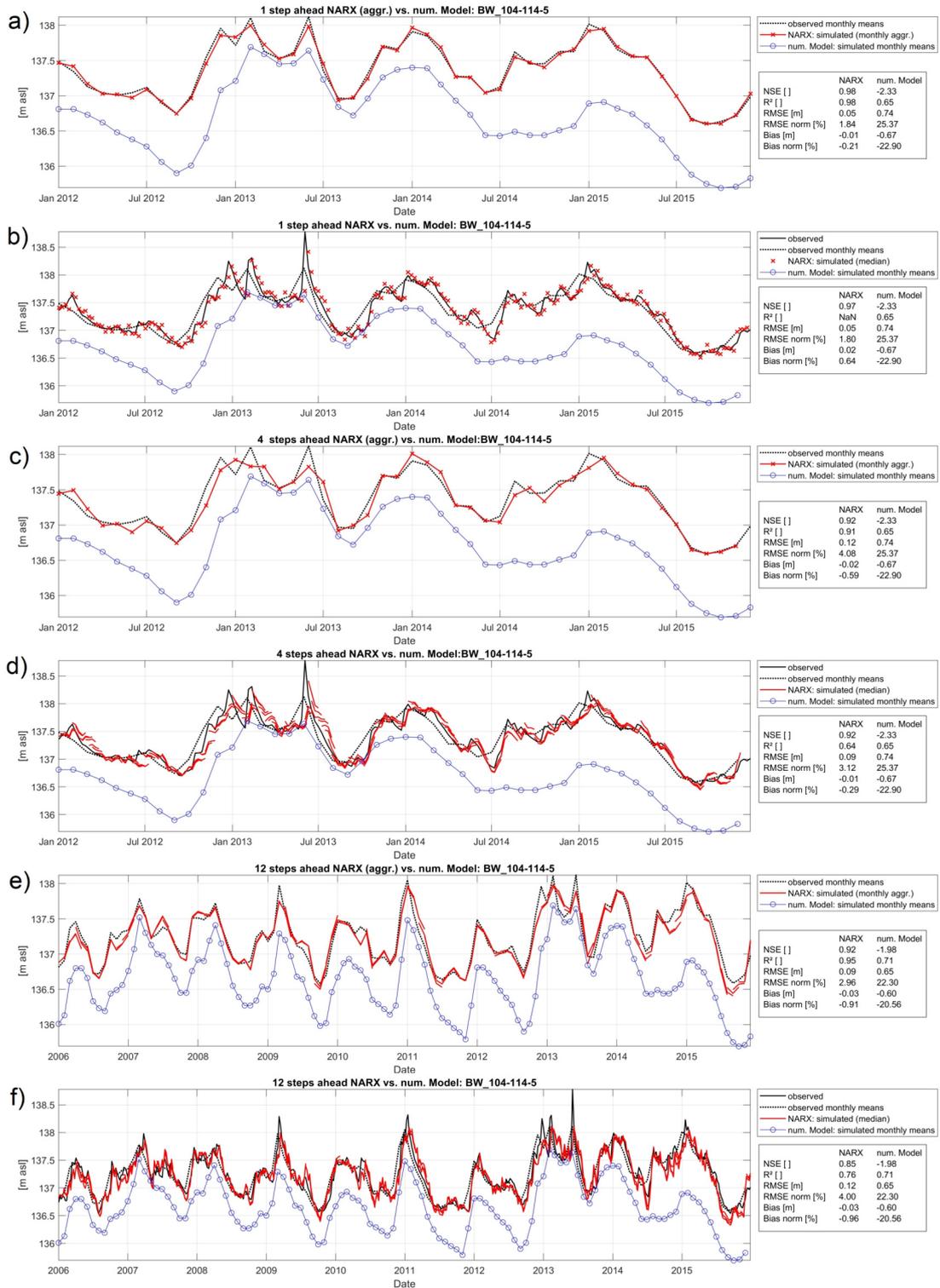


Abbildung 4.4: Vergleich Vorhersagegüte an Messstelle BW_104-114-5

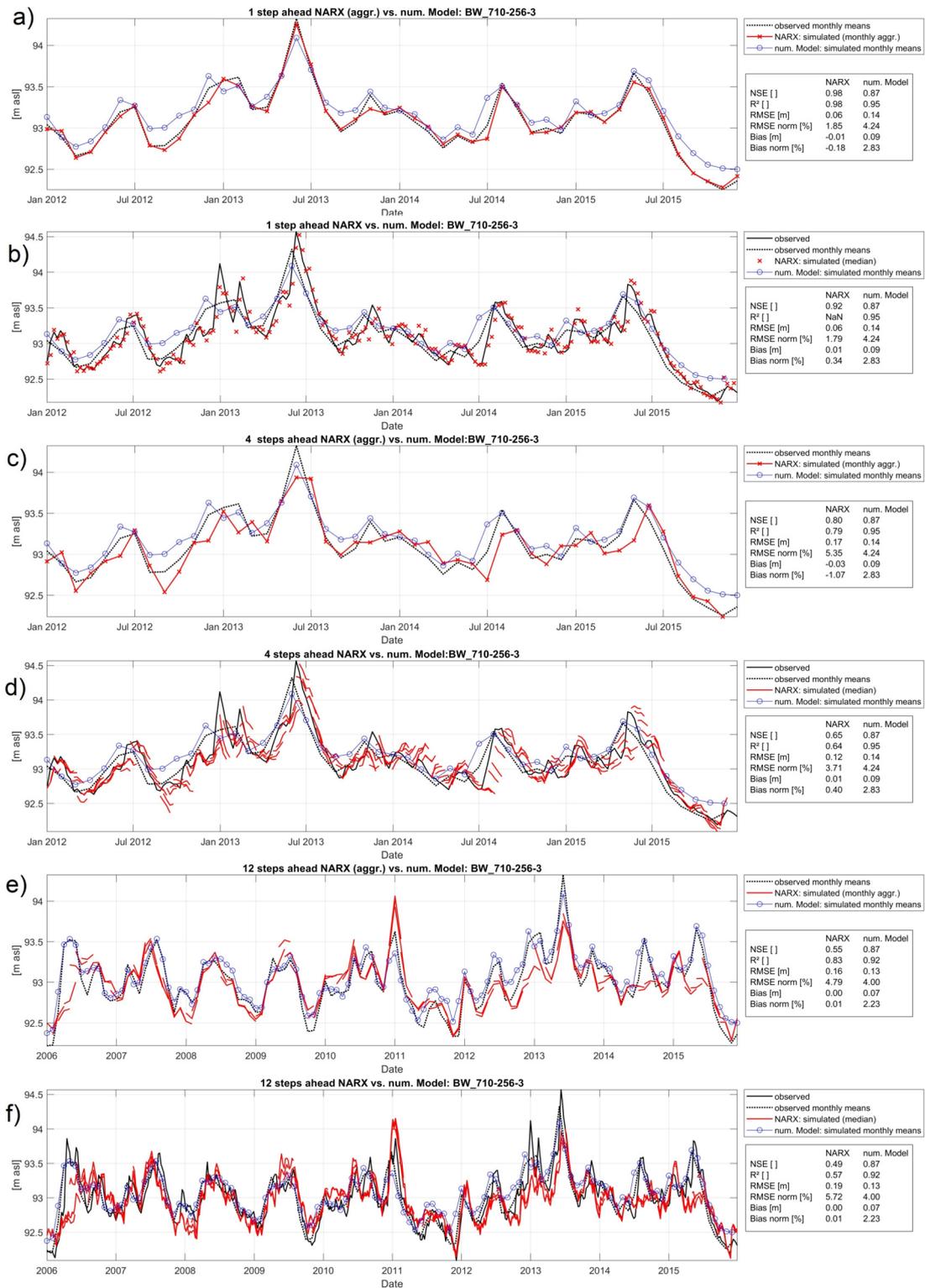


Abbildung 4.5: Vergleich Vorhersagegüte an Messstelle BW_710-256-3

4 Vergleich mit numerischen Grundwasserströmungsmodellen

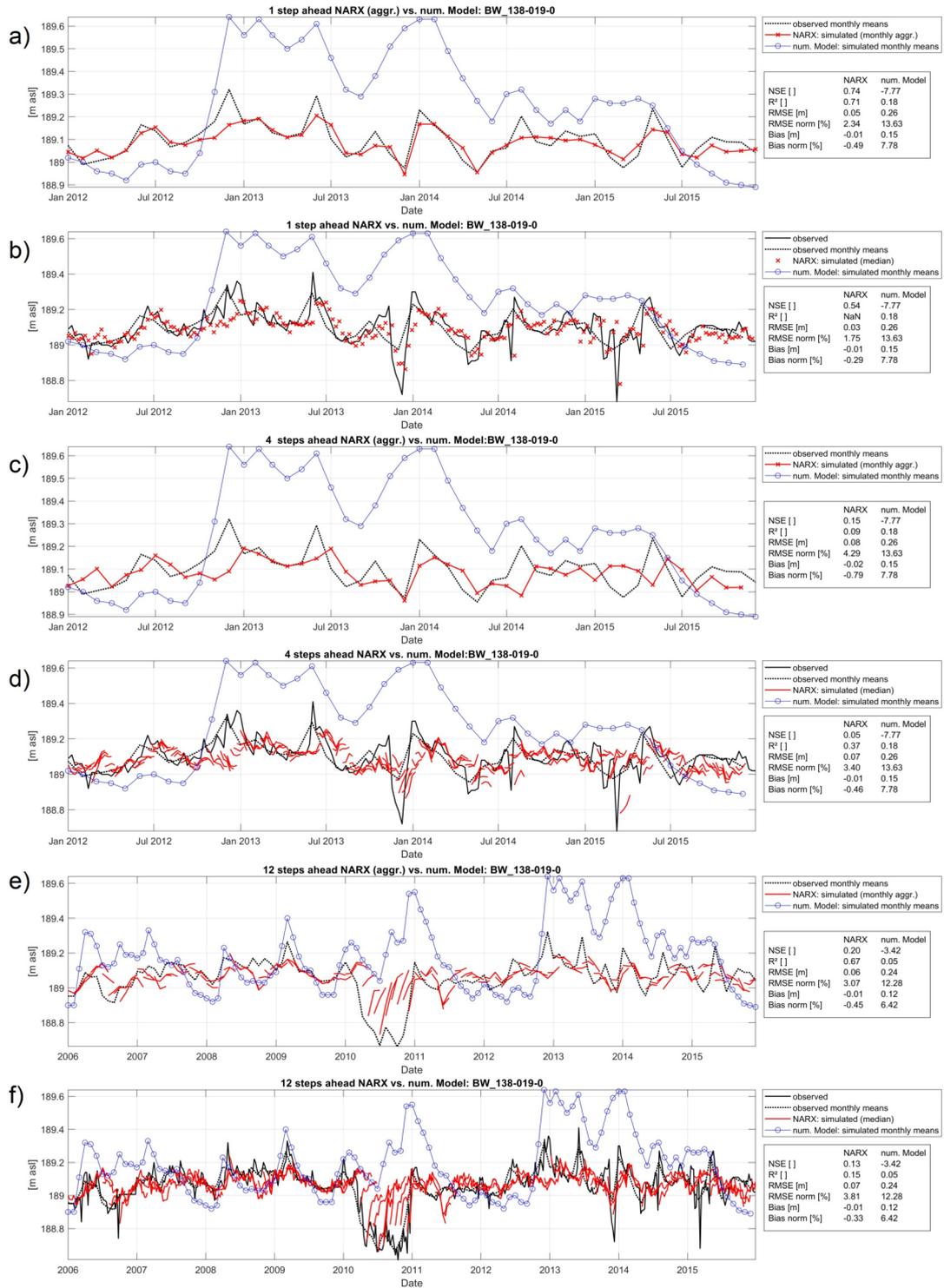


Abbildung 4.6: Vergleich Vorhersagegüte an Messstelle BW_138-019-0

4.3.2 Vergleich mit Grundwassermodell Lenzen an der Elbe

Wie in Abschnitt 4.2.4 dargelegt, kann ein Vergleich der Vorhersagegüte für das Modell Lenzen a.d. Elbe nur für eine Grundwassermessstelle durchgeführt werden. Die Ergebnisse sind in Abbildung 4.7 zusammengefasst. Aufgrund der hohen zeitlichen Auflösung des numerischen Modells im Vergleich zu den Modellen des Oberrheingrabens, ist hier keine Aggregation der NARX-Vorhersagen auf Monatsmittel notwendig. Vielmehr wurden die wochenbasierten Vorhersagen des numerischen Modells aus vorhandenen Tageswerten extrahiert (Stichtags-Selektion). Der Validierungszeitraum wurde als Vergleichszeitraum gewählt und umfasst nur ca. 1,5 Jahre, lässt jedoch diesmal einen direkten Vergleich von Vorhersagen zu. Das numerische Modell ist hierbei den NARX-basierten Vorhersagen im Falle von Monats- und Jahreszeitemvorhersagen überlegen. Es erreicht wie auch im Falle der Oberrheingrabenmodelle gute R^2 -Werte, weist aber auch im vorliegenden Fall nur einen knapp negativen NSE-Wert sowie einen deutlichen negativen Bias auf. Absolut gesehen ist der Bias aber vor allem in Hinblick auf die vertikale Mächtigkeit des Modells und die Gesamtvariabilität der Zeitreihe nur gering. Die isolierte Betrachtung des Validierungszeitraums vermittelt hier ein unvollständiges Bild. Die Güte der NARX-Vorhersagen ist an dieser Messstelle generell verbesserungswürdig. Hierfür kommen drei Hauptursachen in Frage. Zunächst ist die kurze historische Datenverfügbarkeit, die für Training des Modells wichtig ist, zu nennen. Eine vergrößerte Trainingsdatenbasis lässt generell auf verbesserte Vorhersageergebnisse hoffen. Im vorliegenden Fall wurde jedoch mitten im Trainingszeitraum (ca. 2009) eine Deichrückverlegung und damit mutmaßlich auch eine signifikante Veränderung der Grundwasserverhältnisse durchgeführt. Dies lässt generell den Schluss zu, dass künstliche neuronale Netze für den betrachteten Anwendungsfall nicht geeignet erscheinen, da diese mit Hilfe von historischen Daten generalisieren, unter den aktuellen Verhältnissen aber keine entsprechende Datenbasis vorhanden ist. Abschließend ist noch die räumliche Nähe zur Elbe zu nennen. Der gewählte Vorhersageansatz für NARX-Modelle basiert nur auf meteorologischen Inputparametern. Am betrachteten Standort sind die Oberflächengewässer jedoch maßgeblich an der Entstehung der Grundwasserdynamik beteiligt. Ohne entsprechende Inputdaten (z.B. Pegeldata der Elbe) lässt sich dieser Zusammenhang über KNN nicht abbilden.

4.4 Zusammenfassung Modellvergleich

Das Projekt *Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze* (KNN) hat übergeordnet zum Ziel herkömmliche physikalisch-numerische Grundwassermodelle zur überregionalen Vorhersage von Grundwasserständen durch Techniken des maschinellen Lernens zu ersetzen, um so die Abhängigkeit von fehlenden oder unvollständigen Felddaten (z.B. Geometrien und physikalische Eigenschaften des Untergrundes) aufzulösen. Dies ist möglich, da hierbei die Fragestellung auf eine Input-Output Beziehung reduziert wird, und die zuvor aufwändige und teure Datenerhebung zur Charakterisierung des Untergrundes entfällt. Konkretes Ziel des Projektes ist die wiederkehrende Berechnung und vor allem Bereitstellung von Grundwasserstandsvorhersagen von bis zu drei Monaten in die Zukunft an über das gesamte Gebiet der Bundesrepublik Deutschland verteilten Messstellen.

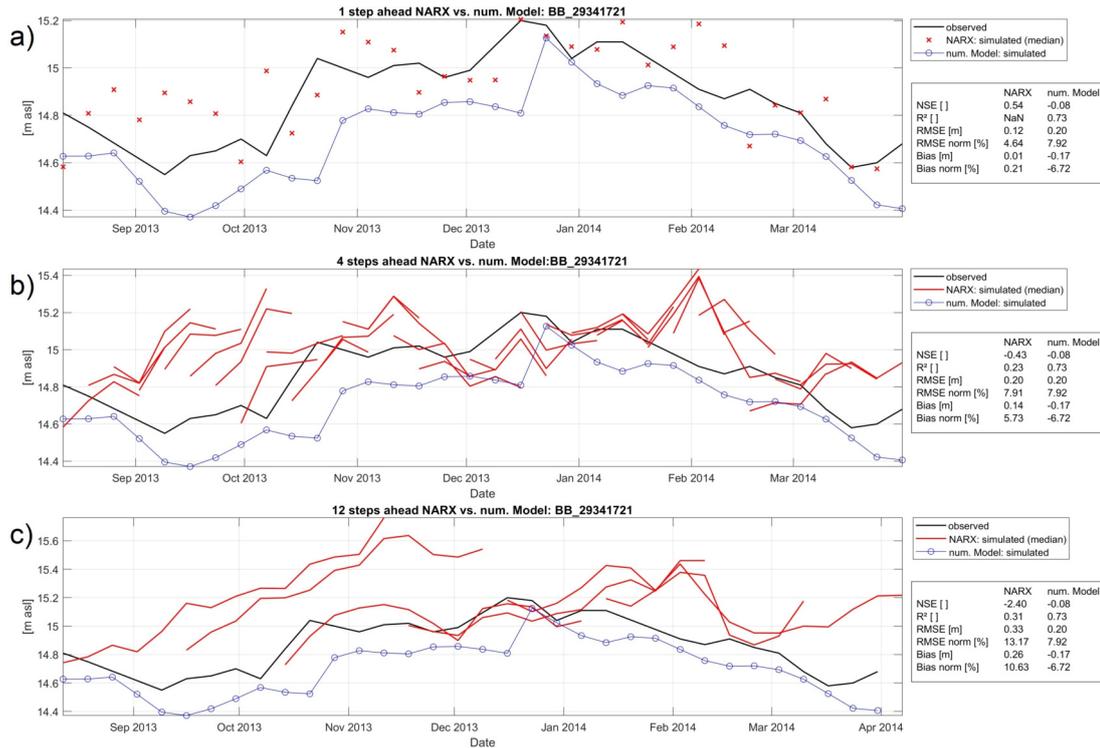


Abbildung 4.7: Vergleich Vorhersagegüte an Messstelle BB_29341721

Das Gesamtprojekt wurde ins Leben gerufen, motiviert durch den hohen Aufwand und die hohen Kosten, welche die Datenerhebung für herkömmliche numerische Grundwassermodelle bei der überregionalen Vorhersage von Grundwasserständen verursacht. Ziel war es, die Abhängigkeit von fehlenden oder unvollständigen Felddaten als Inputdaten für numerische Grundwasserströmungsmodelle aufzulösen und die Fragestellung mithilfe von Techniken des maschinellen Lernens auf eine Input-Output Beziehung zu reduzieren. Dies scheint den vorliegenden Vergleichsdaten nach geglückt. Für den größten Anteil der im Vergleich betrachteten Messstellen liefern die KNN-basierten Vorhersagen bessere Anpassungsgüten als die Daten der numerischen Grundwassermodelle. Dies gilt, trotz des Vergleichs der numerischen Kalibrierdaten mit echten Vorhersagen der KNN. Dabei benötigt der KNN-Ansatz weniger Daten und ist schneller zu implementieren als ein vergleichbares numerisches Modell. Da der entwickelte Ansatz auf einer vereinfachten Betrachtung beruht, die nur meteorologische Inputdaten zur Berechnung der Grundwasserstände einbezieht, wird vor allem im Vergleich mit dem kleinräumigen Modell nahe Lenzen a.d. Elbe jedoch deutlich, wo die Grenzen des KNN-Ansatzes liegen. Die dortige Messstelle befindet sich im Einflussbereich vielfältiger Faktoren (mehrere Oberflächengewässer), zusätzlich zu einer vor wenigen Jahren durchgeführten Deichrückverlegung. Letztere wird, auch wenn lange historische Zeitreihen zum Aufbau fundierter KNN vorhanden wären, deren Eignung fragwürdig machen. Ein numerisches Grundwassermodell ist an diesem Standort also nicht durch KNN zu ersetzen, da eine umfassende Einbeziehung der Einflussfaktoren und Randbedingungen nur in ersterem möglich ist.

Generell gilt, dass der KNN-basierte Ansatz für Messstellen, deren Dynamik über die Zeit grundsätzlich unverändert bleibt und die hauptsächlich von der meteorologischen

Situation abhängt, klare Vorteile bietet und daher zu bevorzugen ist. Hierzu zählen die bessere Vorhersagegüte, die schnellere Implementierung und auch die geringere notwendige Datenmenge. Für gewöhnlich liegt auch die Berechnungsgeschwindigkeit der KNN weit über der numerischer Grundwassermodelle, hierbei spielt aber natürlich die Größe und die räumliche Auflösung des numerischen Modells eine entscheidende Rolle. Sind hingegen keine historischen Datenreihen der Einflussgrößen vorhanden oder sollen spezielle Situationen nachgebildet werden, die neben den meteorologischen Faktoren weitere Einflussgrößen abbilden und die beispielsweise auch eine Änderung der Grundwasserdynamik zur Folge haben, ist die Anwendung numerischer Grundwassermodelle weiterhin klar zu empfehlen. KNN sind in diesen speziellen Fällen nicht zur Vorhersage der Grundwasserstände geeignet.

5 Regionalisierung des mittleren Grundwasserflurabstands

5.1 Zielsetzung

Im Folgenden werden Karten des mittleren Grundwasserflurabstands sowie der absoluten Grundwasseroberfläche in mNN für ganz Deutschland vorgestellt, welche auf Basis der in Kapitel 2 gesammelten und geclusterten Daten erstellt wurden. Dies umfasst demnach sowohl die Referenzmessstellen aus 2 sowie alle diesen zugeordneten Clustermessstellen. Zur Erstellung der Karten wird auf bewährte Interpolationsverfahren zur Regionalisierung (Co-Kriging) und auf die hydrogeologische Raumgliederung von Deutschland (BGR und SGD, 2015) zur Abgrenzung räumlicher Einheiten zurückgegriffen.

Die Ergebnisse der im Folgenden durchgeführten Regionalisierung sind aufgrund einer meist unzureichenden Datenbasis überwiegend nicht belastbar und damit nur mit Vorsicht zu interpretieren.

5.2 Material und Methoden

5.2.1 Datengrundlage

Als Datengrundlage für die Erstellung der Karten des mittleren Grundwasserflurabstands dienen die Grundwasserdaten der in Kapitel 2 geclusterten Messstellen. Hierfür wurden neben den 166 Referenzmessstellen auch alle diesen zugeordneten Clustermessstellen genutzt. Die Berechnungen basieren auf dem mittleren Grundwasserstand der Jahre 2007 bis 2014. Dieser Zeitraum ist der für den vorhandenen Datensatz optimale Kompromiss aus Datenverfügbarkeit und Länge und ermöglicht die Berechnung eines stabilen langjährigen mittleren Grundwasserstandes. Insgesamt basiert die Berechnung damit auf Messwerten von >4800 Messstellen (Abbildung 5.1a), da nicht alle Clustermessstellen im genannten Zeitraum über (ausreichend) Messwerte verfügen. Eine Ausweitung des genannten Zeitraumes würde teils die Datensätze ganzer Bundesländer ausschließen, da dort in Einzelfällen die Datenverfügbarkeit vergleichsweise gering war.

Neben den genannten Grundwasserhöhen wurde zur Berechnung der Flurabstände auf das digitale SRTM Höhenmodell (JPL NASA, 2013a,b) zurückgegriffen (Abbildung 5.1b). Dies kam sowohl bei der Regionalisierung selbst als auch bei der Umrechnung zwischen absoluter Grundwasserhöhe und Flurabstand zum Einsatz.

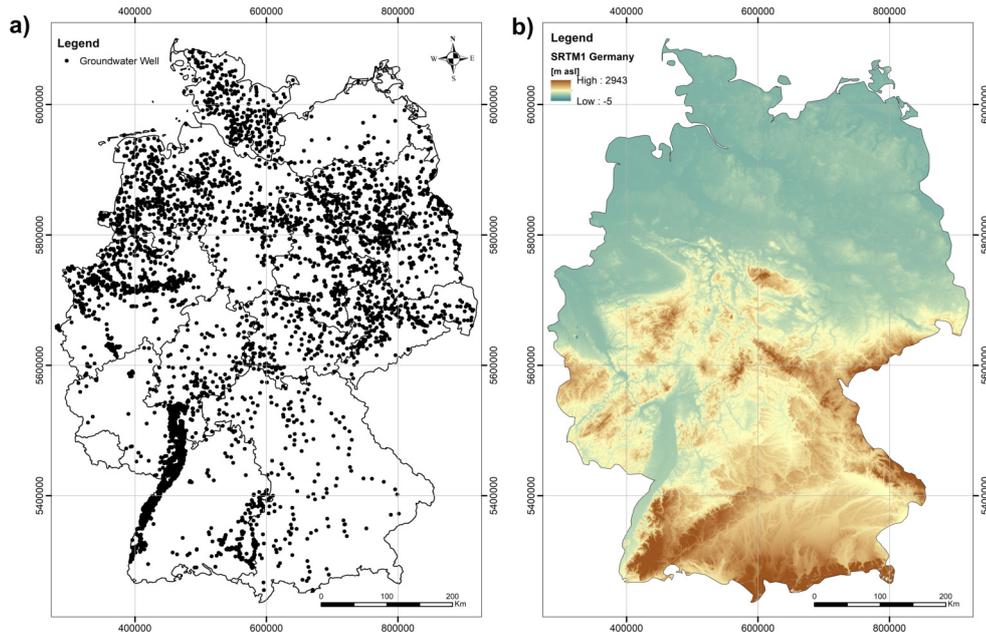


Abbildung 5.1: (a) Lage der verwendeten Grundwassermessstellen (b) SRTM Höhenmodell

5.2.2 Regionalisierung

Die Regionalisierung wird mittels der etablierten Methode des Co-Kriging durchgeführt. Da es sich bei Hydrogeologie um einen komplexen dreidimensionalen Sachverhalt handelt, ist eine direkte Interpolation der Grundwasseroberfläche über Aquifergrenzen hinweg nicht oder nur sehr bedingt sinnvoll. Die höchstauflösende und auf Bundesebene verfügbare räumliche (zweidimensionale) Gliederungen sind die hydrogeologischen Teilräume und die hydrogeologischen Einheiten von Deutschland (BGR und SGD, 2015). Eine kleinräumige Interpolation auf Basis dieser Gliederung ist für den genutzten Datensatz jedoch nicht möglich, da die Messstellendichte meist nicht ausreichend hoch für eine solche Interpolation ist und für viele Einheiten dann keine oder nur sehr wenige Messstellen zur Interpolation herangezogen werden könnten (Abb. 5.1). Selbst auf hydrogeologischer Raumebene (Abb. 2.4) könnte für einen Großteil der Einheiten keine Aussage getroffen werden. Trotz der äußerst begrenzten Aussagekraft basieren daher alle im Folgenden dargestellten Interpolationen auf der hydrogeologischen Großraumgliederung von Deutschland, da nur hier eine einigermaßen flächenhafte Abdeckung der Messstellen und damit der Interpolationsergebnisse erzielt werden kann. Nur vereinzelt werden hydrogeologische Räume getrennt betrachtet, wenn diese beispielsweise räumlich getrennt sind oder eine signifikant andere Messstellendichte aufweisen.

Als Regionalisierungsverfahren wurde Co-Kriging zur Interpolation der Grundwasserhöhen genutzt, bei dem zusätzlich Informationen aus einer oder mehreren korrelierten Sekundärvariablen in die Interpolation einfließen. Neben den Grundwasserhöhen selbst wird dabei im vorliegenden Fall auch die Geländeoberfläche in Form der oben genannten SRTM Datensätze mit einbezogen (Abb. 5.1b), da hierdurch, wie von Ohmer et al. (2017) gezeigt wurde, eine deutliche Verbesserung der Interpolationsergebnisse zu

erwarten ist. Da beim Co-Kriging keine Interpolations-Grenzen vorgegeben werden können, wurden basierend auf der hydrogeologischen Raumgliederung, Gebiete abgetrennt, einzeln betrachtet und interpoliert.

5.3 Ergebnisse und Diskussion

Im Folgenden werden die Ergebnisse der Regionalisierung mittels Co-Kriging dargelegt. Hierbei wurde neben den langjährigen mittleren Grundwasserhöhen auch das SRTM Geländemodell als Sekundärvariable verwendet. Für jeden Großraum wurde eine gesonderte Interpolation durchgeführt, wo notwendig, z.B. weil geographisch nicht zusammenhängend, wurden auch hydrogeologische Räume getrennt betrachtet. Einen Überblick über die Berechnungsgebiete geben Tabelle 5.1 und die Abbildungen 5.5, 5.8 und 5.9 sowie die Abbildungen in Anhang M. Mit Ausnahme einiger Bereiche ohne ausreichend Messstellen wurde demnach für das gesamte Gebiet von Deutschland eine Interpolation der Grundwasseroberfläche durchgeführt. Die berechneten Werte entsprechen sicherlich tendenziell dem groben Verlauf der überregionalen wahren Grundwasseroberfläche, dennoch sollte das Ergebnis mit äußerster Vorsicht betrachtet werden, da die Aussagekraft in vielen Bereichen sicherlich begrenzt ist. Hauptgrund hierfür ist die für eine Regionalisierung fast durchweg unzureichende Messstellendichte. Gerade für geologisch, hydrogeologisch und topographisch komplexere Bereiche wird eine höhere Messstellendichte für eine aussagekräftige Regionalisierung benötigt, als für etwas einfacher zu charakterisierende große Porengrundwasserleiter wie sie z.B. in weiten Teilen Norddeutschlands zu finden sind. Das Gegenteil ist im vorliegenden Datensatz jedoch der Fall. So ergibt sich vor allem für weite Teile des mittel- und süddeutschen Raums, vielleicht mit Ausnahme des Oberrheingrabens, keine belastbare Interpolation der Grundwasseroberfläche. Aus diesem Grund wurde auf eine bundesweite Darstellung der Grundwasseroberfläche an dieser Stelle verzichtet, eine entsprechende Karte findet sich dennoch in Anhang L. Ebenso wurde mit der aus dieser Karte berechneten bundesweiten Flurabstandskarte verfahren, für deren Berechnung die Grundwasseroberfläche von der SRTM Geländeoberfläche abgezogen wurde und für die neben den erwähnten Unsicherheiten noch zusätzlich die Unsicherheit der digitalen Geländemodells eine Rolle spielt. Auch diese sollte nicht für weitergehende Interpretationen herangezogen werden, da wegen der groben und fehlerbehafteten Grundwasseroberfläche oft hauptsächlich ein Artefakt der Topographie abgebildet wird.

Nachfolgend werden die Interpolationen der Grundwasseroberflächen von ausgewählten Gebieten dargestellt und diskutiert. Karten zu allen nicht aufgeführten Gebieten finden sich in Anhang M. Tabelle 5.1 gibt einen Überblick über die Messstellendichten der interpolierten Gebiete sowie die Kreuzvalidierungsfehler der Interpolationen. Die Angabe der Messstellendichte differiert zur Angabe in Tabelle 2.7 wegen der hier angewandten Einschränkung auf Messstellen mit ausreichend Daten zwischen 2007 und 2014. Im Folgenden werden nur die Ergebnisse der drei Gebiete mit den höchsten Messstellendichten aufgeführt, welche auch die bereits in Kapitel 2 (Tabelle 2.10) identifizierten, mutmaßlich am geeignetsten Teilräume beinhalten. Hierzu zählen Großraum 1 (Nord- und mitteldeutsches Lockergesteinsgebiet: 1,7 Mst. pro 100 km²), Großraum 2 (Rheinisch-Westfälisches Tiefland: 2,6 Mst. pro 100 km²) sowie der Oberrheingraben in Form des Großraumes 3 ohne Raum 33 (Oberrheingraben mit Mainzer Becken: 9,7 Mst. pro 100 km²). Es handelt sich bei diesen Gebieten gleichzeitig um Bereiche, in denen Porengrundwasserleiter in

unterschiedlicher Komplexität dominieren. Für die restlichen untersuchten Gebiete liegt die Messstellendichte teils deutlich unter 1 Messstelle auf 100 km², gleichzeitig handelt es sich meist um Gebiete mit hohem Festgesteinsanteil sowie komplexerer Hydrogeologie und Topographie, was eine deutlich höhere Messstellendichte erforderlich machen würde, um belastbare Regionalisierungen durchführen zu können. Ein gegenteiliger Trend ist jedoch offensichtlich. Auch in den drei ausgewählten Gebieten schwankt die Messstellendichte regional teils deutlich und führt dazu, dass alle Grundwasseroberflächen nur mit äußerster Zurückhaltung interpretiert werden sollten. Dort wo eine entsprechende Datengrundlage verfügbar ist, kann durchaus von einer groben Übereinstimmung der Form der interpolierten mit der wahren Grundwasseroberfläche ausgegangen werden. Absolute Fehler können jedoch trotzdem lokal recht hoch sein (Meter bis Zehner-Meter), weshalb sich die erzeugten Karten nicht für eine detaillierte Interpretation oder regionale und lokale Betrachtungen eignen. Insgesamt zeigen die Großräume 1 bis 3 (ohne R33) jedoch im Vergleich zu den restlichen Gebieten (Anhang M) die mit Abstand niedrigsten Kreuzvalidierungsfehler (RMSE) und durchweg hohe Korrelationswerte (R²) (siehe Tabelle 5.1 und Abbildungen 5.2-5.4).

Tabelle 5.1: Co-Kriging Gebiete mit Messstellendichte und Kreuzvalidierungsfehler

Gebiet	Fläche [km ²]	Mst.	Mst. pro 100 km ²	R ²	RMSE [m]	Bias [m]
Großraum 1	128183	2142	1,7	0,95	5,80	0,19
Großraum 2	19020	486	2,6	0,89	9,97	0,47
Großraum 3 ohne Raum 33	9343	907	9,7	0,98	4,73	0,17
Raum 33	3258	22	0,7	0,59	66,07	2,07
Großraum 4	31278	266	0,9	0,98	10,01	0,15
Großraum 5 mit Raum 83	43941	442	1,0	0,80	43,59	1,51
Großraum 6 ohne Raum 61	52581	208	0,4	0,89	41,15	-0,15
Raum 61 Süd	3139	8	0,3	0,07	33,75	-6,22
Großraum 7	6230	10	0,2	0,04	102,29	6,44
Großraum 8 ohne Raum 83	29823	113	0,4	0,91	43,36	-1,66
Großraum 9 ohne Raum 95	19205	205	1,0	0,89	37,98	0,81
Großraum 10 Nord	5385	29	0,5	0,05	44,88	-2,28

Vor allem in Großraum 1 (Abbildung 5.5) ist allein aufgrund der überdurchschnittlichen Fläche und der schwankenden Messstellendichte von regional deutlichen Güteunterschieden in der Qualität der Regionalisierung auszugehen. Um eine Einordnung zu erleichtern, werden im Folgenden Grundwasseroberflächen aus Mecklenburg-Vorpommern und Brandenburg zu Vergleichszwecken herangezogen.

Im Falle von Brandenburg wurden Grundwasser-Isolinien von April 2011 (Möhler et al. 2014) zum Vergleich der Grundwasserhöhen verwendet. Um die Plausibilität des Ver-

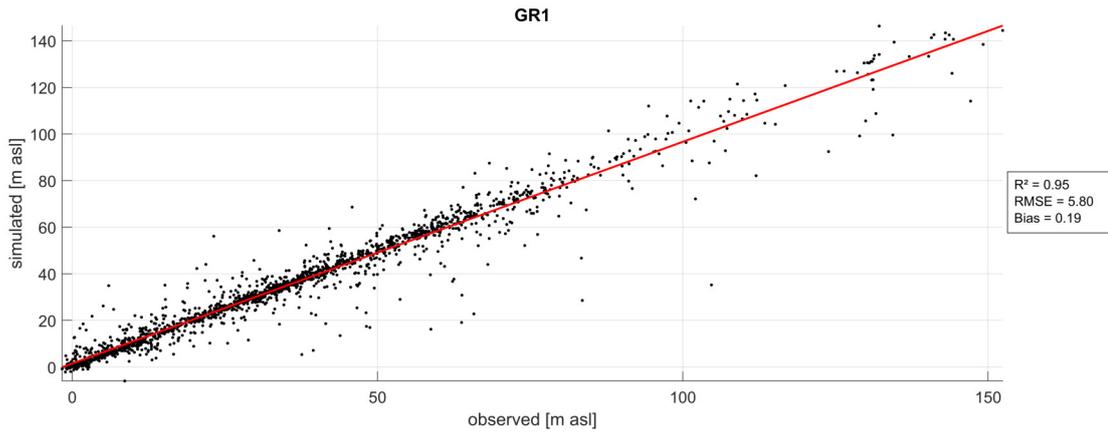


Abbildung 5.2: Scatterplot der Kreuzvalidierung für die Co-Kriging Grundwasseroberfläche in Großraum 1

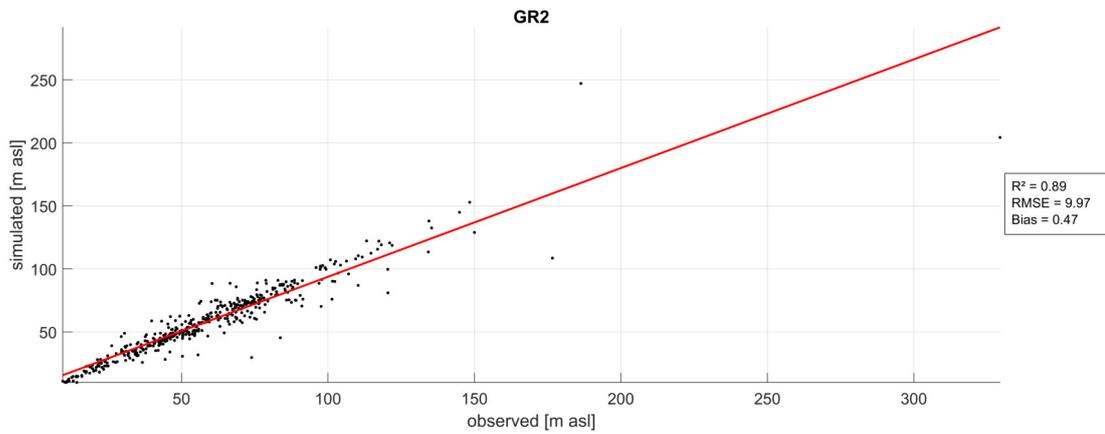


Abbildung 5.3: Scatterplot der Kreuzvalidierung für die Co-Kriging Grundwasseroberfläche in Großraum 2

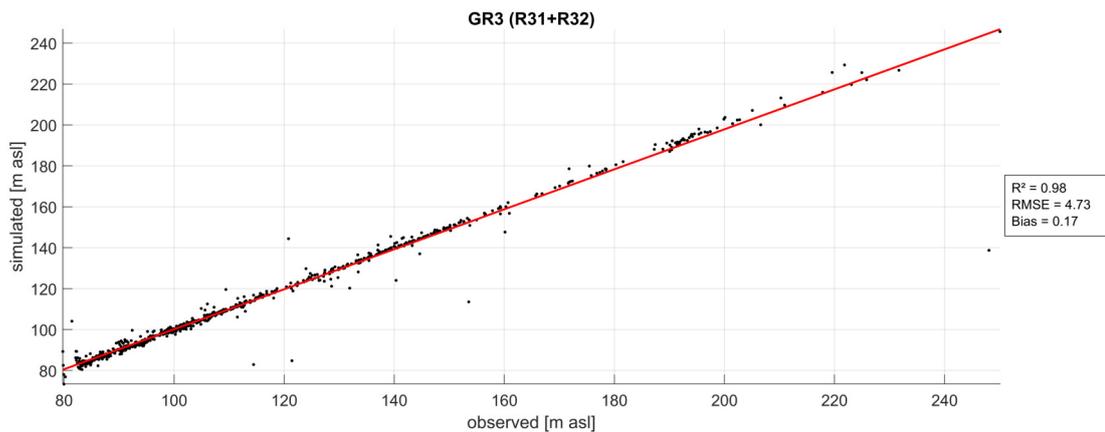


Abbildung 5.4: Scatterplot der Kreuzvalidierung für die Co-Kriging Grundwasseroberfläche in Großraum 3 ohne Raum 33

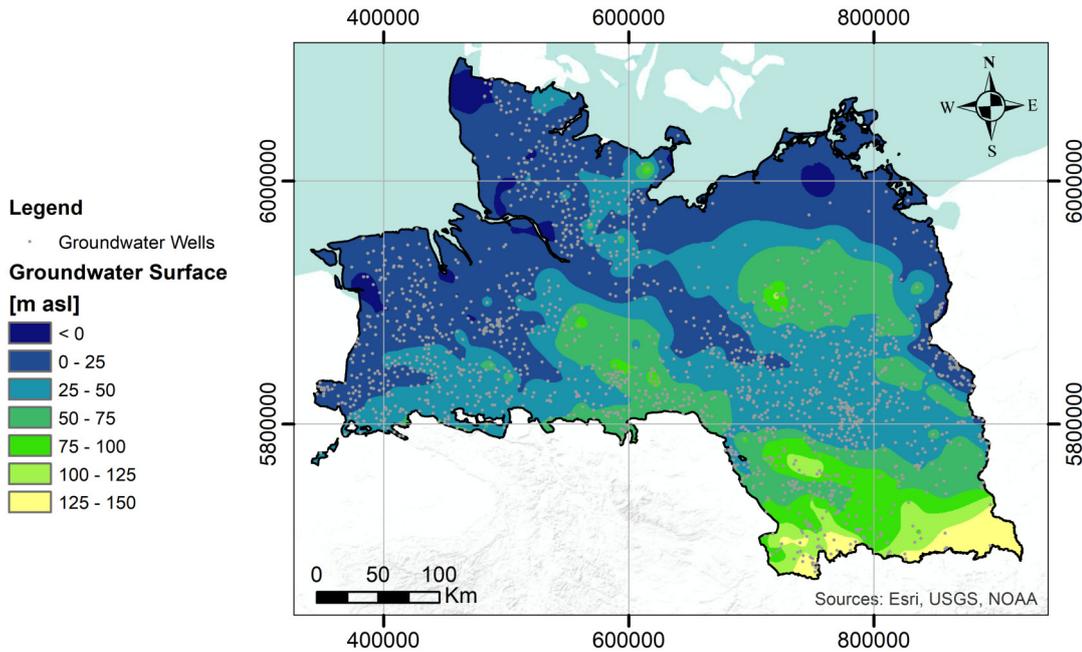


Abbildung 5.5: Co-Kriging Grundwasseroberfläche für Grobraum 1

gleichs zu erhöhen, wurde für Brandenburg zusätzlich eine gesonderte Co-Kriging Interpolation der Grundwasseroberfläche auf Basis der Daten von April 2011 durchgeführt. Abbildung 5.6a zeigt die hier erstellte Co-Kriging Grundwasseroberfläche, Abbildung 5.6b die auf Möhler et al. (2014) basierenden Grundwasser-Isolinien. In Abbildung 5.6c sind a und b überlagernd dargestellt. Im Detail sind nach wie vor größere Abweichungen der Absolutwerte zu finden (>10 m), dort wo ausreichend Messstellen vorhanden sind, kann die generelle Form der Grundwasseroberfläche zumindest auf regionalem Maßstab recht genau wiedergegeben werden, der Detailgrad ist jedoch deutlich niedriger als bei Möhler et al. (2014). Überhaupt nicht erfasst werden beispielsweise im Südosten die Tagebaubereiche, die in Abbildung 5.6b als dunkle Umrisse (Grundwasserabsenkungen) gut erkennbar sind.

In Mecklenburg-Vorpommern ergibt sich ein völlig anderes Bild im Vergleich mit der Grundwasseroberfläche aus Hilgert und Hennig (2017). Diese mittels geohydraulischer Modellierung in Verbindung mit Detrended Kriging erstellte Grundwasseroberfläche weicht in vielen Bereichen um mehrere Zehner Meter von der hier grob erstellten Oberfläche ab. Auch die allgemeine Form der beiden Grundwasseroberflächen stimmt kaum überein. Die Interpolation von Hilgert und Hennig (2017) basiert auf einem Datensatz mit durchschnittlich einem Messwert pro $2,4 \text{ km}^2$. Für das hier durchgeführte Co-Kriging konnte im Bereich von Mecklenburg-Vorpommern auf lediglich 51 Messstellen zurückgegriffen werden (entspricht einer Messstelle auf ca. 450 km^2). In den angrenzenden Gebieten benachbarter Bundesländer sind zwar noch Messstellen vorhanden, die das Ergebnis sicherlich positiv beeinflussen, insgesamt ist die erstellte Grundwasseroberfläche von Grobraum 1 im Bereich von Mecklenburg-Vorpommern jedoch in keinsten Weise belastbar oder interpretierbar. Ein Übertrag der Ansätze von Möhler et al. (2014) und

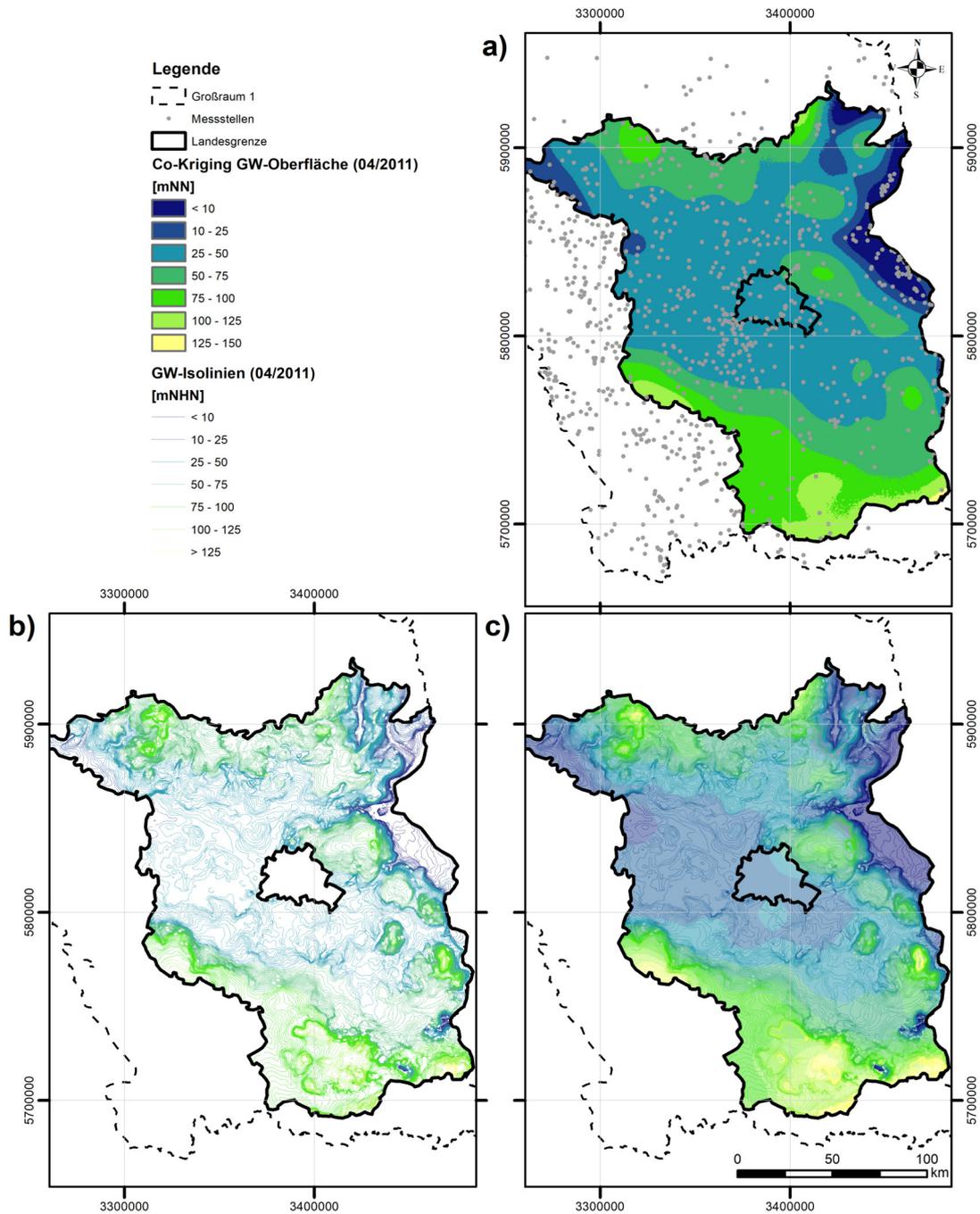


Abbildung 5.6: Vergleich der für April 2011 gesondert erstellten Co-Kriging Grundwasser-oberfläche (a) mit den auf Möhler et al. (2014) basierenden Grundwasser-Isolinien von April 2011 in Brandenburg (b). Abbildung (c) überlagert (a) und (b)

Hilgert und Hennig (2017) auf dieses Projekt ist wegen einer fehlenden Datenbasis nicht möglich.

Für die übrigen Bereiche von Großraum 1 sowie die Großräume 2 (Abb. 5.8) und 3 ohne Raum 33 (Abb. 5.9) sind die Ergebnisse der Regionalisierung mit ähnlicher Vorsicht zu betrachten. Auch wenn regional ausreichend Messstellen vorhanden sein sollten, um wie in Brandenburg die grobe Form der Grundwasseroberfläche gut abzubilden, so ist durchweg von einem geringen Detailgrad und lokal hohen Abweichungen der Absolutwerte auszugehen. Die belastbarsten Ergebnisse dürften sich für den Oberrheingraben ergeben. Dort ist die Messstellendichte recht hoch (ca. 1 Messstelle auf 10 km²) und die Hydrogeologie und die Topographie sind vergleichsweise simpel, sodass von einer annehmbaren Qualität der Regionalisierung ausgegangen werden kann. Eine Ausnahme hiervon bildet vor allem das Mainzer Becken (Nordwesten), wieder aufgrund von vergleichsweise wenigen Messstellen und komplexerer Geologie. Auch für den Oberrheingraben muss lokal jedoch mit signifikanten Unterschieden in der absoluten Höhe der Grundwasseroberfläche ausgegangen werden.

Auf eine detaillierte Darstellung der Flurabstände (analog Anhang L), wird verzichtet, da sich hierbei die Unsicherheit der Darstellung durch die Unsicherheit des digitalen Geländemodells noch erhöht. Für weitere Betrachtungen sowie regionale und lokale Interpretationen ist der hier berechnete Flurabstand demnach nicht geeignet.

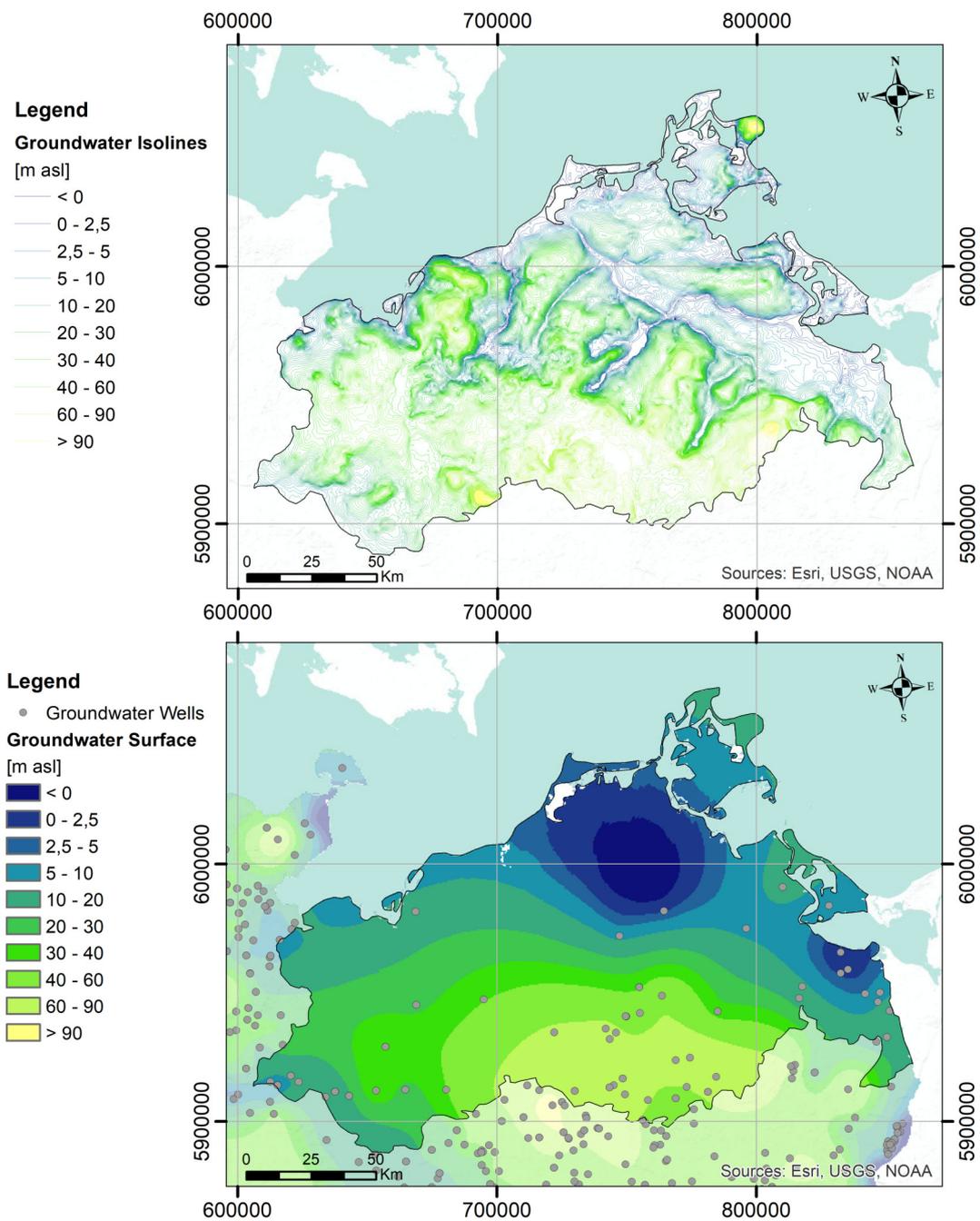


Abbildung 5.7: Vergleich der Grundwasserisolines aus Hilgert und Hennig (2017) (oben) mit der Co-Kriging Grundwasseroberfläche im Bereich von Mecklenburg-Vorpommern (unten).

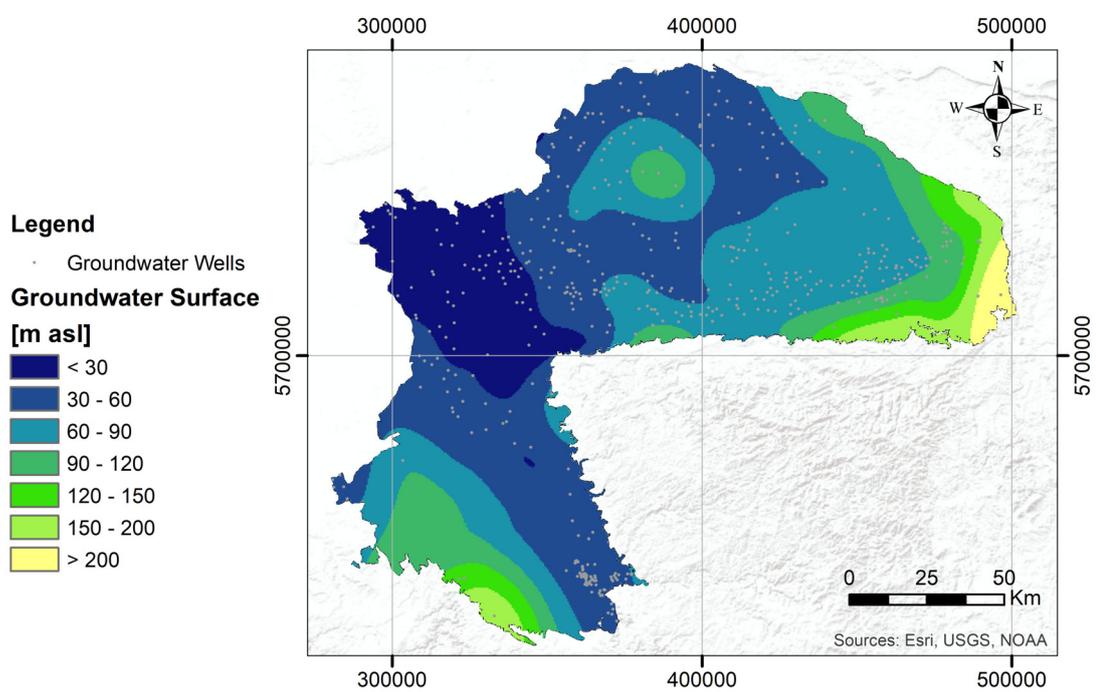


Abbildung 5.8: Co-Kriging Grundwasseroberfläche für Großraum 2

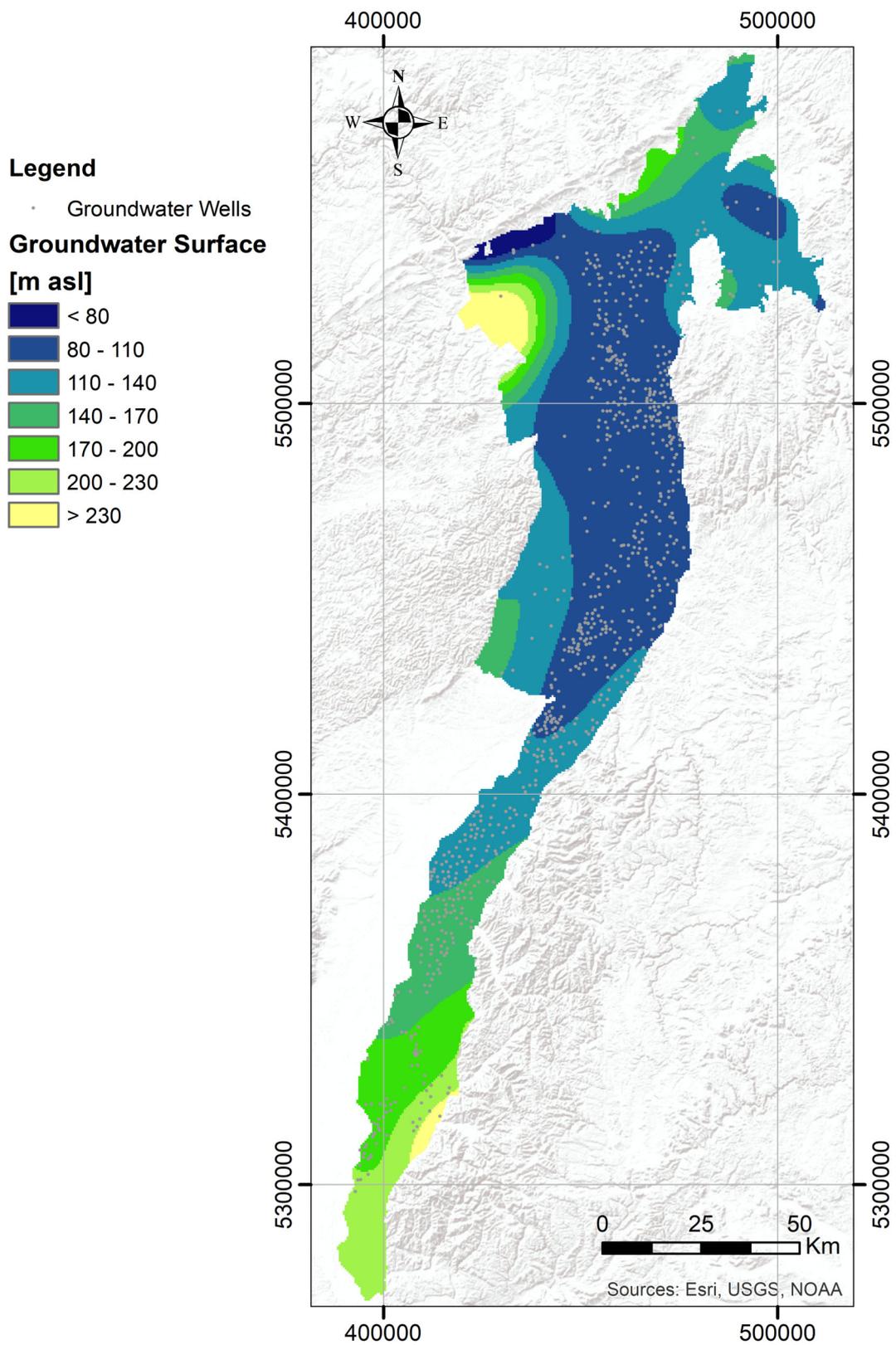


Abbildung 5.9: Co-Kriging Grundwasseroberfläche für die Räume 31 und 32 (Großraum 3 ohne Raum 33)

5.4 Zusammenfassung der Regionalisierung des mittleren Grundwasserflurabstands

In vorliegendem Zwischenbericht werden die Ergebnisse der Regionalisierung der mittleren Grundwasserhöhe und des mittleren Grundwasserflurabstands dargelegt. Als Datengrundlage für die Erstellung der Karten für ganz Deutschland dienen die langjährigen mittleren Grundwasserhöhen der in Kapitel 2 festgelegten Referenzmessstellen sowie der jeweils zugeordneten Clustermessstellen. Die Berechnung basiert auf dem mittleren Grundwasserstand der Jahre 2007 bis 2014. Zur Regionalisierung der Grundwasserstände wird auf die etablierte Interpolationsmethode des Co-Kriging mit einem SRTM Höhenmodell als Sekundärvariable zurückgegriffen. Da eine Interpolation in der Hydrogeologie über Aquifergrenzen hinweg meist nicht sinnvoll ist, dies jedoch aufgrund der unzureichenden Datenlage nicht umgesetzt werden konnte, wurden die Grenzen der hydrogeologischen Großräume (und teilweise auch Räume) als Interpolationsgrenzen festgelegt. Die entsprechenden Gebiete wurden einzeln interpoliert und evaluiert.

Die erzeugte Grundwasseroberfläche ist nur mit Vorsicht für Interpretationen heranzuziehen, da die Qualität der Interpolation regional stark schwankt und mit der Messstellendichte zusammenhängt. Die Belastbarkeit der Ergebnisse ist fast durchweg nicht ausreichend für eine detaillierte Interpretation oder regionale und lokale Betrachtungen sind. Auch in Bereichen mittlerer bis hoher Messstellendichte ist davon auszugehen, dass die absolute Abweichung der Interpolation lokal teils deutlich >10 m liegen kann. Wie am Beispiel von Brandenburg gezeigt werden konnte, ist die generelle Form der Grundwasseroberfläche regional im Bereich von Gebieten mit höherer Messstellendichte angemessen abbildbar, der Detailgrad bleibt jedoch unzufriedenstellend. Faktoren welche die Genauigkeit der Interpolation mutmaßlich negativ beeinflussen sind vor allem eine unzureichende Datendichte, komplexere Geologie oder auch stark variable Topographie. Die beiden letztgenannten Faktoren führen im Umkehrschluss dazu, dass vor allem dort eine höhere Datendichte nötig wäre, um ein realistischeres Bild von der Grundwasseroberfläche zu zeichnen. Im Datensatz ist jedoch eine gegenteilige Tendenz zu beobachten. Generell sind existierende Grundwassergleichenpläne nicht durch die Ergebnisse dieser Arbeit zu ersetzen.

Die Berechnung des Flurabstandes erfolgte durch Umrechnung der Grundwasseroberfläche mittels SRTM Höhenmodell. Zusätzlich zur Unsicherheit der Grundwasserhöhe ist auch der Fehler des Geländemodells zu berücksichtigen. Für zahlreiche Bereiche werden deutlich (>30 m) negative Grundwasserflurabstände berechnet. Es ist davon auszugehen, dass es sich dabei nicht um durchweg artesisch gespannte Verhältnisse, sondern um Interpolations- und in geringerem Maße Geländemodellfehler handelt. Aufgrund der Unsicherheit der Ergebnisse, sind diese nur in Anhang L dargestellt.

6 Zusammenfassung und Ausblick

Vorliegender Bericht beschreibt die Ergebnisse des Projektes Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze. Neben der Zusammenstellung und Homogenisierung einer umfangreichen Datengrundlage von insgesamt ca. 13.500 Messstellen umfasste die erste Projektphase (Kapitel 2) hauptsächlich die Entwicklung einer Methodik zur Auswahl von Referenzmessstellen. Diese basiert auf Self-Organizing-Maps und ähnelt einem klassischen Clusteransatz, bei dem Gruppen von Grundwasserganglinien mit ähnlicher Dynamik definiert werden. Grundlage sind hierbei für jede Ganglinie berechnete Features, welche bestimmte Aspekte der Zeitreihendynamik beschreiben. Weiterhin wurde auf die hydrogeologische Raumgliederung von Deutschland zurückgegriffen, um einen Kausalzusammenhang zwischen Ganglinien ähnlicher Dynamik zu gewährleisten. Der Gesamtdatensatz wurde in mehreren Arbeitsschritten auf ca. 5.500 Messstellen reduziert, die in 166 homogenen Clustern zusammengefasst sind. Entsprechend wurden auch 166 Referenzmessstellen mit repräsentativer Dynamik definiert, welche als Grundlage für die Entwicklung und Evaluierung des Vorhersageansatzes dienen. Die restlichen Messstellen wurden entweder im Verlauf des Preprocessings aufgrund mangelnder Datenqualität oder im Verlauf des Clusterings als Teil eines Clusters mit z.B. mutmaßlich anthropogener Beeinflussung, Ausreißern o.ä. aussortiert. Für die verbleibenden Cluster ist später eine Übertragung der Vorhersage von der RM auf die verbleibenden Clustermessstellen mit verschiedenen Ansätzen denkbar. Weiterhin erlaubt die räumliche Verteilung der Cluster weitergehende Interpretationen über die Dynamik beeinflussende Faktoren wie z.B. Oberflächengewässer, Grundwasserentnahmen etc. .

Zur Vorhersage der Grundwasserstände (Kapitel 3) wurde ein leicht übertragbarer Vorhersageansatz gewählt, welcher auf flächendeckend verfügbaren klimatischen Eingangsdaten (Niederschlag, Temperatur und relative Luftfeuchte) und rekurrenten KNN, genauer sog. NARX-Netzen basiert. Für jede der 166 RM wurden gesondert KNN-Modelle optimiert und trainiert. Die Evaluierung der Vorhersagegüte erfolgte für drei Vorhersagehorizonte: 1 Woche, 1 Monat und 3 Monate. Die Evaluierungszeiträume für die verschiedenen Horizonte waren 4 Jahre (2012-2015) für Wochen- und Monatsvorhersagen sowie 10 Jahre (2006-2015) im Falle von Jahreszeitemvorhersagen (3-Monatsvorhersagen). Die maximal erreichbare Vorhersagegüte für Grundwasserstände rein aufgrund der Modellanpassung wurde dabei auf Basis von beobachteten klimatischen Eingangsdaten (HYRAS-Daten) berechnet und für die Jahreszeitemvorhersagen der Güte von Vorhersagen auf Basis modellierter klimatischer Eingangsdaten (EPISODES Hindcasts) gegenübergestellt. Letztere erlauben die Beurteilung einer im operationellen Betrieb zu erwartenden Güte. Insgesamt können mit dem gewählten Ansatz für den überwiegenden Teil der Messstellen zufriedenstellende bis sehr gute Grundwasserstandsvorhersagen über alle Vorhersagehorizonte berechnet werden. Es existiert dabei sowohl eine Tendenz zu besseren Vorhersagen bei längeren Vorhersagehorizonten, als auch eine variable Güte mit saisonalem Verlauf. Im Falle von operationellen Vorhersagen ist ein deutlicher, jedoch auch häufig (je nach

Verwendungszweck) verkraftbarer Güterverlust der Grundwasserstandsvorhersagen zu erwarten. Für einen kleinen Teil der Modelle treten zudem in diesem Fall Modellinstabilitäten auf, die jedoch in der Gesamtbetrachtung über alle Messstellen kaum ins Gewicht fallen. Vor allem operationelle Vorhersagen für 1 und 3 Monate erscheinen mit dem entwickelten Ansatz auf überregionalem Maßstab sinnvoll. Für kürzere Vorhersagehorizonte ist der zuletzt gemessene Grundwasserstand ein starker Prädiktor der im Mittel nur schwer zu schlagen ist.

Im direkten Vergleich mit numerischen Modellen (Kapitel 4) zeigt sich, dass der entwickelte Ansatz deutliche Vorteile bietet. Für Messstellen, deren Dynamik über die Zeit grundsätzlich unverändert bleibt und die hauptsächlich von der meteorologischen Situation abhängt, kann der KNN-Ansatz eine bessere Vorhersagegüte bei gleichzeitig schnellerer Implementierung und geringerer notwendiger Datenmenge erzielen. Für gewöhnlich liegen auch der nötige Arbeitsaufwand für den Modellaufbau sowie die Berechnungsgeschwindigkeit der KNN weit über der numerischer Grundwassermodelle, hierbei spielt aber natürlich die Größe und die räumliche Auflösung des numerischen Modells eine entscheidende Rolle. Sind hingegen keine historischen Datenreihen der Einflussgrößen vorhanden oder sollen spezielle Situationen nachgebildet werden, die neben den meteorologischen Faktoren weitere Einflussgrößen abbilden und die beispielsweise auch eine Änderung der Grundwasserdynamik zur Folge haben, ist die Anwendung numerischer Grundwassermodelle weiterhin klar zu empfehlen. KNN sind in diesen speziellen Fällen nicht zur Vorhersage der Grundwasserstände geeignet.

In einem abschließenden Schritt (Kapitel 5) wurden Karten des mittleren Grundwasserflurabstandes sowie der absoluten Grundwasserhöhe in mNN für Deutschland erstellt. Dabei wurden auf Co-Kriging als bewährtes Interpolationsverfahren zur Regionalisierung und auf die hydrogeologische Raumlagerung von Deutschland (BGR und SGD, 2015) zur Abgrenzung räumlicher Einheiten zurückgegriffen. Die Datenbasis waren die Messstellen der 166 Cluster aus Kapitel 2. Während in Gebieten mit hoher Messstellendichte die generelle Form der Grundwasseroberfläche inklusive regionaler Strukturen teils gut abgebildet werden konnte, so war es für die meisten Bereiche nicht möglich eine belastbare Regionalisierung zu erzeugen. Die Genauigkeit ist demnach nicht ausreichend für eine detaillierte Interpretation oder regionale und lokale Betrachtungen. Da für die Berechnung des Flurabstandes noch zusätzlich der Fehler des Höhenmodells eine Rolle spielt und damit die Unsicherheit noch zusätzlich ansteigt, sind die Karten generell nur mit Vorsicht zu betrachten.

Obwohl mit NARX in diesem Projekt gute Ergebnisse erzielt wurden, ist die Einbeziehung neuester Erkenntnisse auf dem Gebiet des Deep-Learning zur weiteren Verbesserung der Grundwasserstandsvorhersagen zukünftig zu empfehlen. Deep-Learning unterscheidet sich von flachen KNN vereinfacht ausgedrückt durch eine höhere Anzahl an verdeckten Schichten. Dies ermöglicht in vielen Anwendungen eine eigenständige Feature-Erkennung im Trainingsdatensatz mit geringen oder sogar gänzlich ohne Vorwissen. Hierdurch kann ein klassisches Feature-Engineering, wie in den Datenwissenschaften üblich und auch in diesem Projekt durchgeführt, überflüssig werden. Hierbei sind in erster Linie Long Short-Term Memory (LSTM) Modelle sowie Convolutional Neural Networks (CNN) zu nennen. LSTMs sind, wie auch NARX, rekurrente KNN, die im Gegensatz zu anderen RNNs, in der Lage sind, Langfristabhängigkeiten in Zeitreihen zu erfassen, da sie explizit zur Überwindung des sog. „vanishing gradient“-Problems entwickelt wurden

(Hochreiter und Schmidhuber, 1997). Dies kann für die Vorhersage von Grundwasserständen von Nutzen sein, da diese häufig von langfristigen klimatischen Schwankungen beeinflusst sind (z.B. Liesch und Wunsch, 2019). CNNs sind neuronale Netze, die vorwiegend zur Bilderkennung und -klassifikation eingesetzt werden, da sie sich vor allem für Daten mit räumlicher Struktur eignen. Sie eignen sich jedoch z.B. als 1D-CNN ebenso für Signalverarbeitungsaufgaben und werden z.B. für Natural Language Processing eingesetzt und haben in der Zeitreihenvorhersage bereits ihre Leistungsfähigkeit unter Beweis gestellt. Insgesamt zeigen Studien in jüngster Vergangenheit (z.B. Afzaal et al. 2020; Bowes et al. 2019; Jeong und Park, 2019; Jeong et al. 2020; Müller et al. 2020), dass sowohl LSTMs als auch CNNs sehr gut für die Vorhersage von Grundwasserständen geeignet sind und damit eine erfolgversprechende Anwendung erwarten lassen. Weiterhin ist auch eine Nutzung der Clusterergebnisse zur Vorhersage direkt denkbar. So beruht der Ansatz von Bandara et al. (2018) ganz konkret auf homogenen Gruppen von Zeitreihen, die durch ein vorangestelltes Clustering identifiziert wurden.

Neben der in diesem Projekt durchgeführten Einzelwertvorhersage (Sequence-to-Value), sollte zukünftig auch auf eine Sequenzvorhersage (Sequence-to-Sequence) gesetzt werden. Damit ist eine direkte Vorhersage einer ganzen Sequenz, ohne explizite Inputs für jeden Vorhersageschritt innerhalb der Sequenz möglich. Dies scheint vor allem im Kontext der operationellen Vorhersage vielversprechend, da keine meteorologischen Modelldaten mehr zur Vorhersage benötigt werden, deren Unsicherheiten die Vorhersageergebnisse beeinflussen. Weiterhin ermöglicht dies auch die Einbeziehung von Eingangsdaten, die an sich nicht vorhergesagt werden, für die aber an der entsprechenden Messstelle deutliche Abhängigkeiten bestehen (z.B. Wasserentnahmeraten, Pegelstände von Oberflächengewässern). Aus diesem Grund ist für den gesamten Vorhersageansatz eine Einbeziehung weiterer Eingabegrößen zur Steigerung der Vorhersagegüte zukünftig empfehlenswert, sodass auch für Messstellen, die bisher eine nicht zufriedenstellende Vorhersagegüte zeigen, eine deutliche Verbesserung erzielt werden kann. Diese Maßnahme lässt eine Vergrößerung der Datenbasis, welche für eine zukünftige Regionalisierung genutzt werden kann erwarten. Ganz unabhängig davon, ist eine solche für diesen Schritt auch unbedingt zu empfehlen. Der offensichtliche Zusammenhang zwischen der Güte der Regionalisierung und der Messstellendichte sollte Anlass sein, die Gesamtdatenbasis durch die Einbeziehung zusätzlicher Messstellen zu vergrößern. Neben den bereits verwendeten etablierten Interpolationsmethoden, erscheinen auch zur Regionalisierung DL-Ansätze vielversprechend und lassen einen deutlichen Mehrwert erwarten. Neben den oben genannten Ansätzen, ist an dieser Stelle vor allem die direkte Kombination von CNN und LSTM (z.B. ConvLSTM) interessant, welche in der Regel zur räumlichen Vorhersage von Zeitreihen zum Einsatz (z.B. Asadi und Regan, 2019) kommt. Ein Einsatz für Grundwasserstandsdaten ist bisher nicht bekannt, erscheint jedoch aufgrund der erfolgreichen Anwendung in methodisch verwandten Problemstellungen (z.B. Vorhersage von Niederschläge oder Abfluss) durchaus erfolgversprechend. Weitere denkbare Ansätze umfassen beispielsweise 3D-CNNs, welche sich besonders zur Analyse von Daten mit Raumstruktur und zeitlicher Abhängigkeit eignen (z.B. Human Action Recognition, Medical Imaging) oder auch LSTM-basierte Frameworks wie das Incremental Dual-Memory LSTM (Jia et al. 2017), welche ebenfalls bereits erfolgreich in methodisch verwandten Problemstellungen wie der räumlichen Vorhersage von Landnutzungen eingesetzt wurde.

Literatur

- Ad-hoc-Arbeitsgruppe Hydrogeologie, Hrsg. (2016). Regionale Hydrogeologie von Deutschland. Bd. A 163. Geologisches Jahrbuch. Stuttgart, Germany: Schweizerbart Science Publishers. ISBN: 978-3-510-96852-7.
- Adamowski, J. und H. F. Chan (2011). A Wavelet Neural Network Conjunction Model for Groundwater Level Forecasting. In: *Journal of Hydrology* 407.1-4, S. 28–40. ISSN: 00221694. DOI: 10/cr77bs.
- Afzaal, H., A. A. Farooque, F. Abbas, B. Acharya und T. Esau (2020). Groundwater Estimation from Major Physical Hydrology Components Using Artificial Neural Networks and Deep Learning. In: *Water* 12.1, S. 5. DOI: 10/ggvmqj.
- Alizamir, M., O. Kisi und M. Zounemat-Kermani (2018). Modelling Long-Term Groundwater Fluctuations by Extreme Learning Machine Using Hydro-Climatic Data. In: *Hydrological Sciences Journal* 63.1, S. 63–73. ISSN: 0262-6667, 2150-3435. DOI: 10/gfvfg2.
- Alqurashi, T. M. (2017). Thesis: Clustering Ensemble Method. Diss. University of East Anglia.
- Alqurashi, T. M. und W. Wang (2018). Clustering Ensemble Method. In: *International Journal of Machine Learning and Cybernetics*. ISSN: 1868-8071, 1868-808X. DOI: 10/gft8dg.
- Asadi, R. und A. Regan (2019). A Spatial-Temporal Decomposition Based Deep Neural Network for Time Series Forecasting. In: *arXiv:1902.00636 [cs, stat]*. arXiv: 1902.00636 [cs, stat].
- Ayad, H. G. und M. S. Kamel (2010). On Voting-Based Consensus of Cluster Ensembles. In: *Pattern Recognition* 43.5, S. 1943–1953. ISSN: 0031-3203. DOI: 10/b9tz3r.
- Bandara, K., C. Bergmeir und S. Smyl (2018). Forecasting Across Time Series Databases Using Recurrent Neural Networks on Groups of Similar Series: A Clustering Approach. In: *arXiv:1710.03222 [cs, econ, stat]*. arXiv: 1710.03222 [cs, econ, stat].
- Barreto, G. A. (2007). Time Series Prediction with the Self-Organizing Map: A Review. In: *Perspectives of Neural-Symbolic Integration*. Springer, S. 135–158.
- Barzegar, R., E. Fijani, A. Asghari Moghaddam und E. Tziritis (2017). Forecasting of Groundwater Level Fluctuations Using Ensemble Hybrid Multi-Wavelet Neural Network-Based Models. In: *Science of The Total Environment* 599-600, S. 20–31. ISSN: 00489697. DOI: 10/gcx4pr.
- Beale, H. M., M. T. Hagan und H. B. Demuth (2016). Neural Network Toolbox™ User's Guide: Revised for Version 9.1 (Release 2016b).
- Bengio, Y., P. Simard und P. Frasconi (1994). Learning Long-Term Dependencies with Gradient Descent Is Difficult. In: *IEEE Trans. Neural Netw.* 5.2, S. 157–166. ISSN: 1045-9227, 1941-0093. DOI: 10/bmf3dr.
- BGR und SGD (2015). Hydrogeologische Raumgliederung von Deutschland. URL: www.bgr.bund.de/hyraum.
- Böttcher, S., C. Merz, G. Lischeid und R. Dannowski (2014). Using Isomap to Differentiate between Anthropogenic and Natural Effects on Groundwater Dynamics in a Com-

- plex Geological Setting. In: *Journal of Hydrology* 519, S. 1634–1641. ISSN: 00221694. DOI: 10/f6tn4x.
- Bowes, B. D., J. M. Sadler, M. M. Morsy, M. Behl und J. L. Goodall (2019). Forecasting Groundwater Table in a Flood Prone Coastal City with Long Short-Term Memory and Recurrent Neural Networks. In: *Water* 11.5, S. 1098. ISSN: 2073-4441. DOI: 10.3390/w11051098.
- Bucher, B. (1999). Die Analyse von Grundwasserganglinien Mit Dem Wiener-Mehrkanal-Filter. In: *Grundwasser* 4.3, S. 113–118. DOI: 10/ctnmbp.
- Buehren, M. (2011). Functions for the Rectangular Assignment Problem - File Exchange - MATLAB Central. <https://de.mathworks.com/matlabcentral/fileexchange/6543>.
- Cabanes, G., Y. Bennani und D. Fresneau (2012). Enriched Topological Learning for Cluster Detection and Visualization. In: *Neural Networks*. Selected Papers from IJCNN 2011 32, S. 186–195. ISSN: 0893-6080. DOI: 10/f3z7z8.
- Caiado, J., E. A. Maharaj und P. D’urso (2015). Time-Series Clustering. In: *Handbook of Cluster Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton London New York, S. 241–264. ISBN: 978-1-4665-5189-3 978-1-4665-5188-6.
- Caliński, T. und J. Harabasz (1974). A Dendrite Method for Cluster Analysis. In: *Communications in Statistics* 3.1, S. 1–27. ISSN: 0090-3272. DOI: 10/dsds7k.
- Chang, F.-J., L.-C. Chang, C.-W. Huang und I.-F. Kao (2016). Prediction of Monthly Regional Groundwater Levels through Hybrid Soft-Computing Techniques. In: *Journal of Hydrology* 541, S. 965–976. ISSN: 00221694. DOI: 10/f8825x.
- Chitsazan, M., G. Rahmani und A. Neyamadpour (2015). Forecasting Groundwater Level by Artificial Neural Networks as an Alternative Approach to Groundwater Modeling. In: *Journal of the Geological Society of India* 85.1, S. 98–106. DOI: 10/f7dsmp.
- Coppola, E., M. Poulton, E. Charles, J. Dustman und F. Szidarovszky (2003). Application of Artificial Neural Networks to Complex Groundwater Management Problems. In: *Natural Resources Research* 12.4, S. 303–320. DOI: 10/cjwg2d.
- Corona, C. R., J. J. Gurdak, J. E. Dickinson, T. Ferré und E. P. Maurer (2018). Climate Variability and Vadose Zone Controls on Damping of Transient Recharge. In: *Journal of Hydrology* 561, S. 1094–1104. ISSN: 00221694. DOI: 10/gdzw3x.
- Coulibaly, P., F. Anctil, R. Aravena und B. Bobée (2001). Artificial Neural Network Modeling of Water Table Depth Fluctuations. In: *Water resources research* 37.4, S. 885–896. DOI: 10/cjq8mg.
- Daliakopoulos, I. N., P. Coulibaly und I. K. Tsanis (2005). Groundwater Level Forecasting Using Artificial Neural Networks. In: *Journal of Hydrology* 309.1-4, S. 229–240. ISSN: 00221694. DOI: 10/crzknh.
- Desgraupes, B. (2018). clusterCrit: Clustering Indices.
- DWD (2017). Testreferenzjahre (TRY). URL: <https://www.dwd.de/DE/leistungen/testreferenzjahre/testreferenzjahre.html>.
- (2018). Seasonal Forecasting. URL: https://www.dwd.de/EN/ourservices/seasonals/_forecasts/project/_description.html?nn=495490&lsbId=619784.
- (2019). EPISODES. URL: <https://www.dwd.de/DE/leistungen/episodes/episodes.html?nn=645608>.
- Frick, C., H. Steiner, A. Mazurkiewicz, U. Riediger, M. Rauthe, T. Reich und A. Gratzki (2014). Central European High-Resolution Gridded Daily Data Sets (HYRAS): Mean Temperature and Relative Humidity. In: *Meteorologische Zeitschrift* 23.1, S. 15–32. ISSN: 0941-2948. DOI: 10/f6n4g3.

- Geman, S., E. Bienenstock und R. Doursat (1992). Neural Networks and the Bias/Variance Dilemma. In: *Neural Computation* 4.1, S. 1–58. ISSN: 0899-7667, 1530-888X. DOI: 10/dz9dd3.
- Ghosh, J. und A. Acharya (2011). Cluster Ensembles. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.4, S. 305–315. ISSN: 1942-4795. DOI: 10/cgwrhm.
- Gorgij, A. D., O. Kisi und A. A. Moghaddam (2017). Groundwater Budget Forecasting, Using Hybrid Wavelet-ANN-GP Modelling: A Case Study of Azarshahr Plain, East Azerbaijan, Iran. In: *Hydrology Research* 48.2, S. 455–467. ISSN: 0029-1277, 2224-7955. DOI: 10/f936sm.
- Grossmann, J. und F. Skowronek (2005). Quantifizierung anthropogener Veränderungen der Grundwasserstände mit dem Wiener-Mehrkanal-Filter. In: *Grundwasser* 10.3, S. 157–166. ISSN: 1430-483X, 1432-1165. DOI: 10/c7w2qc.
- Gupta, H. V., H. Kling, K. K. Yilmaz und G. F. Martinez (2009). Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling. In: *Journal of Hydrology* 377.1, S. 80–91. ISSN: 0022-1694. DOI: 10/c2hqxw.
- Guzman, S. M., J. O. Paz und M. L. M. Tagert (2017). The Use of NARX Neural Networks to Forecast Daily Groundwater Levels. In: *Water Resources Management* 31.5, S. 1591–1603. ISSN: 0920-4741, 1573-1650. DOI: 10/f92hfc.
- Guzman, S. M., J. O. Paz, M. L. M. Tagert und A. E. Mercer (2019). Evaluation of Seasonally Classified Inputs for the Prediction of Daily Groundwater Levels: NARX Networks Vs Support Vector Machines. In: *Environ. Model. Assess.* 24.2, S. 223–234. ISSN: 1420-2026. DOI: 10/gf37xk.
- Hanczar, B. und M. Nadif (2013). Study of Consensus Functions in the Context of Ensemble Methods for Biclustering.
- Hilgert, T. und H. Hennig (2017). Grundwasserfließgeschehen Mecklenburg-Vorpommerns – Geohydraulische Modellierung mit Detrended Kriging. In: *Grundwasser* 22.1, S. 17–29. ISSN: 1430-483X, 1432-1165. DOI: 10/gg8qx9.
- Hochreiter, S. und J. Schmidhuber (1997). Long Short-Term Memory. In: *Neural Computation* 9.8, S. 1735–1780. ISSN: 0899-7667, 1530-888X. DOI: 10/bxd65w.
- Hubert, L. und J. Schultz (1976). Quadratic Assignment as a General Data Analysis Strategy. In: *British Journal of Mathematical and Statistical Psychology* 29.2, S. 190–241. ISSN: 2044-8317. DOI: 10/dm98dn.
- Izady, A., K. Davary, A. Alizadeh, A. Moghaddamnia, A. N. Ziaei und S. M. Hashemina (2013). Application of NN-ARX Model to Predict Groundwater Levels in the Neishaboor Plain, Iran. In: *Water Resources Management* 27.14, S. 4773–4794. ISSN: 0920-4741, 1573-1650. DOI: 10/f5fngx.
- Jain, A. K. und R. C. Dubes (1988). Algorithms for Clustering Data. Prentice Hall Advanced Reference Series. Englewood Cliffs, N.J: Prentice Hall. ISBN: 978-0-13-022278-7.
- Jeong, J. und E. Park (2019). Comparative Applications of Data-Driven Models Representing Water Table Fluctuations. In: *Journal of Hydrology* 572, S. 261–273. ISSN: 00221694. DOI: 10/gfx567.
- Jeong, J., E. Park, H. Chen, K.-Y. Kim, W. Shik Han und H. Suk (2020). Estimation of Groundwater Level Based on the Robust Training of Recurrent Neural Networks Using Corrupted Data. In: *Journal of Hydrology* 582, S. 124512. ISSN: 00221694. DOI: 10/ggkkqw.

- Jha, M. K. und S. Sahoo (2015). Efficacy of Neural Network and Genetic Algorithm Techniques in Simulating Spatio-Temporal Fluctuations of Groundwater: In: *Hydrological Processes* 29.5, S. 671–691. ISSN: 08856087. DOI: 10/f6388q.
- Jia, X., A. Khandelwal, G. Nayak, J. Gerber, K. Carlson, P. West und V. Kumar (2017). Incremental Dual-Memory LSTM in Land Cover Prediction. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, S. 867–876. ISBN: 978-1-4503-4887-4. DOI: 10/gg38b7.
- JPL NASA (2013a). NASA Shuttle Radar Topography Mission Global 1 Arc Second. DOI: 10.5067/MEASURES/SRTM/SRTMGL1.003.
- (2013b). NASA Shuttle Radar Topography Mission Global 3 Arc Second. DOI: 10.5067/MEASURES/SRTM/SRTMGL3.003.
- Khalil, B., S. Broda, J. Adamowski, B. Ozga-Zielinski und A. Donohoe (2015). Short-Term Forecasting of Groundwater Levels under Conditions of Mine-Tailings Recharge Using Wavelet Ensemble Neural Network Models. In: *Hydrogeology Journal* 23.1, S. 121–141. ISSN: 1431-2174, 1435-0157. DOI: 10/f6ws98.
- Kisi, O., M. Alizamir und M. Zounemat-Kermani (2017). Modeling Groundwater Fluctuations by Three Different Evolutionary Neural Network Techniques Using Hydroclimatic Data. In: *Natural Hazards* 87.1, S. 367–381. ISSN: 0921-030X, 1573-0840. DOI: 10/gcx4qn.
- Kitanidis, P. K. und R. L. Bras (1980). Real-Time Forecasting with a Conceptual Hydrologic Model: 2. Applications and Results. In: *Water Resour. Res.* 16.6, S. 1034–1044. ISSN: 00431397. DOI: 10/d4tw78.
- Kohonen, T. (1987). Adaptive, Associative, and Self-Organizing Functions in Neural Computing. In: *Applied Optics* 26.23, S. 4910–4918. DOI: 10/dfgkbb.
- (1990). The Self-Organizing Map. In: *Proceedings of the IEEE* 78.9, S. 1464–1480. ISSN: 0018-9219. DOI: 10/ckgfr3.
- (2014). Matlab Implementations and Applications of the Self-Organizing Map. Helsinki. ISBN: 978-952-60-3679-3.
- Kong A Siou, L., A. Johannet, V. Borrell und S. Pistre (2011). Complexity Selection of a Neural Network Model for Karst Flood Forecasting: The Case of the Lez Basin (Southern France). In: *Journal of Hydrology* 403.3-4, S. 367–380. ISSN: 00221694. DOI: 10/bsg7sv.
- Krause, P., D. P. Boyle und F. Bäse (2005). Comparison of Different Efficiency Criteria for Hydrological Model Assessment. In: *Advances in geosciences* 5, S. 89–97. DOI: 10/czzds4.
- Kreienkamp, F., A. Paxian, B. Früh, P. Lorenz und C. Matulla (2019). Evaluation of the Empirical–Statistical Downscaling Method EPISODES. In: *Clim Dyn* 52.1, S. 991–1026. ISSN: 1432-0894. DOI: 10/gfxvfm.
- Kuhn, H. W. (1955). The Hungarian Method for the Assignment Problem. In: *Naval Research Logistics Quarterly* 2.1-2, S. 83–97. ISSN: 1931-9193. DOI: 10/b2k5tg.
- Lallahem, S., J. Mania, A. Hani und Y. Najjar (2005). On the Use of Neural Networks to Evaluate Groundwater Levels in Fractured Media. In: *Journal of Hydrology* 307.1-4, S. 92–111. ISSN: 00221694. DOI: 10/cdk48d.
- Liesch, T. und A. Wunsch (2019). Aquifer Responses to Long-Term Climatic Periodicities. In: *Journal of Hydrology* 572, S. 226–242. ISSN: 00221694. DOI: 10/gfwjhb.
- Lin, T., B. G. Horne, P. Tiño und C. L. Giles (1996). Learning Long-Term Dependencies in NARX Recurrent Neural Networks. In: *IEEE Transactions on Neural Networks* 7.6, S. 1329–1338.

- Lischeid, G., J. Steidl und C. Merz (2012). Funktionalanalyse versus Trendanalyse zur Abschätzung anthropogener Einflüsse auf Grundwasserganglinien. In: *Grundwasser* 17.2, S. 79–89. ISSN: 1430-483X, 1432-1165. DOI: 10/fzmn9j.
- LUBW (2018). Grundwassermodell Karlsruhe - Worms: Analyse Und Prognose Der TFA-Belastung Im Rhein-Neckar-Raum. Techn. Ber. Karlsruhe: LUBW Landesanstalt für Umwelt Baden-Württemberg, S. 109.
- (2020). LOGAR. URL: <https://www.lubw.baden-wuerttemberg.de/wasser/logar/#Modellsystem>.
- Mathworks Inc. (2018). Matlab 2018b.
- McClain, J. O. und V. R. Rao (1975). CLUSTISZ: A Program to Test for the Quality of Clustering of a Set of Objects. In: *Journal of Marketing Research* 12.4, S. 456–460. ISSN: 0022-2437.
- Merk, M. (2015). (unveröffentl.) Quantifizierung der Austauschdynamik zwischen Grund- und Oberflächenwasser im Anschluss an eine Deichrückverlegung.
- Mohammadi, K. (2008). Groundwater Table Estimation Using MODFLOW and Artificial Neural Networks. In: *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*. Water Science and Technology Library v. 68. Berlin: Springer. ISBN: 978-3-540-79880-4.
- Möhler, F., S. Dinse und A. Hermsdorf (2014). Grundwassergleichenplan für Brandenburg – Interpolation mittels Kriging mit externer Drift. In: *Grundwasser* 19.3, S. 189–199. ISSN: 1432-1165. DOI: 10/gg85cp.
- Moriasi, D. N., J. G. Arnold, M. W. V. Liew, R. L. Bingner, R. D. Harmel und T. L. Veith (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. In: *Transactions of the ASABE* 50.3, S. 885–900. ISSN: 2151-0040. DOI: 10/gcpcmf.
- Moriasi, D. N., M. W. Gitau, N. Pai und P. Daggupati (2015). Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. In: *Transactions of the ASABE* 58.6, S. 1763–1785. ISSN: 2151-0032. DOI: 10/f774s6.
- Müller, J., J. Park, R. Sahu, C. Varadharajan, B. Arora, B. Faybishenko und D. Agarwal (2020). Surrogate Optimization of Deep Neural Networks for Groundwater Predictions. In: *J Glob Optim*. ISSN: 1573-2916. DOI: 10/gg7z7g. arXiv: 1908.10947.
- Munkres, J. (1957). Algorithms for the Assignment and Transportation Problems. In: *Journal of the Society for Industrial and Applied Mathematics* 5.1, S. 32–38. ISSN: 0368-4245. DOI: 10/dxnjhs.
- Nayak, P. C., Y. R. S. Rao und K. P. Sudheer (2006). Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. In: *Water Resources Management* 20.1, S. 77–90. ISSN: 0920-4741, 1573-1650. DOI: 10/d4kqh2.
- Ohmer, M., T. Liesch, N. Goeppert und N. Goldscheider (2017). On the Optimal Selection of Interpolation Methods for Groundwater Contouring: An Example of Propagation of Uncertainty Regarding Inter-Aquifer Exchange. In: *Advances in Water Resources* 109, S. 121–132. ISSN: 03091708. DOI: 10/gcnbs8.
- Pakhira, M. K., S. Bandyopadhyay und U. Maulik (2004). Validity Index for Crisp and Fuzzy Clusters. In: *Pattern Recognition* 37.3, S. 487–501. ISSN: 0031-3203. DOI: 10/c78kjs.
- Rajae, T., H. Ebrahimi und V. Nourani (2019). A Review of the Artificial Intelligence Methods in Groundwater Level Modeling. In: *Journal of Hydrology*. ISSN: 00221694. DOI: 10/gfvfg3.

- Ratkowsky, D. und G. Lance (1978). A Criterion for Determining the Number of Groups in a Classification. In: *Australian Computer Journal* 10.3, S. 115–117.
- Rauthe, M., H. Steiner, U. Riediger, A. Mazurkiewicz und A. Gratzki (2013). A Central European Precipitation Climatology – Part I: Generation and Validation of a High-Resolution Gridded Daily Data Set (HYRAS). In: *Meteorol. Z.*, S. 22. DOI: 10/f5gf49.
- Région Alsace (2012). Abschlussbericht Projekt Interreg IV "Länderübergreifende Organisation Für Grundwasserschutz Am Rhein (LOGAR)". Techn. Ber.
- Ritter, A. und R. Muñoz-Carpena (2013). Performance Evaluation of Hydrological Models: Statistical Significance for Reducing Subjectivity in Goodness-of-Fit Assessments. In: *Journal of Hydrology* 480, S. 33–45. ISSN: 00221694. DOI: 10/f4pksm.
- Sahoo, S., T. A. Russo, J. Elliott und I. Foster (2017). Machine Learning Algorithms for Modeling Groundwater Level Changes in Agricultural Regions of the U.S.: MACHINE LEARNING GROUNDWATER MODEL. In: *Water Resources Research* 53.5, S. 3878–3895. ISSN: 00431397. DOI: 10/gbk22w.
- Shestakov, A. (2017). Consensus Clustering Experiment Framework.
- Sinharay, S. (2010). Jackknife Methods. In: *International Encyclopedia of Education (Third Edition)*. Hrsg. von P. Peterson, E. Baker und B. McGaw. Oxford: Elsevier, S. 229–231. ISBN: 978-0-08-044894-7. DOI: 10.1016/B978-0-08-044894-7.01338-5.
- Sreekanth, P. D., N. Geethanjali, P. D. Sreedevi, S. Ahmed, N. R. Kumar und P. K. Jayanthi (2009). Forecasting Groundwater Level Using Artificial Neural Networks. In: *Current science*, S. 933–939.
- Taormina, R., K.-w. Chau und R. Sethi (2012). Artificial Neural Network Simulation of Hourly Groundwater Levels in a Coastal Aquifer System of the Venice Lagoon. In: *Engineering Applications of Artificial Intelligence* 25.8, S. 1670–1676. ISSN: 09521976. DOI: 10/f4fz7w.
- Trichakis, I. C., I. K. Nikolos und G. P. Karatzas (2011). Artificial Neural Network (ANN) Based Modeling for Karstic Groundwater Level Simulation. In: *Water Resources Management* 25.4, S. 1143–1152. ISSN: 0920-4741, 1573-1650. DOI: 10/fjq8zx.
- Vega-Pons, S. und J. Ruiz-Shulcloper (2011). A Survey of Clustering Ensemble Algorithms. In: *Int. J. Patt. Recogn. Artif. Intell.* 25.03, S. 337–372. ISSN: 0218-0014. DOI: 10/bewtkr.
- Vesanto, J. (2000). SOM Toolbox for Matlab 5. Espoo: Helsinki University of Technology. ISBN: 978-951-22-4951-0.
- (2002). Data Exploration Process Based on the Self-Organizing Map. Acta Polytechnica Scandinavica Mathematics and Computing Series 115. Espoo: Finnish Acad. of Technology. ISBN: 978-951-666-596-5.
- (2005). SOM Toolbox: Implementation of the Algorithm. URL: <http://www.cis.hut.fi/projects/somtoolbox/documentation/somalg.shtml>.
- Wang, X., K. A. Smith und R. J. Hyndman (2005). Dimension Reduction for Clustering Time Series Using Global Characteristics. In: *International Conference on Computational Science*. Springer, S. 792–795. DOI: 10/fgcj2h.
- (2006). Characteristic-Based Clustering for Time Series Data. In: *Data Mining and Knowledge Discovery* 13.3, S. 335–364. ISSN: 1384-5810, 1573-756X. DOI: 10/c7zhtf.
- Wang, X., K. A. Smith, R. J. Hyndman und D. Alahakoon (2004). A Scalable Method for Time Series Clustering. In: *Technical Report, Monash University*.
- Wingering, M. (1999). Die Anwendung Der Clusteranalyse Bei Der Auswahl Repräsentativer Grundwassermeßstellen in Baden-Württemberg. In: *HW* 43.4. ISSN: 1439-1783.

- Wriedt, G. (2017). Verfahren zur Analyse klimatischer und anthropogener Einflüsse auf die Grundwasserstandsentwicklung. In: *Grundwasser* 22.1, S. 41–53. ISSN: 1430-483X, 1432-1165. DOI: 10/gcx4qv.
- Wunsch, A., T. Liesch und S. Broda (2018). Forecasting Groundwater Levels Using Nonlinear Autoregressive Networks with Exogenous Input (NARX). In: *Journal of Hydrology* 567, S. 743–758. ISSN: 00221694. DOI: 10/gcx5k4.
- Yue, S.-h., P. Li, J.-d. Guo und S.-g. Zhou (2004). Using Greedy Algorithm: DBSCAN Revisited II. In: *J. Zhejiang Univ.-Sci.* 5.11, S. 1405–1412. ISSN: 1673-565X, 1862-1775. DOI: 10/bdmcv2.

Anhang

Gliederung:

- **Anhang A:** Algorithmenkonfigurationen Referenzmessstellenauswahl
- **Anhang B:** Berechnung der Cluster-Validierungs-Indizes
- **Anhang C:** Liste aller Referenzmessstellen
- **Anhang D:** Feature-Boxplots
- **Anhang E:** Initiale und verbleibende Messstellendichten
- **Anhang F:** Potentielle Eignung der räumlichen Einheiten zur späteren Vorhersage
- **Anhang G:** Ergebnisse der bayesischen Hyperparameteroptimierung
- **Anhang H:** Fehlerübersicht aller Referenzmessstellen und aller Vorhersagen
- **Anhang I:** Übersichtskarten bez. der Vorhersagequalität einzelner (ausgewählter) Monate für alle Vorhersagehorizonte
- **Anhang J:** Liste von Messstellen mit Instabilitäten bei EPISODES-basierten Jahreszeitemvorhersagen
- **Anhang K:** Modellvergleich der Vorhersagegüte an den Messstellen
- **Anhang L:** Co-Kriging Ergebnisse Deutschland
- **Anhang M:** Co-Kriging Ergebnisse der Einzelgebiete

Übersicht elektronischer Anhang:

- *01_Daten*
 - *EPISODES Data:* Meteorologische Modelldaten für die Referenzmessstellen, Grundlage für die Evaluierung der operativen Vorhersagequalität
 - *HYRAS Data:* Meteorologische Beobachtungsdaten für die Referenzmessstellen, Grundlage für KNN-Training und Evaluierung der Vorhersagequalität
 - *GW-Rohdaten:* Enthält im Format vereinheitlichte Originaldaten aller Grundwassermessstellen, sowie die Grundwasserdaten nach dem Preprocessing
 - *GW-Data_Referenzmessstellen:* für die Vorhersage nötigen Grundwasserdaten der Referenzmessstellen (interpoliert und gekürzt)
 - *GW-Stammdaten:* enthält Stammdaten aller Grundwassermessstellen
 - *pretrained_networks:* enthält Matlab Dateien der trainierten NARX zur schnelleren Reproduktion der Ergebnisse

- **02_Skripte** notwendige (Matlab-) Skripte zur Reproduktion der Ergebnisse, sowie zugehörige Matlab-Dateien und Funktionen Dritter (siehe Lizenzfiles in den jeweiligen Ordnern)
- **03_Ergebnisse**

- *01_Clusterergebnisse*

Sammlung folgender Ergebnisse für jedes der 11 untersuchten Gebiete einzeln:

- * Ordner *Clusterergebnisse vollständig*: alle Ergebnisse vor dem Kappen und Aussortieren der Cluster
 - Ordner *Cluster Messstellenlisten*: *.txt-Dateien mit Messstellenlisten je Cluster inkl. R_w -Werten
 - Ordner *Feature Boxplots*: Werteverteilungen der Features innerhalb der einzelnen Cluster dargestellt als Boxplots - geliefert als .png und Matlab-Figure (*.fig)
 - Ordner *Shapefiles*: Shapefiles für jedes Cluster
 - Ordner *Summary_Graphics*: grafische Darstellung der Ergebnisse als Karten, gestapelte, z-normierte Ganglinien und gestapelte Ganglinien mit Originalamplitude (mean-shift)
 - Grafik *Cluster_hits*: Clustergrößen als Balkendiagramm
- * Ordner *Clusterergebnisse bereinigt*: Ergebnisse der nicht-aussortierten, gekappten Cluster
 - Ordner *Cluster Messstellenlisten*: *.txt-Dateien mit Messstellenlisten je Cluster inkl. R_w -Werten
 - Ordner *Shapefiles*: Shapefiles für jedes bereinigte Cluster
 - Ordner *Summary_Graphics*: grafische Darstellung der bereinigten Ergebnisse als Karten, gestapelte, z-normierte Ganglinien und gestapelte Ganglinien mit Originalamplitude (mean-shift)

Zusätzlich liefert der Ordner *Referenzmessstellen* folgende Inhalte:

- * Shapefile mit allen RM-Messstellen
- * *RM-Bearbeiter-Auswahl.csv*: Aufschlüsselung welche der Messstellen im Cluster als RM ausgewählt wurde
- * *RM-Liste.csv*: Liste aller RM mit Projekt-ID, Name, Gebiet und Clusternummer

Im Ordner *Sonstiges* sind folgende Inhalte zusammengefasst:

- * Shapefiles aller Messstellen im Originaldatensatz, der Messstellen nach dem Preprocessing, der zum Clustering genutzten Messstellen und aller am Ende (nach Bereinigung der Ergebnisse) verbleibenden Messstellen
- * *Rw-Cutoff-Werte(alleCluster).csv*: für alle Gebiete Angabe des R_w -cutoff Wertes für behaltene Cluster und Angabe ob Cluster aussortiert wurden

- *02_Grundwasserstandsvorhersage*
 - * *01_HYRAS-basiert*: Evaluierung der Vorhersagegüte aller HYRAS-basierten Vorhersagen sowie monatsweise Auswertung der PIop-Werte
 - * *02_EPISODES-basiert*: Ergebnisse der operativen Jahreszeitemvorhersagen sowie Konfidenzintervall-Auswertungen
- *03_Regionalisierung*

Rasterdatensätze alle berechneten Co-Kriging Oberflächen sowie die deutschlandweiten Flurabstände in Form einer Geodatabase ("CoKriging_Raster"). Weiterhin ist ein Shapefile (meanGWH_2007-14) mit den langjährigen mittleren Grundwasserhöhen der genutzten Messstellen beigefügt.

Anhang A - Algorithmenkonfigurationen

Referenzmessstellenauswahl

Gebiet	Mst.	Feature Vorauswahl	DS2L Parameter	beste Feature-Konfiguration	Validierungsindizes	Cluster	Ensemblegröße / Wdh.
GR1	5291	1 RangeRatio	SOM_size big	2 Schiefe	CH	63.096	237 FeatureConfigs 4088
		2 Schiefe	DR Yes	8 SeasonalBehaviour	MR	0.483	Jackknifing 10000
		3 Periodizität	DM No	9 Yearly Variance	PBM	0.020	Voting
		4 F _{diff}	NTH 1	3 Periodizität	RL	0.056	
		5 jumps	HMTH 0	1 RangeRatio	C	0.066	
		6 SEM		7 ExVals			
		7 ExVals		6 SEM			
		8 SeasonalBehaviour					
		9 Yearly Variance					
GR2	888	1 Schiefe	SOM_size big	5 SeasonalBehaviour	CH	115.556	17 FeatureConfigs 330
		2 Periodizität	DR Yes	2 Periodizität	MR	0.513	Jackknifing 5000
		3 jumps	DM No	4 SEM	PBM	1.915	Voting
		4 SEM	NTH -1	6 Yearly Variance	RL	0.200	
		5 SeasonalBehaviour	HMTH 0	1 Schiefe	C	0.107	
		6 Yearly Variance					
R31,32	888	1 Schiefe	SOM_size big	5 SeasonalBehaviour	CH	115.556	17 FeatureConfigs 330
		2 Periodizität	DR Yes	2 Periodizität	MR	0.513	Jackknifing 5000
		3 jumps	DM No	4 SEM	PBM	1.915	Voting
		4 SEM	NTH -1	6 Yearly Variance	RL	0.200	
		5 SeasonalBehaviour	HMTH 0	1 Schiefe	C	0.107	
		6 Yearly Variance					
R33	80	1 Schiefe	SOM_size normal	3 ExVals	CH	29.638	4 FeatureConfigs 58
		2 Periodizität	DR Yes	2 Periodizität	MR	0.616	Jackknifing 5000
		3 ExVals	DM No	4 Yearly Variance	PBM	2.509	Voting
		4 Yearly Variance	NTH -1		RL	0.367	
GR4	624		HMTH 0		C	0.168	
		1 Schiefe	SOM_size big	3 F _{diff}	CH	59.491	11 FeatureConfigs 8715
		2 Periodizität	DR Yes	2 Periodizität	MR	0.640	Jackknifing 5000
		3 F _{diff}	DM No	9 diffsum	PBM	0.805	Voting
		4 jumps	NTH -1	10 m _{up} m _{down}	RL	0.212	
		5 SEM	HMTH 0	6 ExVals	C	0.155	
		6 ExVals		4 jumps			
		7 SeasonalBehaviour		1 Schiefe			
		8 Yearly Variance		5 SEM			
		9 diffsum					
10 m _{up} m _{down}							
GR5	1175	1 RangeRatio	SOM_size normal	9 Yearly Variance	CH	53.962	43 FeatureConfigs 4063
		2 Schiefe	DR Yes	6 SEM	MR	0.588	Jackknifing 5000
		3 Periodizität	DM No	2 Schiefe	PBM	0.192	Voting
		4 F _{diff}	NTH 3	3 Periodizität	RL	0.125	
		5 jumps	HMTH 0	4 F _{diff}	C	0.123	
		6 SEM		8 SeasonalBehaviour			
		7 ExVals		7 ExVals			
		8 SeasonalBehaviour					
		9 Yearly Variance					
GR6	567	1 RangeRatio	SOM_size normal	1 RangeRatio	CH	34.305	27 FeatureConfigs 8707
		2 Schiefe	DR Yes	2 Schiefe	MR	0.595	Jackknifing 2000
		3 Periodizität	DM Yes	3 Periodizität	PBM	0.321	Voting
		4 F _{diff}	NTH 2	5 Longest Recession	RL	0.152	
		5 Longest Recession	HMTH 0	4 F _{diff}	C	0.121	
		6 jumps		7 SEM			
		7 SEM		6 jumps			
		8 ExVals					
		9 SeasonalBehaviour					
		10 m _{up} m _{down}					
GR7	16	1 Schiefe	SOM_size normal	1 Schiefe	CH	6.379	7 FeatureConfigs 44
		2 Periodizität	DR Yes	3 SEM	MR	0.463	Jackknifing 2000
		3 SEM	DM No	4 SeasonalBehaviour	PBM	2.189	Voting
		4 SeasonalBehaviour	NTH 0		RL	0.340	
GR8	436		HMTH 0		C	0.128	
		1 Schiefe	SOM_size normal	2 Periodizität	CH	115.906	31 FeatureConfigs 15
		2 Periodizität	DR Yes	3 SEM	MR	0.264	Jackknifing 5000
		3 SEM	DM Yes	1 Schiefe	PBM	0.664	Voting
			NTH 3		RL	0.170	
			HMTH 0		C	0.039	
GR9	405	1 RangeRatio	SOM_size big	3 Periodizität	CH	67.855	26 FeatureConfigs 171
		2 Schiefe	DR Yes	2 Schiefe	MR	0.437	Jackknifing 5000
		3 Periodizität	DM No	4 ExVals	PBM	0.449	Voting
		4 ExVals	NTH -1	5 Yearly Variance	RL	0.177	
		5 Yearly Variance	HMTH 0		C	0.082	
GR10	47	1 RangeRatio	SOM_size normal	1 RangeRatio	CH	17.412	5 FeatureConfigs 491
		2 Schiefe	DR Yes	2 Schiefe	MR	0.545	Jackknifing 1000
		3 Periodizität	DM Yes	5 SEM	PBM	2.705	Voting
		4 jumps	NTH 0	4 jumps	RL	0.353	
		5 SEM	HMTH 0	6 diffsum	C	0.122	
		6 diffsum					

Anhang B - Berechnung der Cluster-Validierungs-Indizes

Die folgende Zusammenstellung basiert vollständig auf Desgraupes (2018).

Calinski-Harabasz Index

$$CH = \frac{N - K}{K - 1} \cdot \frac{BGSS}{WGSS}$$

N : Observations

K : Groups / Clusters

$BGSS$: Between-group dispersion

$WGSS$: Within-cluster dispersion

McClain-Rao-Index

$$MR = \frac{N_B}{N_W} \cdot \frac{S_W}{S_B}$$

N_B : Total number of distances between pairs of points which do not belong to the same cluster

N_W : Total number of distances between pairs of points of the same cluster

S_W : Sum of within-cluster distances

S_B : Sum of Between-Cluster-Distances

PBM-Index

$$PBM = \left(\frac{1}{K} \cdot \frac{E_T}{E_W} \cdot D_B \right)^2$$

K : Groups / Clusters

E_T : Sum of the distances of all the points to the barycenter G of the entire data set

E_W : Sum of the distances of the points of each cluster to their barycenter

D_B : Largest distance between two cluster barycenters

Ratkowsky-Lance-Index

$$RL = \frac{\bar{c}}{\sqrt{K}}$$

with:

$$\bar{c}^2 = \frac{1}{p} \sum_{j=1}^p \frac{BGSS_j}{TSS_j}$$

K : Groups / Clusters

p : Dimension

$BGSS$: Between-group dispersion

TSS : Total Sum of Squares (Total Scatter)

C-Index

$$C = \frac{S_W - S_{min}}{S_{max} - S_{min}}$$

S_W : Sum of the N_W distances between all the pairs of points inside each cluster

S_{min} : Sum of the N_W smallest distances between all the pairs of points in the entire data set.

S_{max} : Sum of the N_W largest distances between all the pairs of points in the entire data set.

N_W : Total number of distances between pairs of points of the same cluster

Anhang C - Liste aller Referenzmessstellen

Gebiet	Cluster	Proj_ID	Name
GR1	1	BB_28410218	Zootzen, 2,7 km oestl.
GR1	2	BB_27381011	Lockstaedt UP
GR1	4	NI_40505301	Dalumer Moor I
GR1	6	ST_34330013	Jeggau
GR1	7	BB_35406375	Moethlitz, Neu Plauer Forst,MP2
GR1	8	BB_36465001	Klein-Beeren,am Grenzgraben, MP
GR1	9	NI_200000876	St. Hülfe I
GR1	10	SH_10L58103001	6107_Meezen
GR1	11	NI_9700272	Voltlage I
GR1	12	NI_40502382	Nördlich Erika II
GR1	15	BB_36505189	Fuerstenwalde
GR1	16	ST_32360043	Osterburg
GR1	17	BB_37451902	Kleinbeuthen, Rehwinkel, UP
GR1	18	BB_37451908	Juetchendorf,Str.Groeben-Gr.Be.
GR1	19	SH_10L58137003	6317_Soenderby_F2
GR1	20	SH_10L60002001	3369_Beckersdorf_F1
GR1	21	NI_400060071	Elm UE 7 FI
GR1	23	BB_35432590	Marquardt, an der AB Bln.-Ring
GR1	24	NI_500000518	UWO 047_1 Abbendorf
GR1	25	ST_34330014	Peckfitz
GR1	26	NI_200002153	Donstorf
GR1	27	BB_34426014	Klein Behnitz,ca.700m suedl.,UP
GR1	28	BB_33392110	Steckelsdorf,ca.300m oestl. See
GR1	29	BB_34396780	Premnitz, Buetzer Feld
GR1	31	NI_200000620	Walsen
GR1	33	BB_36471636	Bohnsdorf 77
GR1	34	SH_10L62020008	4775_Granderheide_III
GR1	35	BB_38441782	Ahrensdorf,Str.n.Hennicken.MP1
GR1	36	ST_38390006	Reppinichen
GR1	37	NI_400061210	Rönndeich UE 121
GR1	38	NI_500003038	LGD 014 N 13 Sothel
GR1	39	NI_500000256	NA 065_2 Marwede
GR1	40	BB_34481534	Bruchmühle
GR1	41	ST_36395811	Karow
GR1	42	NI_9700159	Langwege
GR1	43	BE_496	402A-3765
GR1	44	ST_41380015	Steckbyer Forst
GR1	45	SH_10L60025003	4895_GLASAU_F2
GR1	49	ST_42410008	Bergwitz
GR1	51	BB_32529245	Tr.Gr.Neuendorf Nord V OP

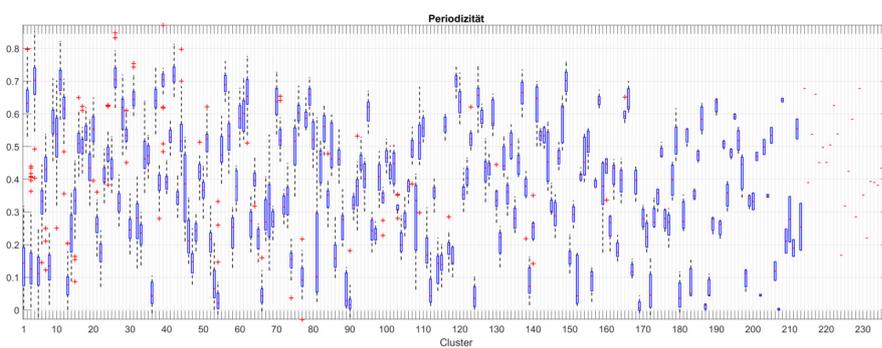
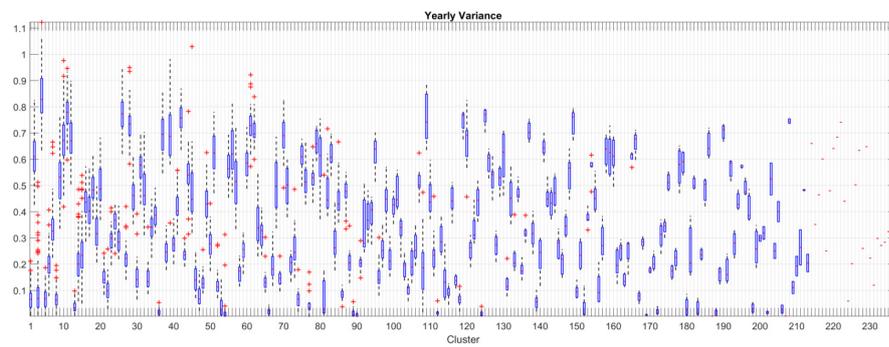
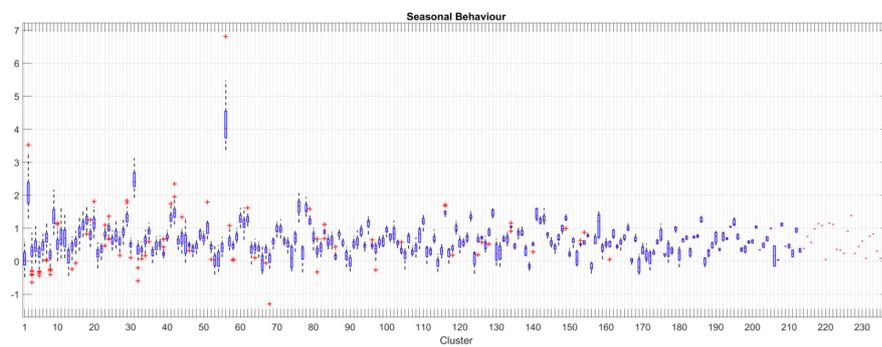
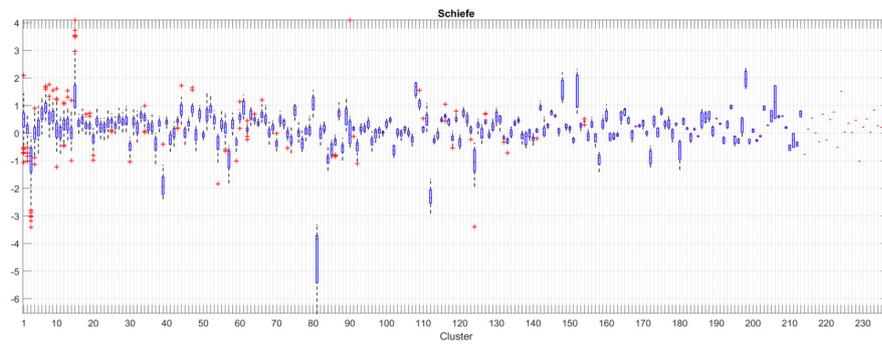
Gebiet	Cluster	Proj_ID	Name
GR1	56	BB_32455305	Hohenbruch, Weg n.Teerofen
GR1	59	SH_10L61042008	3923_Hohenfiert
GR1	60	BB_35416179	Brielow, OT Seehof, Seestr. UP
GR1	61	ST_35370071	Schelldorf
GR1	62	BB_29370010	Uenze, noerdl. d. Ortslage
GR1	64	ST_31350030	Gagel Fl
GR1	70	BB_32449203	Kremmen, Klein Asien
GR1	71	ST_42410003	Bergwitz
GR1	75	ST_41400587	Riesigk
GR1	76	SH_10L56010003	3593_Bokelsess_F3
GR1	78	SH_10L53089002	4389_Möhnsen_2
GR1	79	NI_500000493	UWO 029 Hastedt
GR1	80	BE_377	425B-4585
GR1	87	ST_41419101	Seegrehna
GR1	91	NI_405161720	Dorfhagen UWO 172
GR1	95	SH_10L59087003	5457_TETENHUSEN_F2
GR1	96	SN_47410412	Naunhof - Lindhardt Straße, B 20412_66
GR1	100	BB_34512420	Neuhardenberg, Richtung Wulkow
GR1	109	SN_46421125	Grubnitz
GR1	110	SH_10L58064003	6138_GR._Buchwald_F3
GR1	116	BB_30370039	Karthan, Rohr 1 (gr. Rohr) OP
GR1	119	NI_9700020	Beverbruch 3_6
GR1	123	BB_33465504	Schildow OP
GR1	125	NI_40507282	Echtelerfeld II
GR1	126	NI_400080170	Drochtersen UE 17
GR1	137	NI_600041661	Vögelsen F1
GR2	1	NW_60100205	RWI 66
GR2	2	NW_70200415	LGD Mondf. RP 8n
GR2	3	NW_40100261	Bislicher Wald 42
GR2	4	NW_80100790	ST Tönis 021
GR2	5	NW_40060500	1 06 050 - HS 104
GR2	7	NW_60072301	HS 89
GR2	8	NW_40100182	Hamminkeln 27
GR2	9	NW_91132502	Frielinghausen ML53
GR2	10	NW_60090169	HS 67
GR2	11	NW_70284313	LGD St.Aug.Ankerstr
GR2	12	NW_91163705	Pöppelsche Eikeloh
GR2	14	NW_91163900	Eikeloh Betonplatte
GR2	15	NW_91154108	Berge de Süd

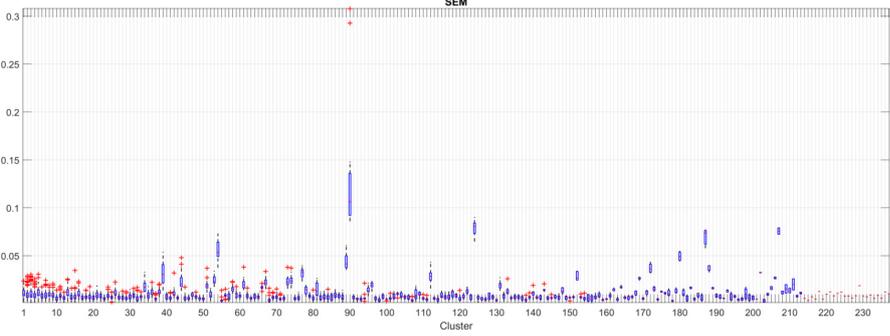
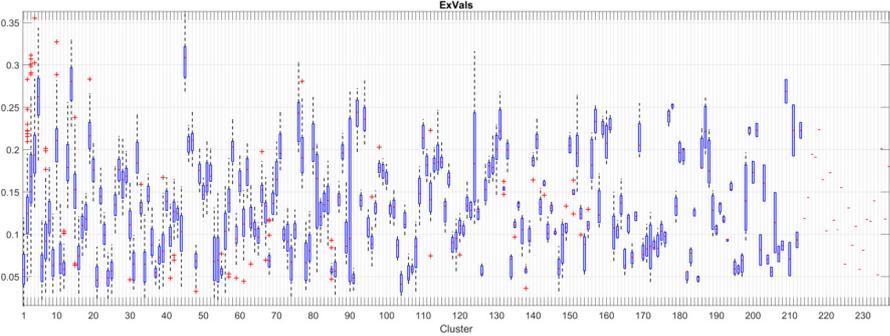
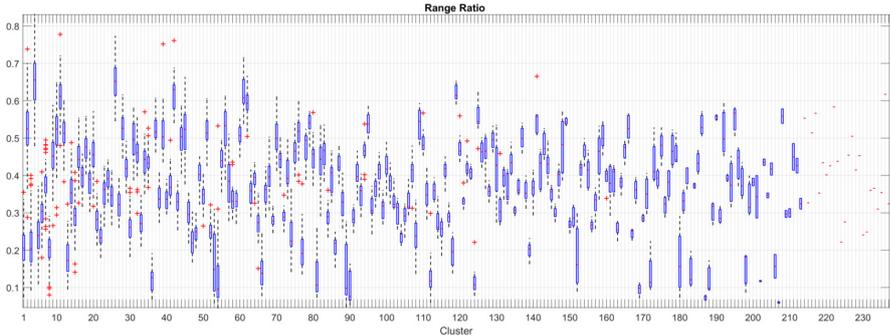
Gebiet	Cluster	Proj_ID	Name
R31,32	1	BW_104-114-5	1807 Renchen 1
R31,32	2	BW_103-261-4	Malsch
R31,32	3	BW_259-259-7	GWM 19-1 T BWV, Blankenloch
R31,32	4	BW_181-304-4	Heddesheim
R31,32	5	BW_765-306-3	Ketsch Schleiweise T
R31,32	6	HE_13711	Heppenheim
R31,32	7	RP_2375178300	1232 III Hördt
R31,32	9	BW_152-308-7	GWM 1 Flach, Forst
R31,32	10	BW_236-115-9	Hohnhurst
R31,32	12	HE_13500	Einhausen
R31,32	13	HE_12447	Stockstadt
R31,32	14	RP_2391187900	1399 Bobenheim-Roxheim
R31,32	16	BW_126-020-2	3534 Hartheim 8
R31,32	17	BW_157-260-8	1555 Bruchhausen
R31,32	19	BW_140-020-6	GWM 3543 Bremgarten 9, Hartheim am Rhein
R31,32	20	HE_12117	Dieburg
R31,32	21	HE_10702	Rüsselsheim
R33	1	HE_7186	Hopfgarten
R33	4	HE_8999	Hutten
GR4	1	BW_120-670-3	GWM Oberessendorf
GR4	3	BW_170-568-9	GWM BO 4-91, Altheim
GR4	4	BW_119-771-0	GWM P 4 Leutkirch
GR4	6	BW_2004-569-8	GWM A7 6-9, Herbertingen
GR4	7	BW_137-770-0	GWM XX Aitrach
GR4	8	BW_165-568-0	GWM B4 Donauwiesen, Daugendorf
GR4	9	BW_106-767-3	GWM Vöhringen
GR4	10	BW_152-772-1	GWM 8-79 Steinthal Urlau
GR5	1	TH_4634900423	GWM Schönewerda (2_2005)
GR5	2	TH_5226240502	Hy Stadtlengsfeld 17_1978 (0581)
GR5	3	TH_4531230792	Uthleben (1_1999)
GR5	6	TH_5138000179	Hy Gera 6_1971 (Frankenthal)
GR5	10	ST_39310015	Eilsdorf
GR5	12	ST_45371005	Röpzig
GR5	14	ST_49390001	Alttröglitz
GR5	15	TH_4826900003	Hy Großburschla 1_2004
GR5	17	TH_4528230791	Breitenworbis
GR5	21	NI_40000555	Mansbergischer Wald
GR5	22	TH_5138000186	Hy Gera 156_1982 (Liebschwitz, z3D)
GR5	24	TH_4626230766	Birkenfelde (1_2002)
GR5	29	ST_45326568	Bennungen
GR5	30	NW_21661108	A 1 Altmiks
GR5	31	ST_49390053	Göbitz
GR6	1	BY_24151	Unterswildenau Q2
GR6	3	BY_5166	Baunach 18
GR6	5	BY_5158	Neuses
GR6	6	BW_102-714-6	GWM, Weidenstetten
GR6	8	BY_10134	Dietersdorf (KMS1)
GR6	10	BW_126-762-3	Oberkochen Bahnwärterhaus Aalen
GR6	12	BW_16-706-8	GWM B2 Stäffelen, Niederstetten-Neuweiler
GR6	15	BW_100-813-7	Giengen Taubental
GR6	17	BW_119-813-3	Oggenhausen Keller
GR6	21	BY_5171	Rothmannsthal 11

Gebiet	Cluster	Proj_ID	Name
GR7	1	BY_25135	Mittenwald-Gerber 801
GR7	2	BY_25668	Oberau 2_3 TR
GR7	4	BY_25111	Weindorf 573
GR7	5	BY_13103	Häuser 3A_TR
GR8	1	RP_2716140900	6039 Neuwied, Friedhof Sohler-Weg
GR8	2	ST_44380061	Dammendorf
GR8	3	HE_8157	Offheim
GR8	4	RP_2718135900	4117 I Üxheim
GR8	6	HE_7373	Herbornseelbach
GR8	8	NW_91174909	Brilon Lederke OL748
GR8	10	HE_6666	Elnhausen
GR8	11	NW_91175604	Umspannwerke OL769
GR8	15	ST_44380072	Niemberg
GR8	16	NW_129660334	Fretter B8A
GR8	17	HE_6615	Dilschhausen
GR8	25	NW_91167504	Kallenhardt DEKRA
GR9	1	SN_48511110	Putzkau
GR9	2	SN_48518085	Kleinpraga,Hy Saf 25a_70 oben
GR9	3	SN_48480903	Berbisdorf
GR9	4	SN_49430964	Arras
GR9	5	SN_49531740	Schönbach
GR9	6	SN_52411556	Schlunzig
GR9	7	SN_50430645	Röllingshain
GR9	8	SN_49486001	Dresden - Niedersedlitz, B 1981
GR9	9	SN_49420959	Weissbach
GR9	10	SN_47420010	Trebsen, B 10_76
GR9	11	SN_56393708	Adorf, B 12_5_62
GR9	12	SN_48421093	Colditz
GR9	16	SN_5143S0010	Chemnitz, P2_94
GR9	17	SN_49451041	Siebenlehn
GR9	23	SN_47500596	Bischheim
GR10	1	HE_10223	Altenmittlau
GR10	2	BW_131-115-0	GWM 3709, Ohlsbach
GR10	3	HE_12142	Langen
GR10	4	BW_2004-166-1	Biberach

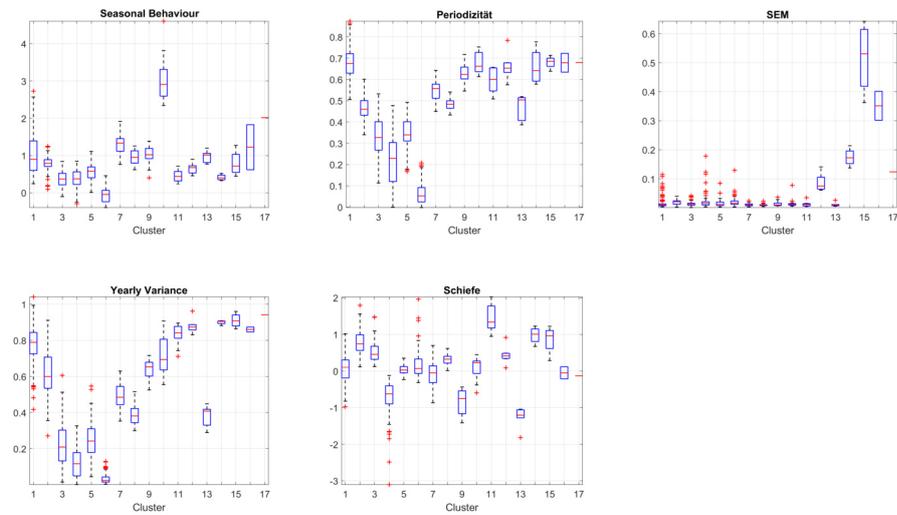
Anhang D - Feature-Boxplots

GR1

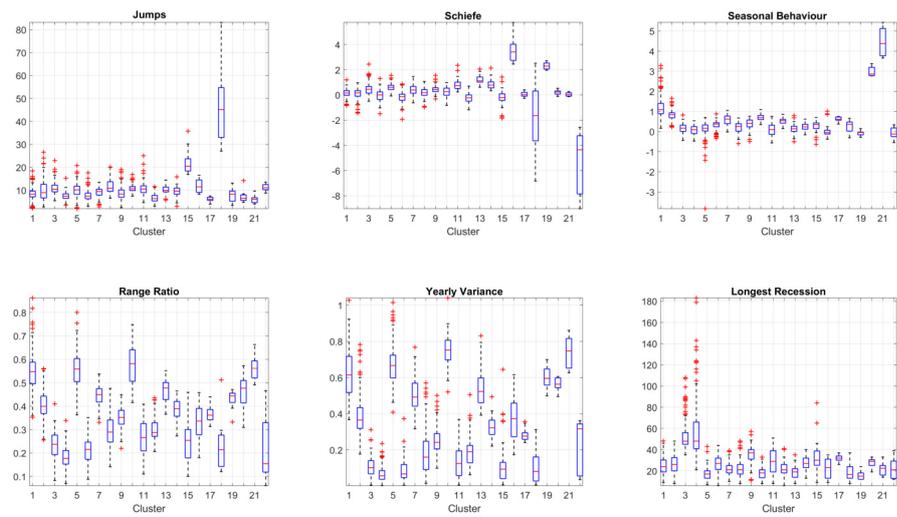




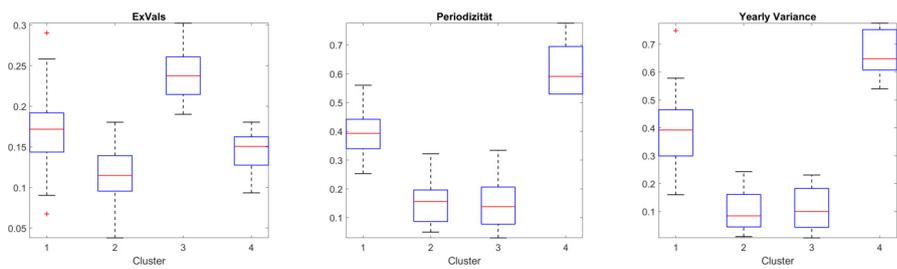
GR2



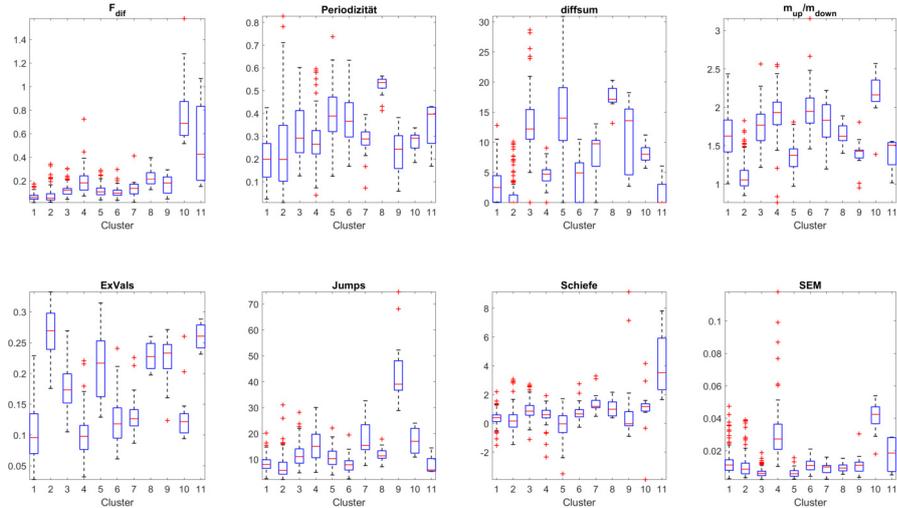
R31,32



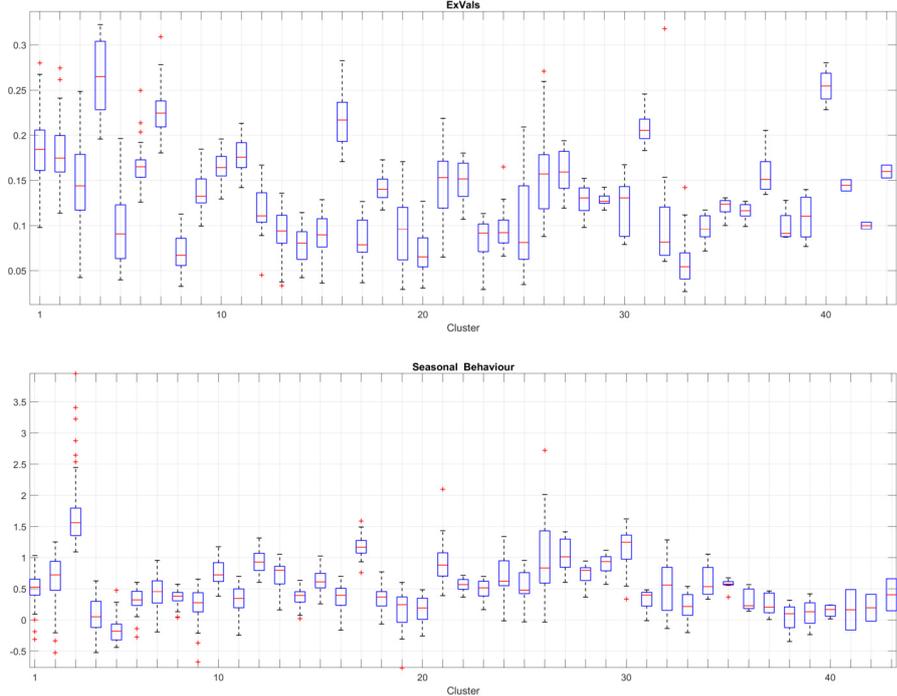
R33

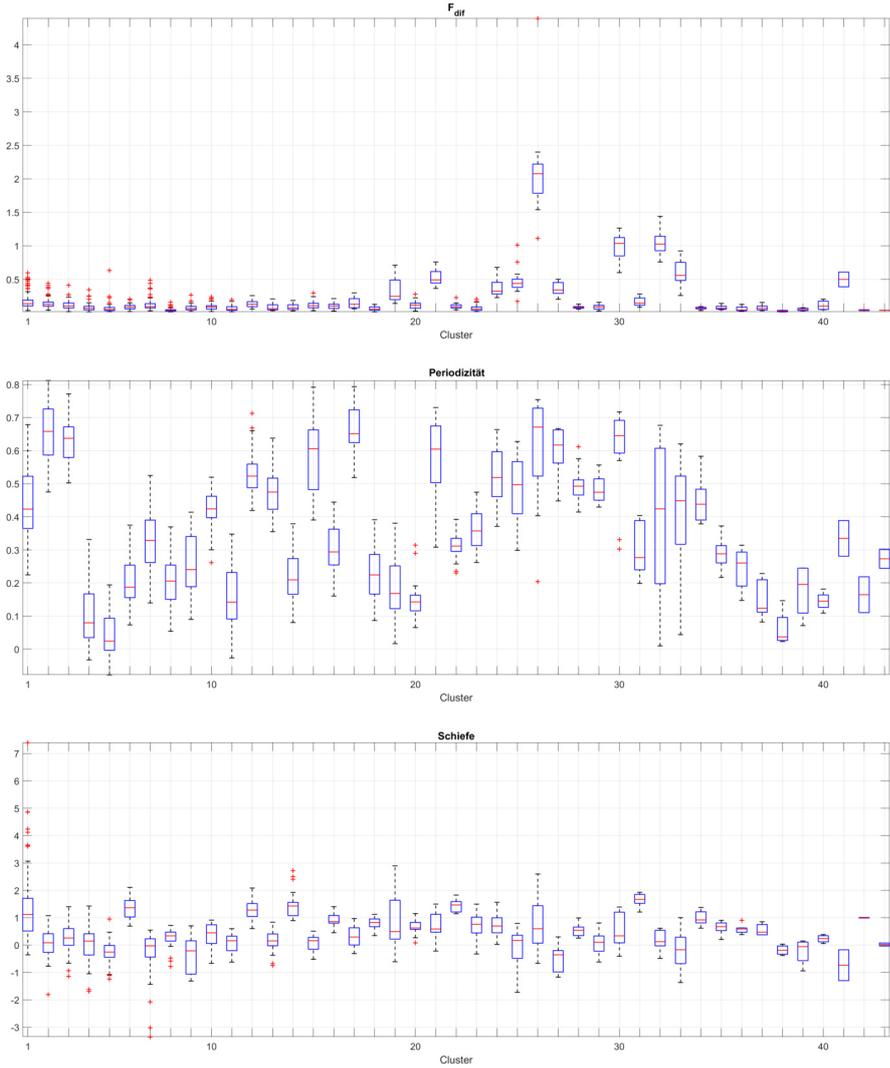


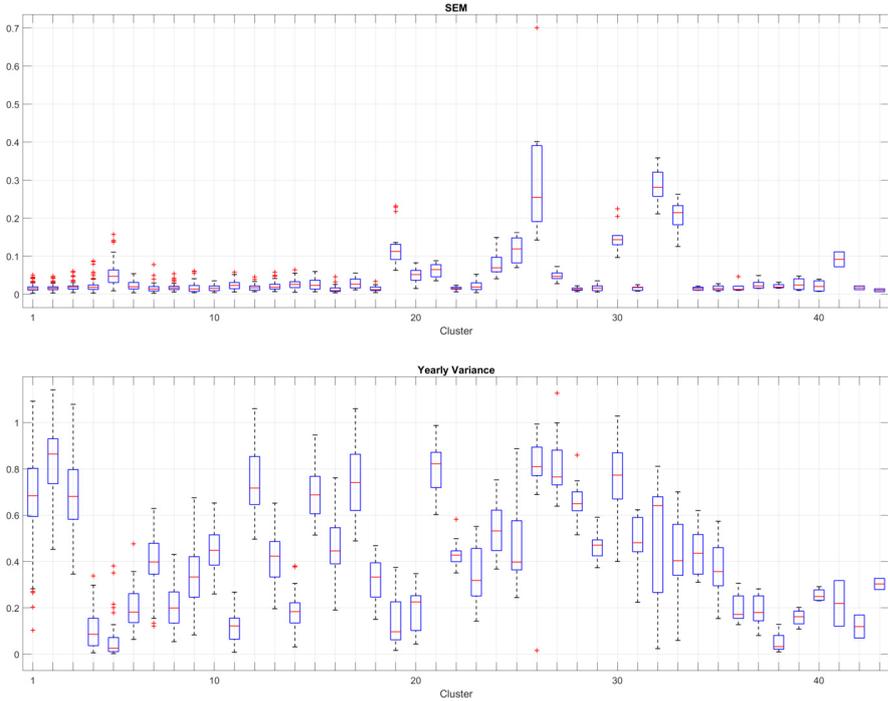
GR4



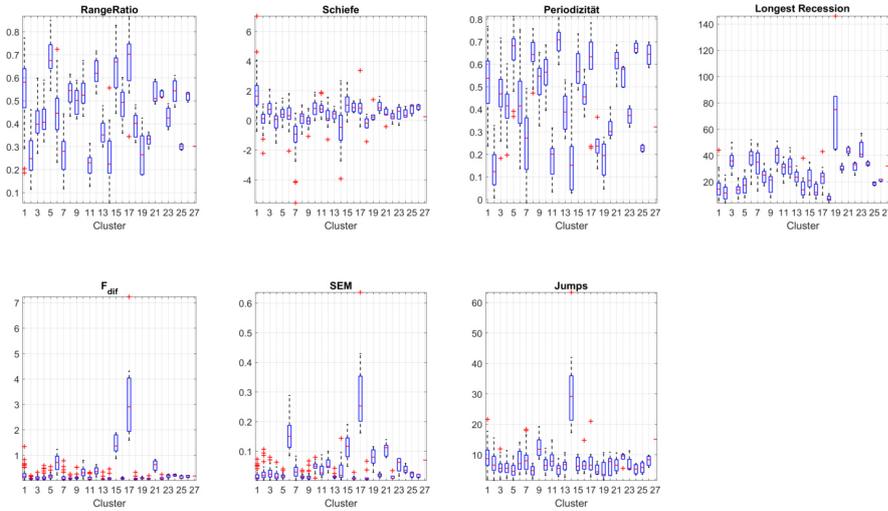
GR5



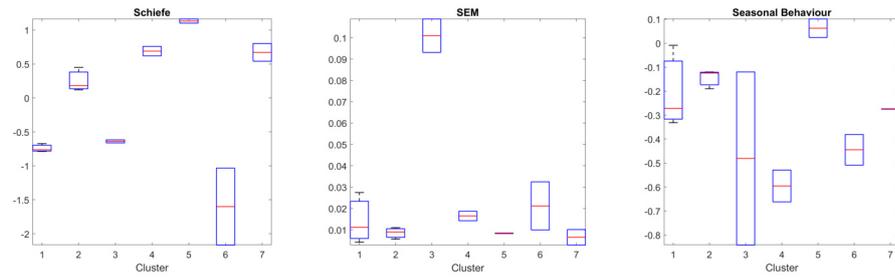




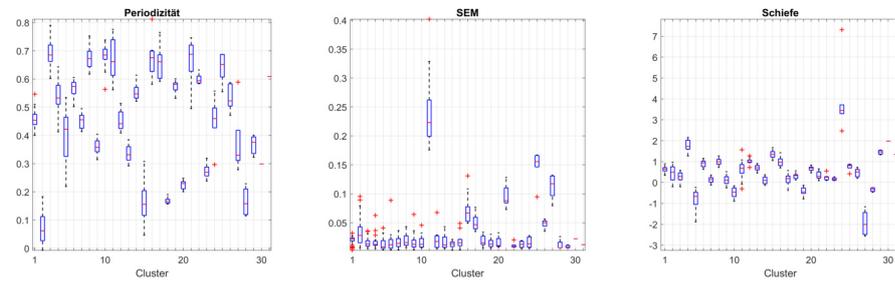
GR6



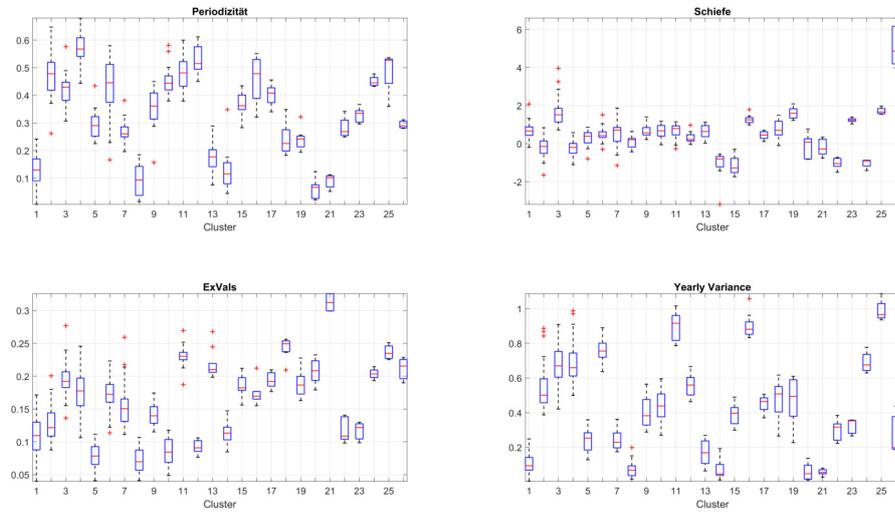
GR7



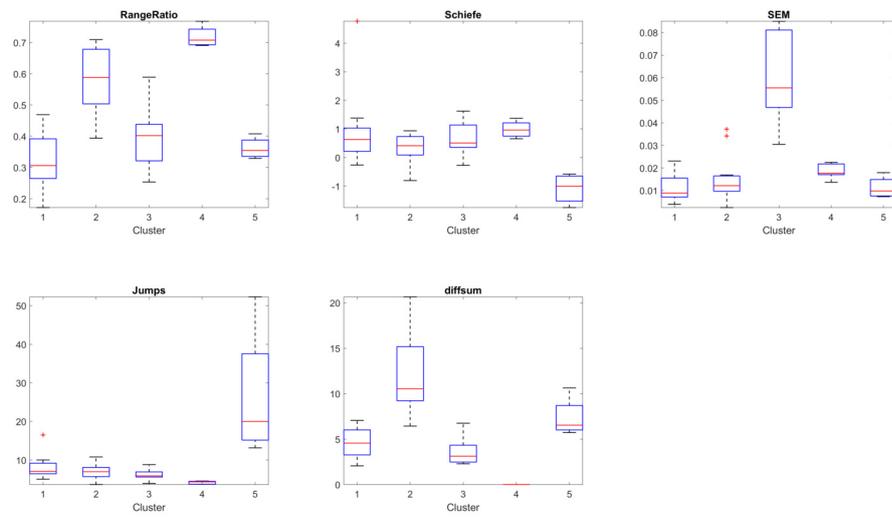
GR8

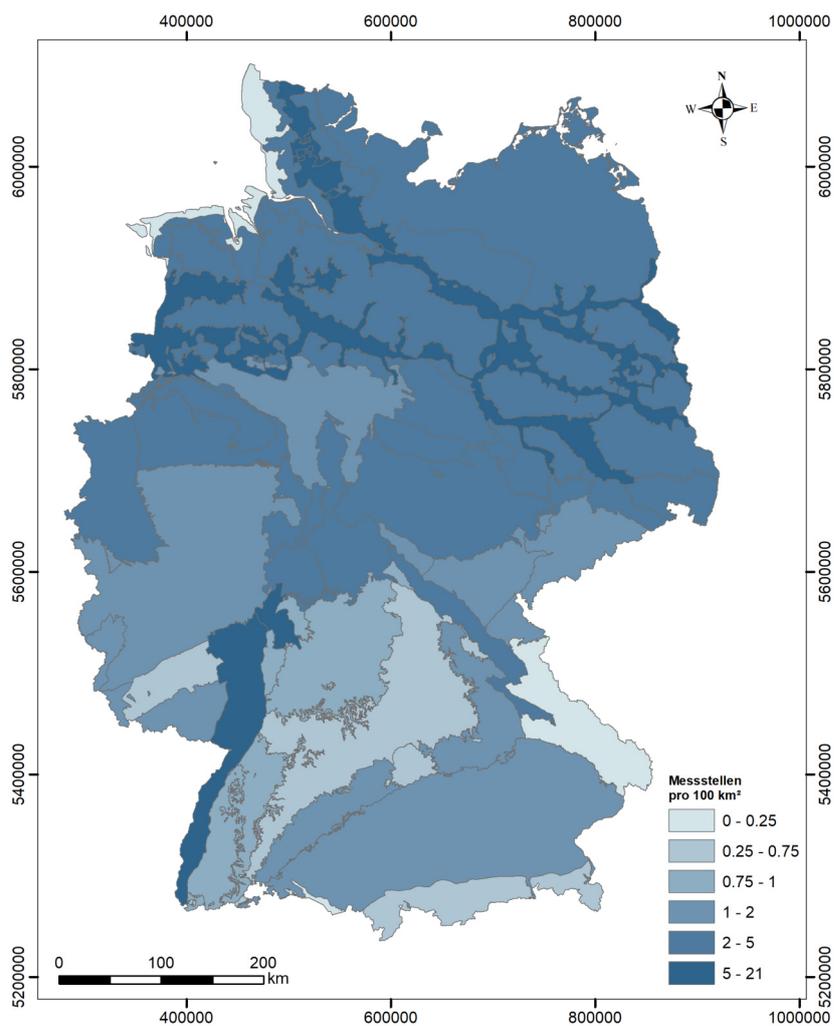


GR9

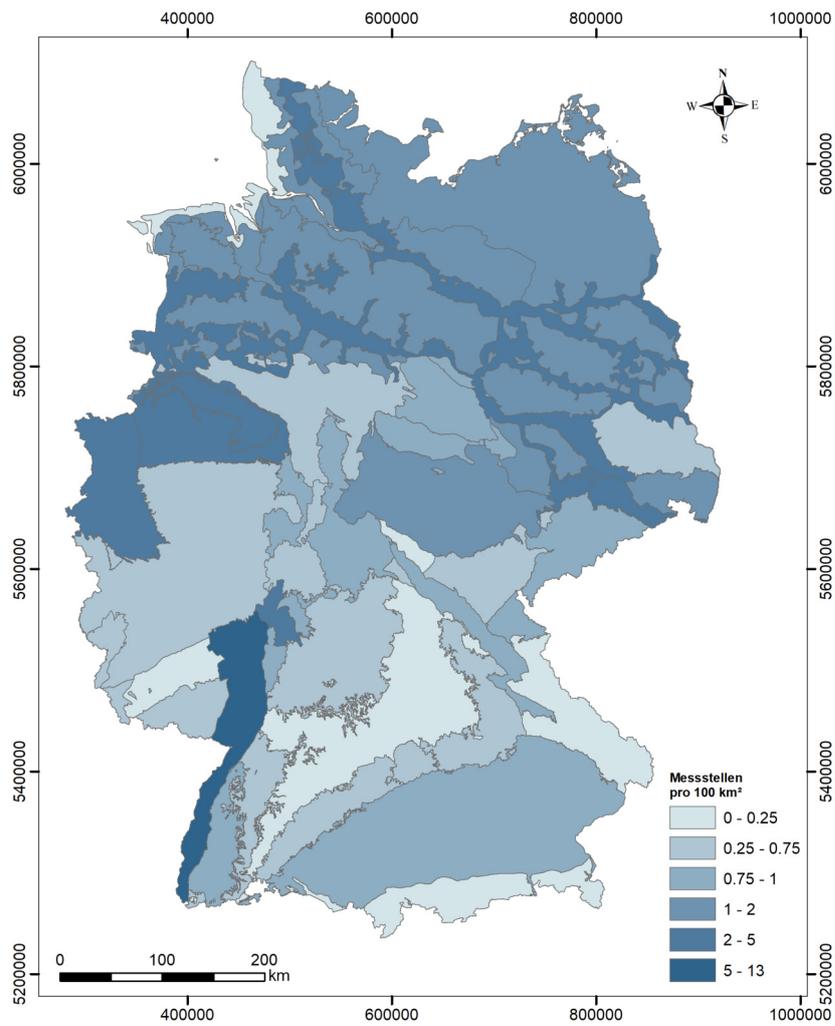


GR10

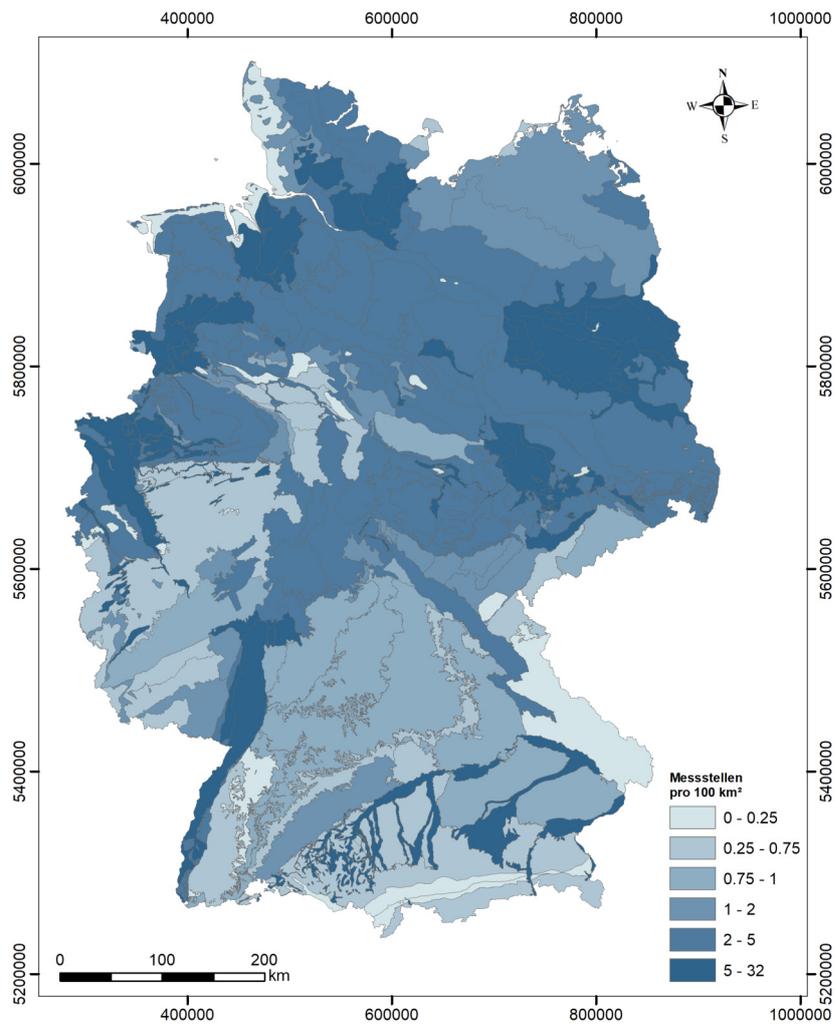


Anhang E - Initiale und verbleibende Messstellendichten

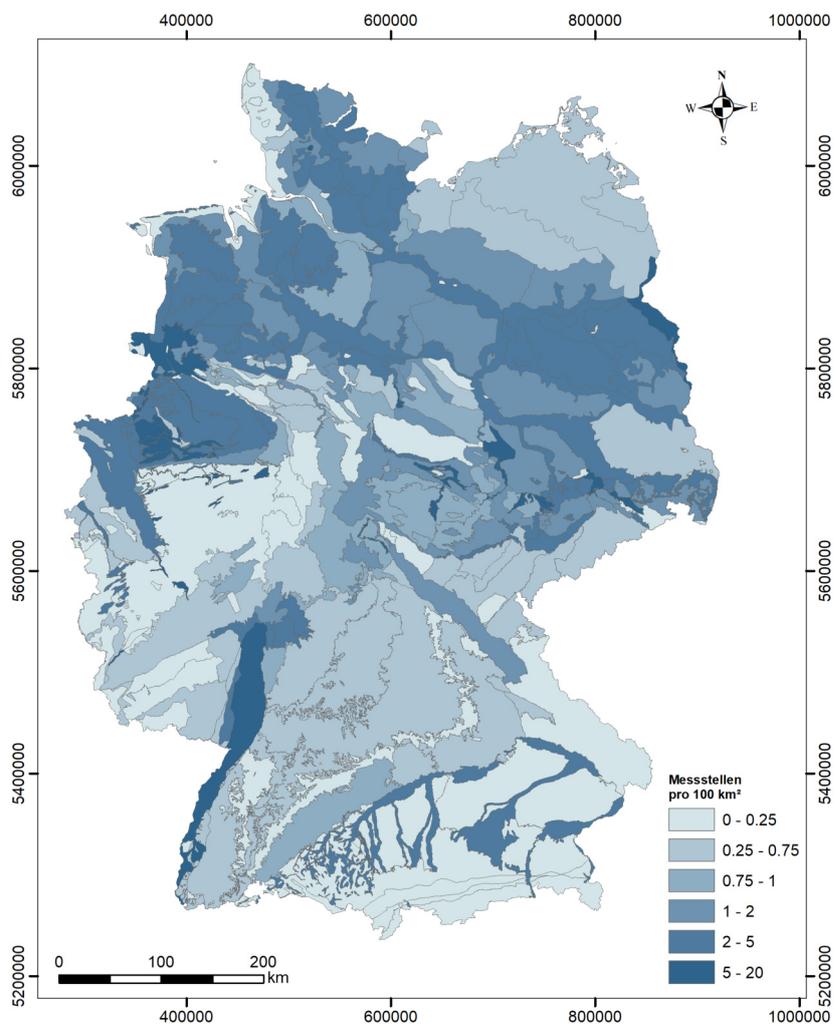
Initiale Messstellendichte der hydrogeologischen Räume in Messstellen pro 100 km^2 .



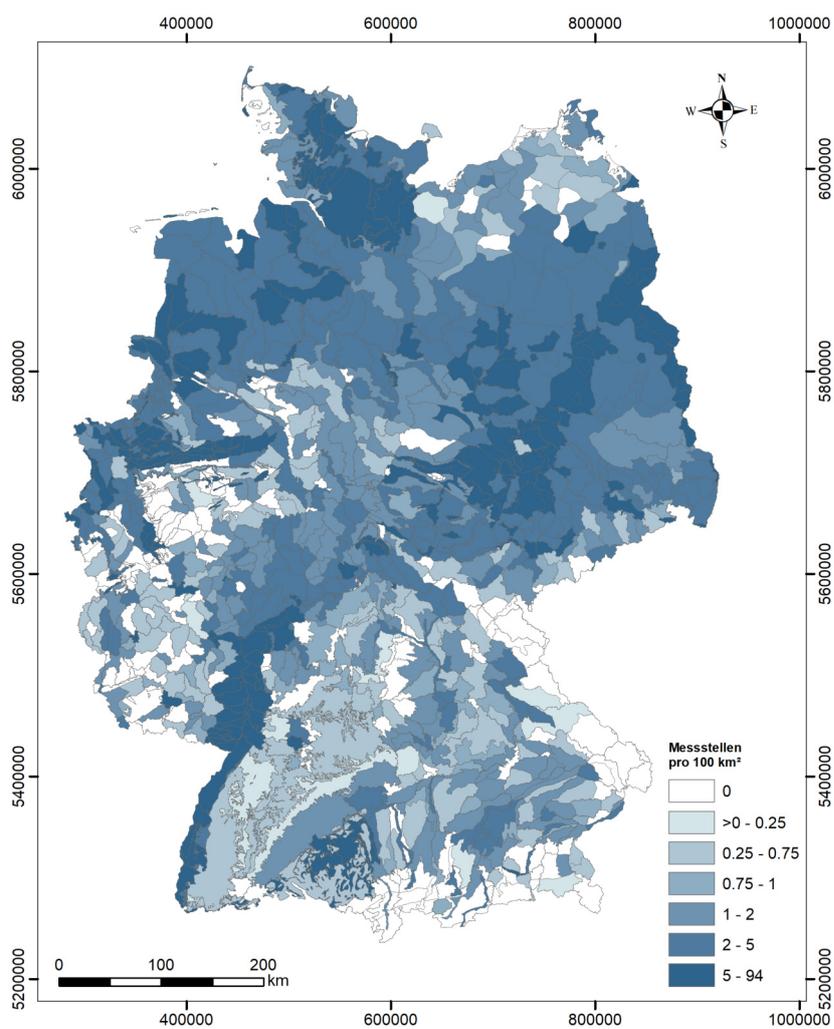
Verbleibende Messstellendichte der hydrogeologischen Räume in Messstellen pro 100 km².



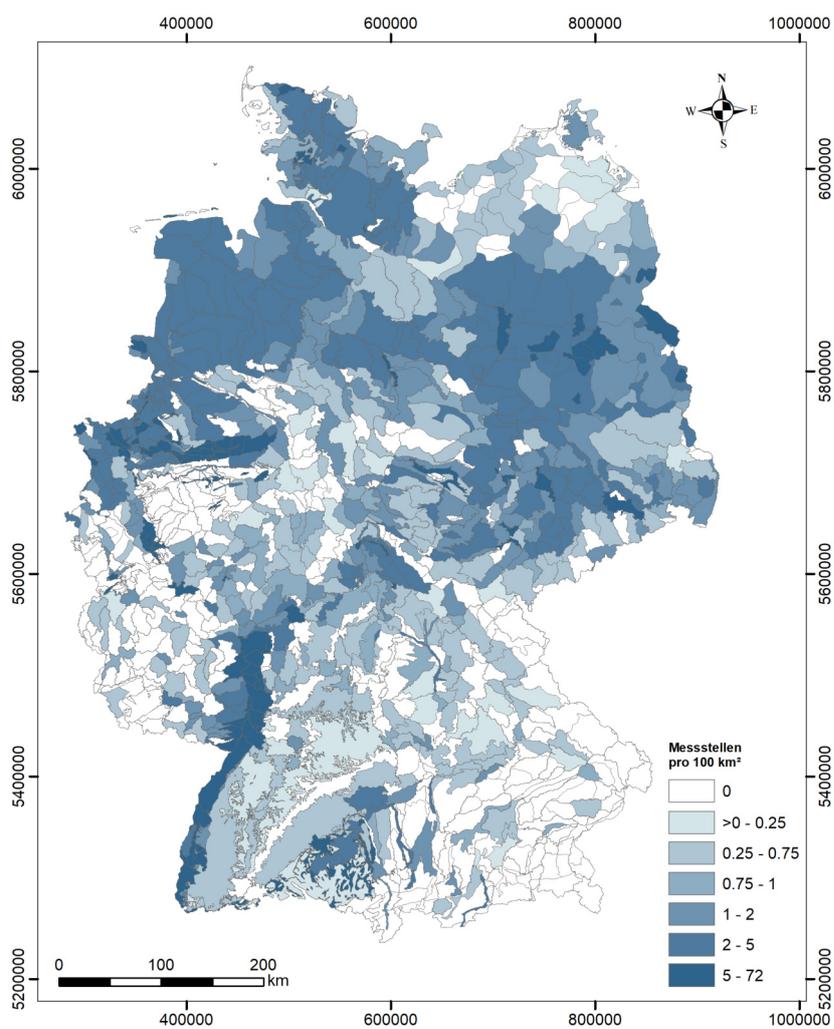
Initiale Messstellendichte der hydrogeologischen Teilräume in Messstellen pro 100 km^2 .



Verbleibende Messstellendichte der hydrogeologischen Teilräume in Messstellen pro 100 km^2 .

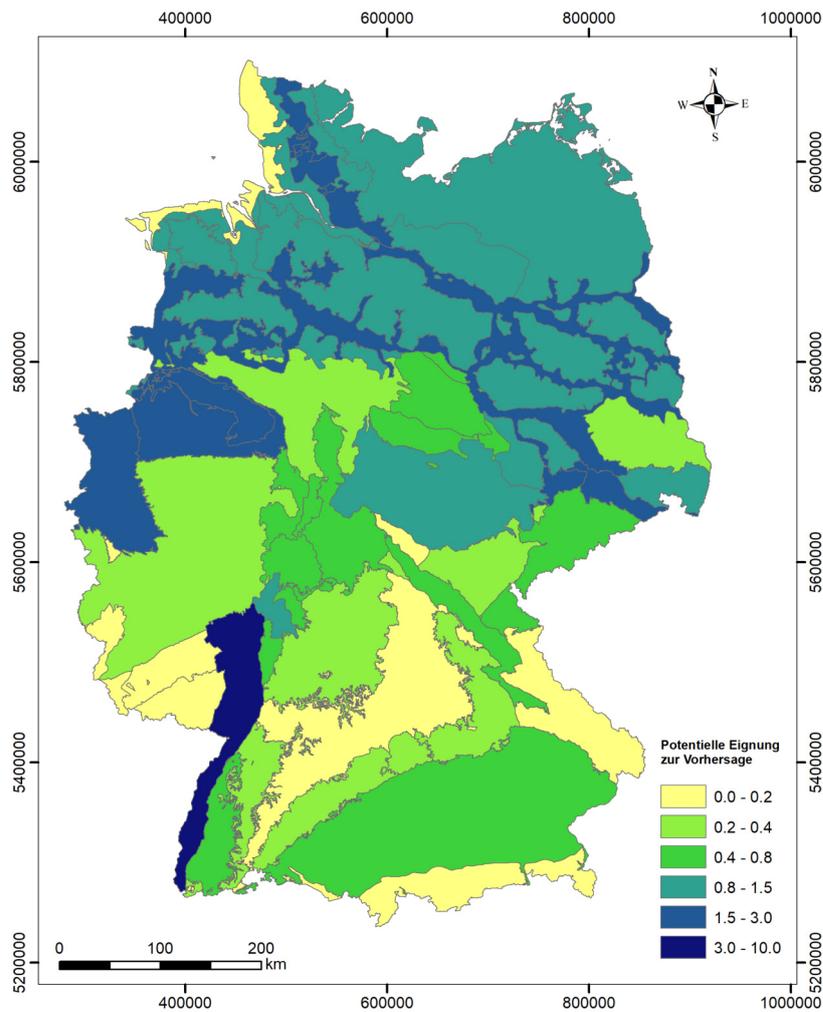


Initiale Messstellendichte der Grundwasserkörper in Messstellen pro 100 km^2 .

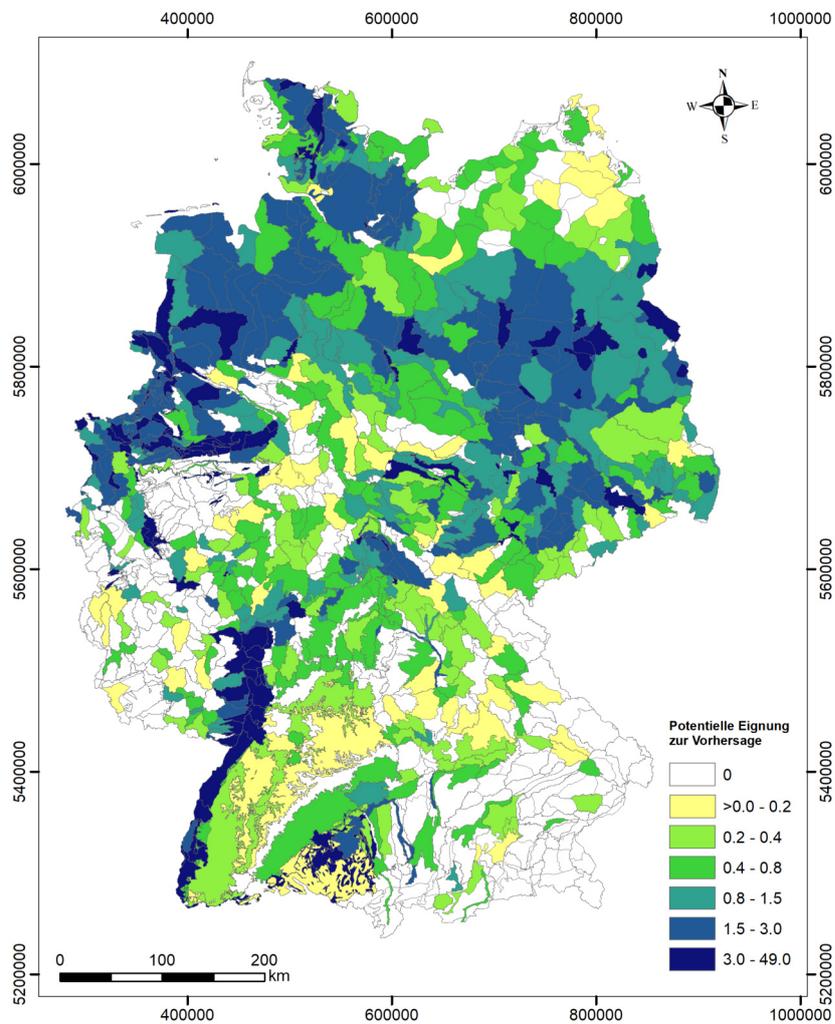


Verbleibende Messstellendichte der Grundwasserkörper in Messstellen pro 100 km².

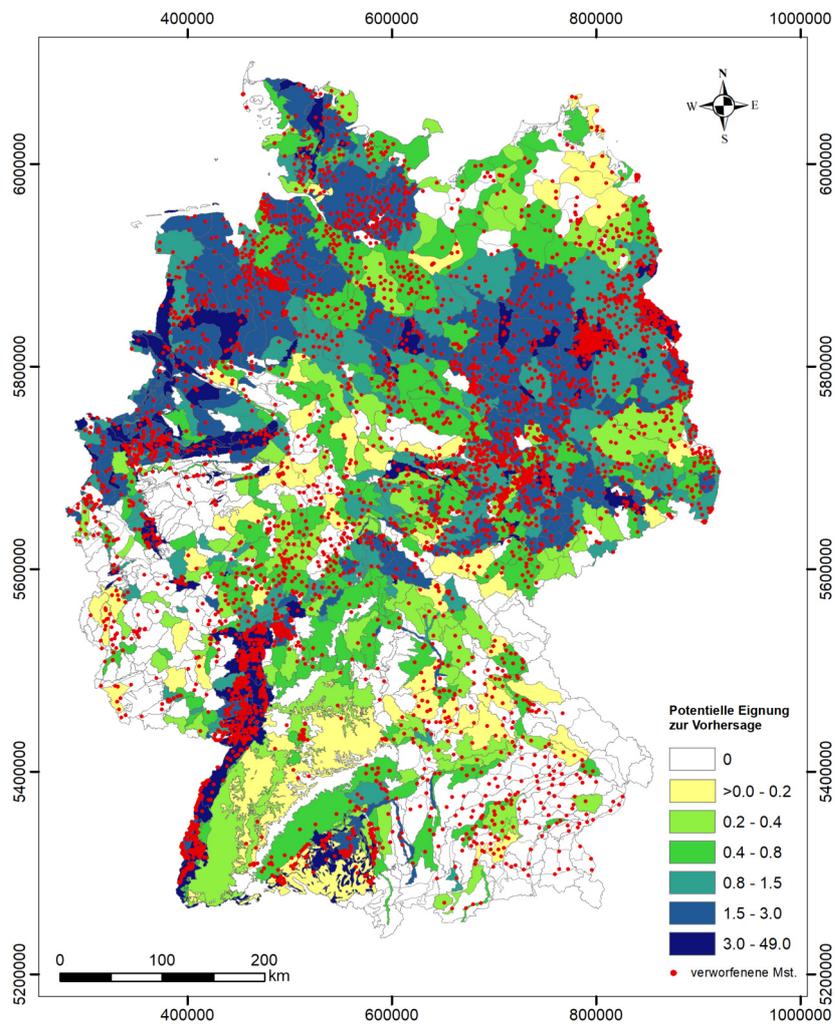
Anhang F - Potentielle Eignung der räumlichen Einheiten zur späteren Vorhersage



Potentielle Eignung der Räume für eine spätere Vorhersage, berechnet aus dem Durchschnitt der gewichteten Korrelationskoeffizienten (\bar{R}_w -Werte) der verbleibenden Messstellen multipliziert mit der verbleibenden Messstellendichte eines Raums. Streng genommen beträgt die Einheit Messstellen pro 100 km^2 , sollte jedoch mehr als relatives Maß zum Vergleich der räumlichen Einheiten untereinander verstanden werden.



Potentielle Eignung der Grundwasserkörper für eine spätere Vorhersage, berechnet aus dem Durchschnitt der gewichteten Korrelationskoeffizienten (\bar{R}_w -Werte) der verbleibenden Messstellen multipliziert mit der verbleibenden Messstellendichte eines Grundwasserkörpers. Streng genommen beträgt die Einheit Messstellen pro 100 km^2 , sollte jedoch mehr als relatives Maß zum Vergleich der räumlichen Einheiten untereinander verstanden werden.



Potentielle Eignung der Grundwasserkörper für eine spätere Vorhersage, berechnet aus dem Durchschnitt der gewichteten Korrelationskoeffizienten (\bar{R}_w -Werte) der verbleibenden Messstellen multipliziert mit der verbleibenden Messstellendichte eines Grundwasserkörpers. Streng genommen beträgt die Einheit Messstellen pro 100 km^2 , sollte jedoch mehr als relatives Maß zum Vergleich der räumlichen Einheiten untereinander verstanden werden. Weiterhin dargestellt sind die im Laufe der Clusterings verworfenen Messstellen.

Anhang G - Ergebnisse der bayesischen Hyperparameteroptimierung

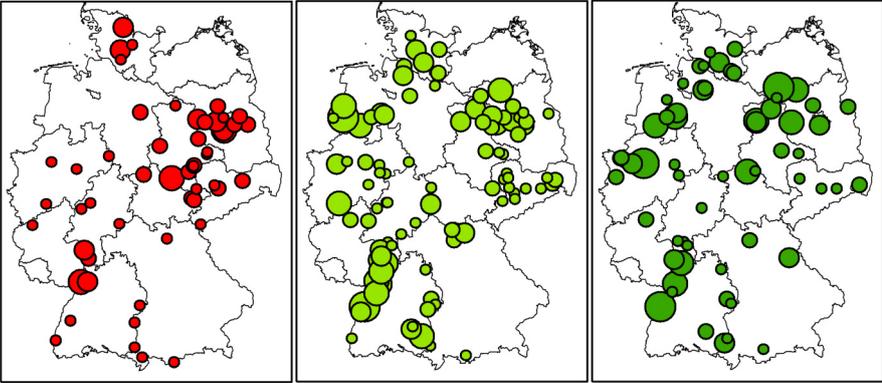
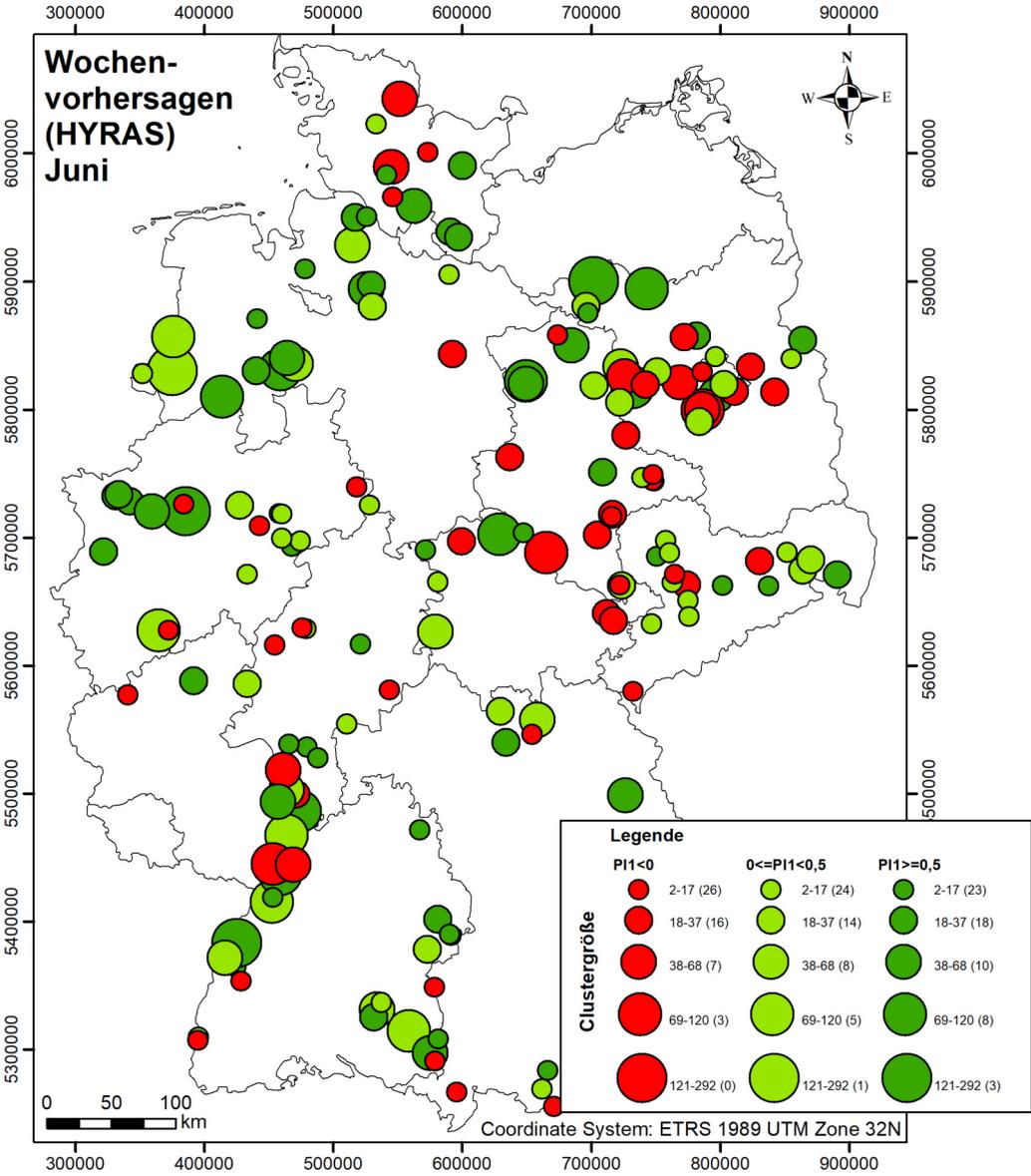
ID	h	Inputs	ID _{max}	FD _{max}	ID	h	Inputs	ID _{max}	FD _{max}	ID	h	Inputs	ID _{max}	FD _{max}	
		rH T Tsin	P rH T Tsin				rH T Tsin	P rH T Tsin				rH T Tsin	P rH T Tsin		
BB_27381011	1	No No Yes	13 51 2		HE_10223	17	Yes No No	48 48 21		SH_10L58137003	17	Yes No No	48 48 21		
BB_28410218	1	No Yes Yes	40 35 51	51	HE_10702	17	Yes No No	48 48 21		SH_10L59087003	6	Yes Yes No	4 29 4	8	
BB_29370010	1	Yes Yes No	42 19 8	2	HE_12117	2	No Yes No	51 10 3		SH_10L60002001	7	Yes Yes No	48 42 46	15	
BB_30370039	6	No Yes Yes	10 8 2	4	HE_12142	1	No Yes No	51 22 16		SH_10L60025003	20	Yes No Yes	19 2 6	27	
BB_32449203	9	No No Yes	5 8 10		HE_12447	19	No No Yes	2 43 2		SH_10L61042008	10	Yes Yes No	16 50 28	24	
BB_32455305	1	Yes Yes No	51 35 25	4	HE_13500	1	Yes Yes Yes	22 28 16	11	1	SH_10L62020008	20	Yes No No	22 20 5	
BB_32529245	2	Yes No Yes	2 51 6	9	HE_13711	2	Yes Yes No	29 42 51	49		SN_46421125	1	Yes No Yes	8 13 45	4
BB_3392110	2	Yes No No	45 16 1		HE_6615	1	Yes Yes Yes	48 6 11 12	1		SN_47410412	1	No Yes Yes	21 11 51	9
BB_33465504	1	Yes Yes Yes	2 52 48 17	9	HE_6666	2	No Yes No	11 4 1		SN_47420010	1	Yes No Yes	6 41 19	5	
BB_34396780	6	Yes Yes Yes	2 22 28 48	51	HE_7186	13	No Yes Yes	50 50 2 17		SN_47500596	1	Yes Yes No	37 37 51	3	
BB_34426014	20	Yes Yes Yes	23 18 37 33	9	HE_7373	6	No Yes Yes	20 5 31 51		SN_48421098	15	Yes Yes Yes	5 50 42 40	12	
BB_34481534	20	No Yes No	21 51 4	10	HE_8157	1	No Yes Yes	7 1 23 8		SN_48480903	5	Yes No Yes	40 3 3	2	
BB_34512420	1	No Yes Yes	25 51 40	8	HE_8999	1	No Yes Yes	1 22 8 15		SN_48511110	8	Yes Yes No	9 42 9	18	
BB_35406375	1	Yes Yes Yes	4 20 43 47	6	NI_200000620	1	No Yes Yes	50 23 35 1		SN_48518085	7	Yes Yes No	8 42 2	35	
BB_35416179	6	Yes Yes Yes	38 3 39 10	50	NI_200000876	2	Yes Yes Yes	16 4 51 7	2		SN_49420959	10	Yes Yes Yes	5 17 3 16	51
BB_35432590	18	No Yes No	14 15 12		NI_200002153	1	No Yes Yes	1 22 8 15		SN_49430964	13	Yes No Yes	6 1 28	8	
BB_36465001	1	No Yes Yes	11 52 20	22	NI_40000555	19	No No Yes	24 50 50		SN_49451041	3	Yes Yes Yes	11 24 1 13	31	
BB_36471636	19	No No Yes	24 50 50		NI_400060071	1	No No Yes	25 10 4		SN_49486001	17	Yes No No	48 48 1	21	
BB_36505189	5	Yes Yes Yes	8 1 27 21	20	NI_400061210	16	Yes No Yes	48 25 17 45		SN_49531740	20	Yes Yes Yes	8 1 32 36	4	
BB_37451902	17	Yes Yes No	32 44 34	40	NI_400080170	4	Yes No Yes	5 4 11 29		SN_50430645	4	No Yes Yes	14 51 6	52	
BB_37451908	1	No Yes Yes	18 32 2	33	NI_40502382	1	Yes Yes No	37 37 51 3		SN_514380010	1	Yes No Yes	19 7 28	16	
BB_38441782	20	Yes Yes No	8 51 52	14	NI_40505301	3	No Yes Yes	17 14 51 36		SN_52411556	6	No Yes Yes	20 5 31 51		
BE_377	5	No Yes Yes	23 34 51	2	NI_40507282	1	No Yes Yes	23 26 8 20		SN_56393708	5	Yes Yes Yes	34 1 51 51	28	
BE_496	6	No Yes Yes	20 5 31 51		NI_405161720	17	No Yes Yes	15 52 16 2		ST_31350030	19	No No Yes	24 50 50		
BW_100-813-7	1	Yes Yes Yes	30 32 24 4	2	NI_500000256	1	Yes Yes Yes	2 52 50 17	9		ST_32360043	18	Yes Yes No	50 52 20	30
BW_102-714-6	2	Yes Yes Yes	24 36 6 18	2	NI_500000493	19	No Yes Yes	39 20 48 5		ST_34330013	1	Yes Yes No	35 4 50	17	
BW_103-261-4	2	Yes No No	52 50 2	1	NI_500000518	2	No Yes Yes	5 10 49 5		ST_34330014	17	Yes No No	48 48 48	21	
BW_104-114-5	1	No Yes Yes	23 40 14	48	NI_500003038	1	No Yes Yes	49 27 36 1		ST_35370071	1	Yes Yes No	33 43 48	5	
BW_106-767-3	6	No Yes Yes	20 5 31 51		NI_600041661	1	No No Yes	51 21 11		ST_36395811	20	No Yes Yes	49 21 2 8	14	
BW_119-771-0	17	Yes Yes Yes	52 23 13 1	10	NI_9700020	1	No No Yes	33 3 16		ST_38390006	18	No Yes No	19 43 14		
BW_119-813-3	1	Yes Yes Yes	23 36 13 14	39	NI_9700159	1	Yes No Yes	14 50 24 1		ST_39310015	1	No Yes No	35 44 10		
BW_120-670-3	17	Yes No No	48 48 21		NI_9700272	3	Yes No No	2 6 1		ST_41380015	20	Yes Yes No	2 9 48	4	
BW_126-020-2	2	No Yes Yes	5 10 49 5		NW_129660934	20	Yes Yes Yes	37 7 3 5 15		ST_41400587	1	No No Yes	8 52 13		
BW_126-762-3	20	Yes Yes Yes	17 2 18 39	33	NW_21661108	2	No Yes No	11 4 1		ST_41419101	1	Yes Yes Yes	16 28 15 36	4	
BW_131-115-0	1	No No Yes	44 45 2		NW_40060500	4	No Yes Yes	52 29 44 49		ST_42410003	6	No Yes Yes	52 30 1 47		
BW_137-770-0	18	No Yes No	2 10 4		NW_40100182	6	No Yes Yes	20 5 31 51		ST_42410008	13	No No Yes	5 2 3		
BW_140-020-6	3	Yes No No	48 52 2		NW_40100261	7	No Yes Yes	32 11 4 2		ST_44380061	1	Yes Yes Yes	38 41 46 3	41	
BW_152-308-7	20	Yes No No	50 42 50		NW_60072301	1	No Yes Yes	16 33 50 36		ST_44380072	10	No No Yes	50 44 34		
BW_152-772-1	1	No Yes Yes	19 10 30	23	NW_60090169	20	Yes Yes Yes	37 9 6 16 10		ST_45326568	11	Yes Yes No	48 1 51 42		
BW_157-260-8	1	Yes Yes Yes	32 38 46 6	43	NW_60100205	1	Yes Yes Yes	2 52 50 17 9		ST_45371005	1	Yes No No	8 23 4		
BW_167-06-8	1	Yes No Yes	6 41 18 5		NW_70200415	16	No Yes No	3 14 1		ST_49390001	20	Yes Yes Yes	37 10 41 50	3	
BW_165-568-0	10	No Yes No	2 4 3		NW_70284313	1	Yes Yes No	5 32 26 1		ST_49390053	17	Yes No No	48 48 21		
BW_170-568-9	6	Yes No Yes	2 3 52 21		NW_80100790	6	No Yes Yes	20 5 31 51		TH_4528230791	7	No Yes Yes	20 51 43 3		
BW_181-304-4	1	Yes No No	41 52 32		NW_91132502	1	No Yes No	30 1 9		TH_4531230792	1	No Yes Yes	24 11 52 9		
BW_2004-166-1	1	Yes Yes Yes	33 48 7 26	1	NW_91154108	1	No No Yes	49 50 6		TH_4626230766	1	No Yes No	13 14 10		
BW_2004-569-8	1	Yes Yes Yes	30 8 3 15 4		NW_91163705	6	No Yes Yes	20 5 31 51		TH_4634900423	6	No No No	14 5		
BW_236-115-9	11	Yes No Yes	1 3 7 10		NW_91163900	5	No No Yes	1 28 4		TH_4826900003	2	No Yes Yes	15 40 46 2		
BW_259-259-7	17	Yes No No	48 48 21		NW_91167504	2	Yes No Yes	4 52 12 15		TH_5138000179	1	Yes Yes Yes	5 20 43 45 6		
BW_765-306-3	2	No No Yes	12 7 1		NW_91174909	1	Yes Yes Yes	49 5 7 12 7		TH_5138000186	8	No Yes No	5 1 5		
BY_10134	1	Yes No Yes	7 38 38 11		NW_91175604	1	Yes Yes Yes	30 8 3 16 4		TH_5226240502	9	No No Yes	13 3 24		
BY_13103	3	No Yes Yes	25 35 14 52		RP_2375178300	5	Yes Yes No	52 6 52 23							
BY_24151	1	No No Yes	8 46 12		RP_2391187900	1	Yes Yes No	42 20 8 2							
BY_25111	17	Yes Yes Yes	13 1 17 5 1		RP_2716140900	1	Yes No Yes	27 52 23 7							
BY_25135	8	Yes No Yes	5 52 25 8		RP_2718135900	1	Yes Yes No	52 2 41 11							
BY_25668	1	Yes Yes No	5 3 5 47		SH_10L53089002	4	Yes Yes Yes	51 52 46 39 12							
BY_5158	1	Yes Yes Yes	16 29 15 36 4		SH_10L56010003	15	Yes Yes No	5 42 28 4							
BY_5166	2	No No Yes	52 51 1		SH_10L58064003	10	No No Yes	39 48 46							
BY_5171	1	Yes Yes Yes	36 44 7 27 1		SH_10L58103001	1	Yes No Yes	6 41 18 5							

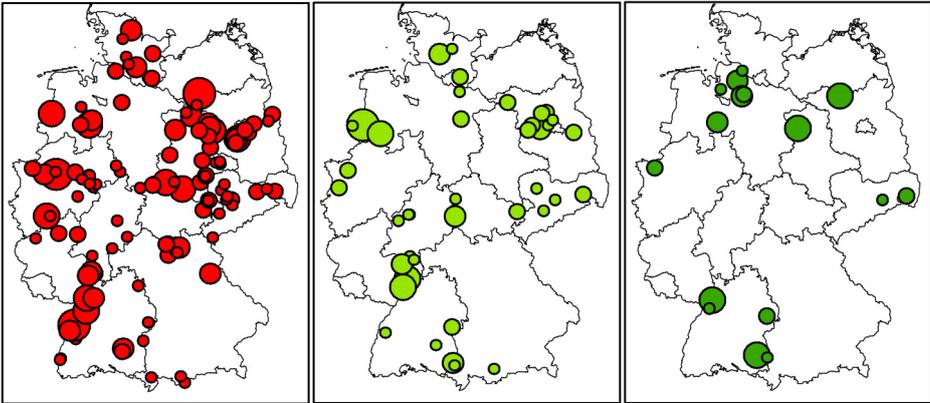
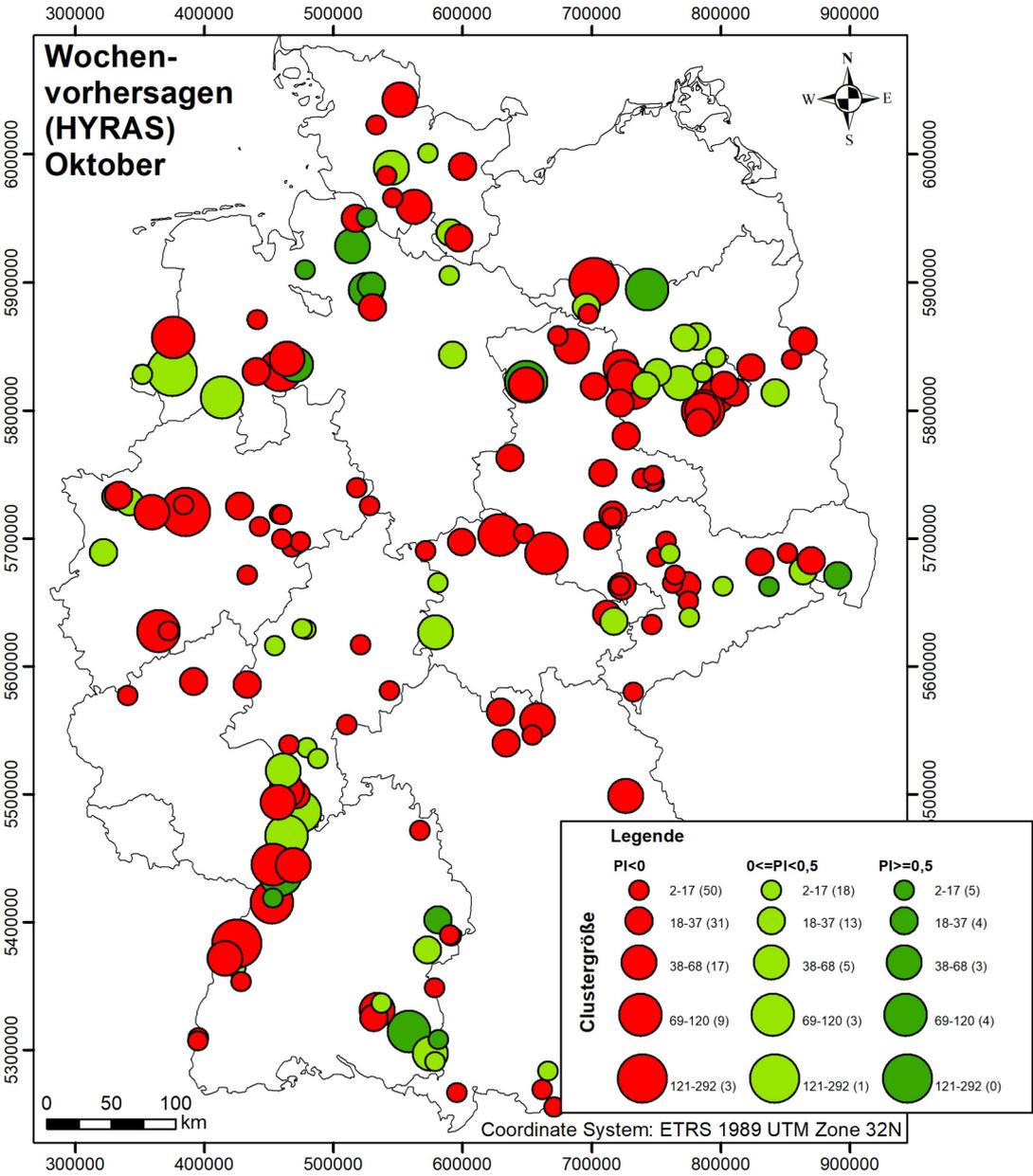
Anhang H - Fehlerübersicht aller Referenzmessstellen und aller Vorhersagen

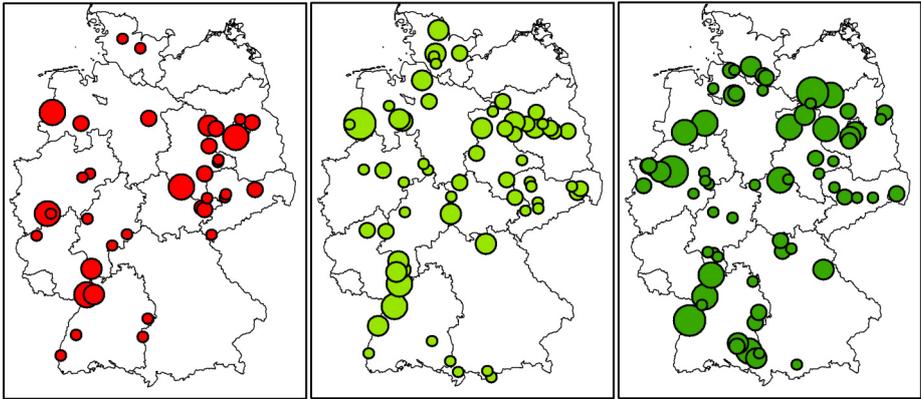
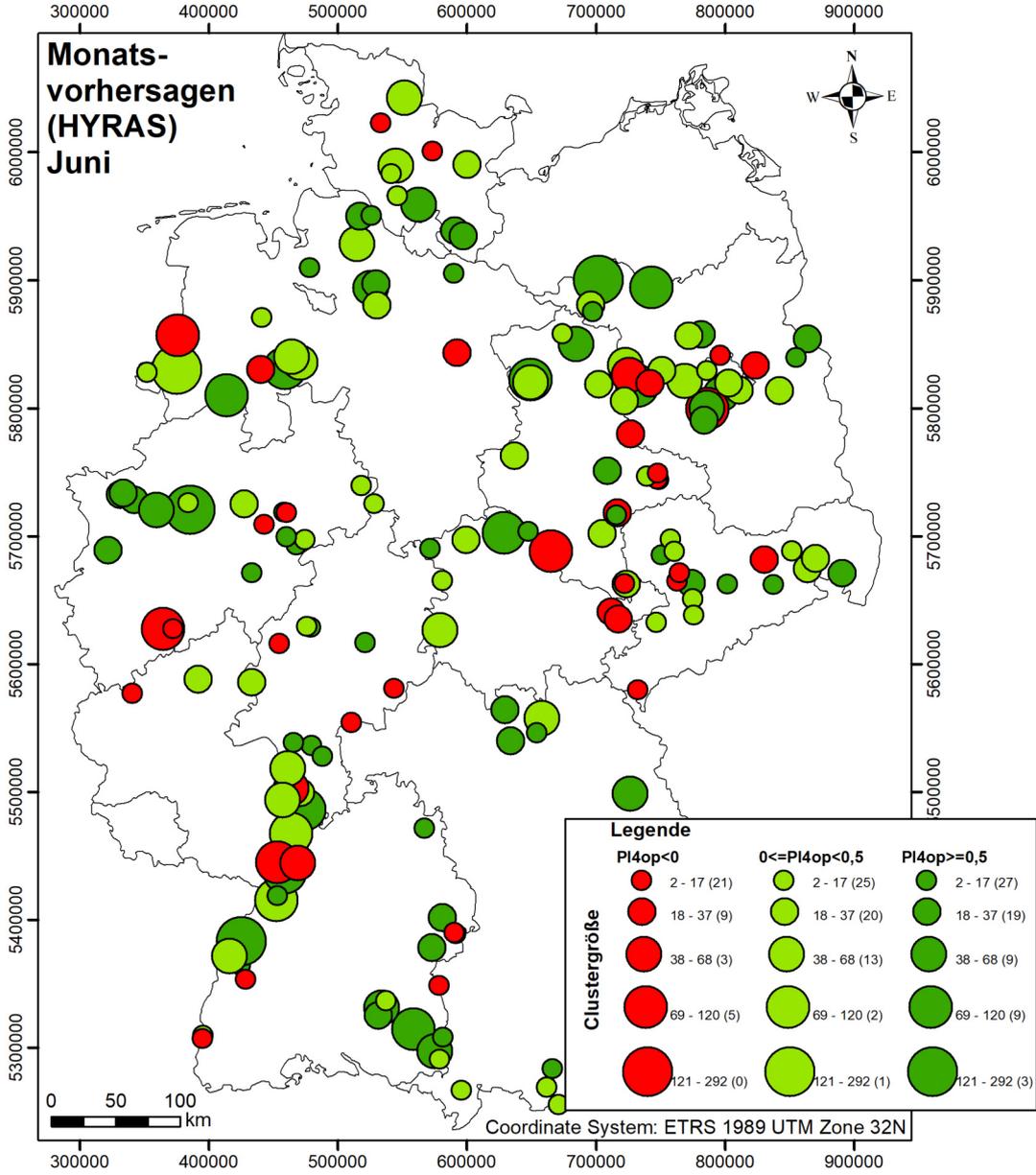
ID	HYRAS														EPISODES															
	Wochenvorhersage							Monatsvorhersage							Saisonale Vorhersage							Saisonale Vorhersage								
	NSE	RMSE	rRMSE	Bias	rBias	P11	P12	NSE	R ²	rRMSE	Bias	rBias	P14	P14op	NSE	R ²	rRMSE	Bias	rBias	P12	P12op	NSE	R ²	rRMSE	Bias	rBias	P12	P12op		
	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]	[]				
BB 27381011	0.99	0.020	1.26	0.009	0.58	0.53	0.97	0.80	0.040	2.57	0.018	1.17	0.88	0.75	0.90	0.82	0.076	4.87	0.011	0.74	0.93	0.86	0.78	0.67	0.128	8.27	0.021	1.34	0.85	0.66
BB 28410218	1.00	0.007	0.32	0.004	0.18	0.44	0.99	0.78	0.016	0.75	0.012	0.53	0.74	0.51	0.97	0.72	0.050	2.26	0.025	1.12	0.80	0.59	0.82	0.48	0.079	3.62	0.034	1.57	0.43	-0.51
BB 29370010	0.95	0.047	2.72	0.012	0.71	0.14	0.88	0.35	0.083	4.88	-0.007	-0.39	0.69	0.45	0.73	0.60	0.118	6.92	-0.004	-0.23	0.88	0.73	0.57	0.41	0.153	8.95	-0.004	-0.23	0.77	0.48
BB 30370039	0.99	0.018	1.04	0.001	0.05	0.56	0.97	0.84	0.038	2.20	0.010	0.60	0.85	0.67	0.89	0.81	0.076	4.42	0.002	0.10	0.93	0.84	0.76	0.71	0.111	6.45	-0.009	-0.50	0.85	0.61
BB 32449203	0.93	0.040	2.65	0.009	0.57	0.13	0.79	0.60	0.078	5.19	0.019	1.25	0.66	0.40	0.78	0.65	0.105	6.99	-0.030	-2.03	0.88	0.72	0.67	0.46	0.138	9.23	-0.050	-3.33	0.76	0.53
BB 32455305	0.99	0.022	1.25	0.008	0.44	0.57	0.97	0.85	0.046	2.60	0.016	0.92	0.86	0.72	0.90	0.86	0.097	5.42	-0.022	-1.21	0.94	0.85	0.81	0.72	0.127	7.09	-0.029	-1.60	0.86	0.73
BB 32529245	0.95	0.074	2.20	0.036	1.07	-0.04	0.76	0.67	0.180	5.35	0.083	2.46	0.53	0.15	0.64	0.53	0.301	8.96	0.005	0.15	0.75	0.44	0.53	0.41	0.119	9.52	-0.047	-1.39	0.70	0.22
BB 33392110	0.92	0.028	2.92	0.009	0.89	-0.16	0.77	0.39	0.051	5.36	0.013	1.35	0.46	0.08	0.67	0.58	0.073	7.64	-0.012	-1.26	0.83	0.54	0.60	0.44	0.088	9.21	-0.026	-2.75	0.76	0.37
BB 33465504	0.98	0.032	1.85	0.018	1.03	0.01	0.94	0.43	0.057	3.34	0.007	0.42	0.62	0.30	0.79	0.54	0.104	6.10	0.012	0.73	0.86	0.70	0.58	0.33	0.136	7.97	0.010	0.58	0.70	0.44
BB 34396780	0.98	0.026	1.70	0.009	0.57	-0.07	0.94	0.45	0.044	2.86	0.020	1.27	0.71	0.44	0.66	0.56	0.102	6.57	0.061	3.93	0.78	0.57	0.52	0.44	0.119	7.65	-0.028	1.78	0.70	0.38
BB 34426014	0.99	0.022	1.49	-0.004	-0.26	-0.08	0.97	0.46	0.035	2.40	-0.008	-0.53	0.45	-0.01	0.93	0.62	0.051	3.49	-0.005	-0.38	0.87	0.63	0.90	0.50	0.061	4.16	-0.004	-0.29	0.76	0.40
BB 34461534	0.99	0.028	1.22	0.015	0.63	-0.64	0.97	0.50	0.054	2.34	0.028	1.20	0.53	-0.12	0.85	0.60	0.107	4.59	0.022	0.95	0.80	0.56	0.82	0.62	0.132	5.65	0.009	0.40	0.73	0.34
BB 34512420	0.99	0.025	1.77	0.003	0.23	0.44	0.95	0.81	0.045	3.24	0.001	0.06	0.61	0.33	0.71	0.49	0.082	5.84	0.000	0.03	0.77	0.54	0.55	0.31	0.102	7.23	-0.014	-1.02	0.60	0.37
BB 35406375	0.99	0.009	0.74	0.003	0.23	0.44	0.95	0.84	0.023	2.95	0.008	0.63	0.76	0.51	0.85	0.78	0.055	4.56	-0.002	-0.13	0.83	0.57	0.59	0.81	0.079	6.55	-0.017	-1.39	0.76	0.41
BB 35416179	0.97	0.032	2.06	0.010	0.65	-0.23	0.88	0.52	0.073	4.68	0.019	1.19	0.52	0.40	0.65	0.59	0.134	8.52	0.009	0.54	0.82	0.59	0.58	0.51	0.152	7.90	0.006	0.38	0.75	0.43
BB 35432590	0.99	0.020	1.33	0.005	0.37	0.17	0.96	0.62	0.038	2.58	0.011	0.73	0.79	0.57	0.83	0.67	0.071	4.81	-0.004	-0.27	0.87	0.67	0.67	0.40	0.098	6.64	-0.023	-1.52	0.74	0.39
BB 36465001	0.99	0.015	0.77	0.003	0.16	0.18	0.98	0.55	0.024	1.17	0.004	0.22	0.73	0.44	0.98	0.69	0.047	2.31	0.006	0.29	0.79	0.55	0.60	0.60	0.080	3.96	0.009	0.45	0.37	-0.64
BB 36471636	0.99	0.016	1.71	-0.009	-0.89	-2.83	0.98	0.58	0.026	2.74	-0.013	-1.31	0.25	-0.35	0.93	0.64	0.046	4.84	-0.021	-2.15	0.76	0.46	0.89	0.57	0.051	5.32	-0.021	-2.14	0.72	0.33
BB 36505189	1.00	0.012	0.63	0.006	0.33	-0.10	0.99	0.64	0.023	1.22	0.014	0.71	0.44	0.03	0.98	0.73	0.044	2.30	0.019	0.98	0.81	0.53	0.97	0.60	0.054	2.83	0.020	1.05	0.69	0.17
BB 37451902	0.96	0.028	2.97	0.019	1.99	-2.11	0.88	0.49	0.050	5.23	0.035	3.66	0.45	-0.07	0.80	0.59	0.059	6.10	0.018	1.90	0.83	0.60	0.66	0.54	0.104	7.68	0.012	1.21	0.76	0.46
BB 37451908	0.99	0.018	1.47	0.008	0.67	-0.16	0.95	0.71	0.043	3.44	0.015	1.24	0.71	0.35	0.85	0.76	0.063	5.11	0.007	0.53	0.92	0.78	0.65	0.65	0.089	7.22	-0.009	-0.76	0.79	0.53
BB 38441782	0.98	0.020	2.31	0.000	-0.03	-0.42	0.92	0.35	0.038	4.44	0.004	0.42	0.45	-0.01	0.76	0.43	0.054	6.39	-0.001	-0.10	0.77	0.49	0.70	0.39	0.063	7.42	0.000	0.00	0.71	0.40
BE 377	0.88	0.051	3.44	0.017	1.12	0.07	0.81	0.31	0.069	4.63	0.005	0.34	0.48	0.18	0.65	0.41	0.097	6.52	0.016	1.08	0.74	0.57	0.64	0.32	0.106	7.11	0.011	0.76	0.69	0.38
BE 496	0.99	0.012	1.58	0.001	0.11	-0.14	0.95	0.52	0.022	2.76	0.001	0.18	0.61	0.33	0.87	0.61	0.035	4.44	-0.014	-1.76	0.81	0.64	0.68	0.42	0.046	5.78	-0.020	-2.47	0.66	0.33
BW 100-813-7	0.98	0.347	1.73	0.033	0.16	0.06	0.92	0.75	0.074	3.51	-0.013	-0.06	0.77	0.47	0.89	0.80	0.896	4.46	0.033	0.16	0.92	0.82	0.49	0.36	0.216	10.63	0.335	1.67	0.58	0.22
BW 102-714-6	0.99	0.394	0.92	0.096	0.22	0.31	0.96	0.84	0.889	2.07	0.081	0.19	0.76	0.51	0.83	0.82	1.788	4.16	0.049	0.11	0.87	0.68	0.63	0.67	2.667	6.21	0.219	0.55	0.75	0.32
BW 103-725-4	0.99	0.028	1.12	0.003	0.11	0.21	0.95	0.81	0.058	3.24	0.008	0.32	0.73	0.36	0.85	0.82	0.099	5.94	-0.010	-0.40	0.88	0.74	0.63	0.78	0.146	5.81	0.014	0.94	0.79	0.47
BW 104-114-5	0.97	0.053	1.80	0.019	0.64	0.28	0.92	0.64	0.091	3.12	-0.009	-0.29	0.77	0.59	0.85	0.76	0.117	4.00	-0.028	-0.96	0.92	0.77	0.58	0.44	0.235	7.68	-0.025	-0.84	0.72	0.46
BW 106-767-3	0.57	0.166	6.67	-0.032	-1.29	-0.07	0.24	0.38	0.282	11.38	-0.064	-2.59	0.47	0.18	0.29	0.23	0.314	12.67	-0.290	0.64	0.39	0.12	0.16	0.16	0.340	1.70	-0.064	-2.58	0.56	0.23
BW 119-771-0	1.00	0.042	0.62	0.016	0.24	0.75	0.99	0.85	0.124	1.83	0.037	0.55	0.88	0.69	0.88	0.82	0.318	4.71	0.123	1.82	0.85	0.65	0.71	0.47	0.514	7.61	0.160	2.36	0.64	0.20
BW 119-813-3	0.97	0.225	1.90	0.309	0.81	-0.17	0.87	0.41	1.469	3.85	0.407	1.07	0.53	0.15	0.77	0.55	1.755	4.60	0.309	0.81	0.79	0.68	0.52	0.68	2.792	7.32	0.459	1.20	0.59	0.23
BW 120-670-3	1.00	0.012	0.50	0.002	0.09	0.62	0.98	0.90	0.040	1.64	0.009	0.38	0.82	0.62	0.91	0.83	0.088	3.64	-0.015	-0.61	0.85	0.62	0.90	0.77	0.112	6.42	0.001	0.60	0.74	0.43
BW 126-020-2	0.87	0.055	1.39	-0.029	-0.74	-0.20	0.46	0.67	0.125	3.17	-0.082	-2.07	0.35	0.03	0.34	0.42	0.191	4.81	-0.092	-2.33	0.63	0.21	0.01	0.24	0.245	4.19	-0.069	-1.74	0.41	-0.04
BW 126-762-3	0.99	0.091	1.04	0.013	0.15	0.41	0.95	0.92	0.244	2.81	0.025	0.29	0.74	0.42	0.70	0.70	0.568	6.55	0.021	0.24	0.85	0.64	0.60	0.67	0.675	7.77	0.115	1.32	0.76	0.45
BW 131-115-0	0.84	0.131	4.41	0.072	2.41	0.13	0.72	0.58	0.206	6.90	0.045	1.52	0.64	0.40	0.70	0.61	0.226	7.60	-0.046	-1.54	0.85	0.69	0.29	0.27	0.403	13.52	0.033	1.12	0.58	0.22
BW 137-770-0	1.00	0.022	0.80	0.010	0.37	0.77	0.96	0.90	0.066	2.44	0.025	0.93	0.89	0.74	0.90	0.88	0.108	3.99	0.002	0.07	0.94	0.86	0.55	0.51	0.219	8.08	0.041	1.52	0.72	0.43
BW 140-020-6	0.66	0.137	3.06	0.005	0.10	-0.13	0.41	0.42	0.225	5.03	-0.082	-1.83	0.50	0.20	0.39	0.31	0.305	6.82	0.027	0.60										

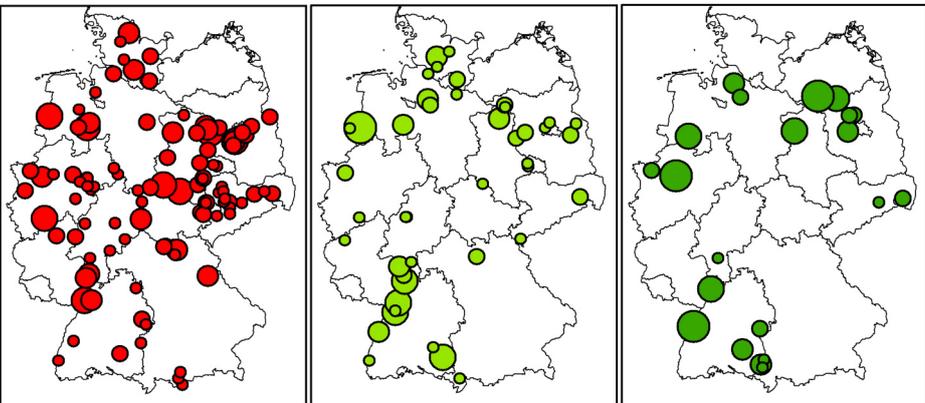
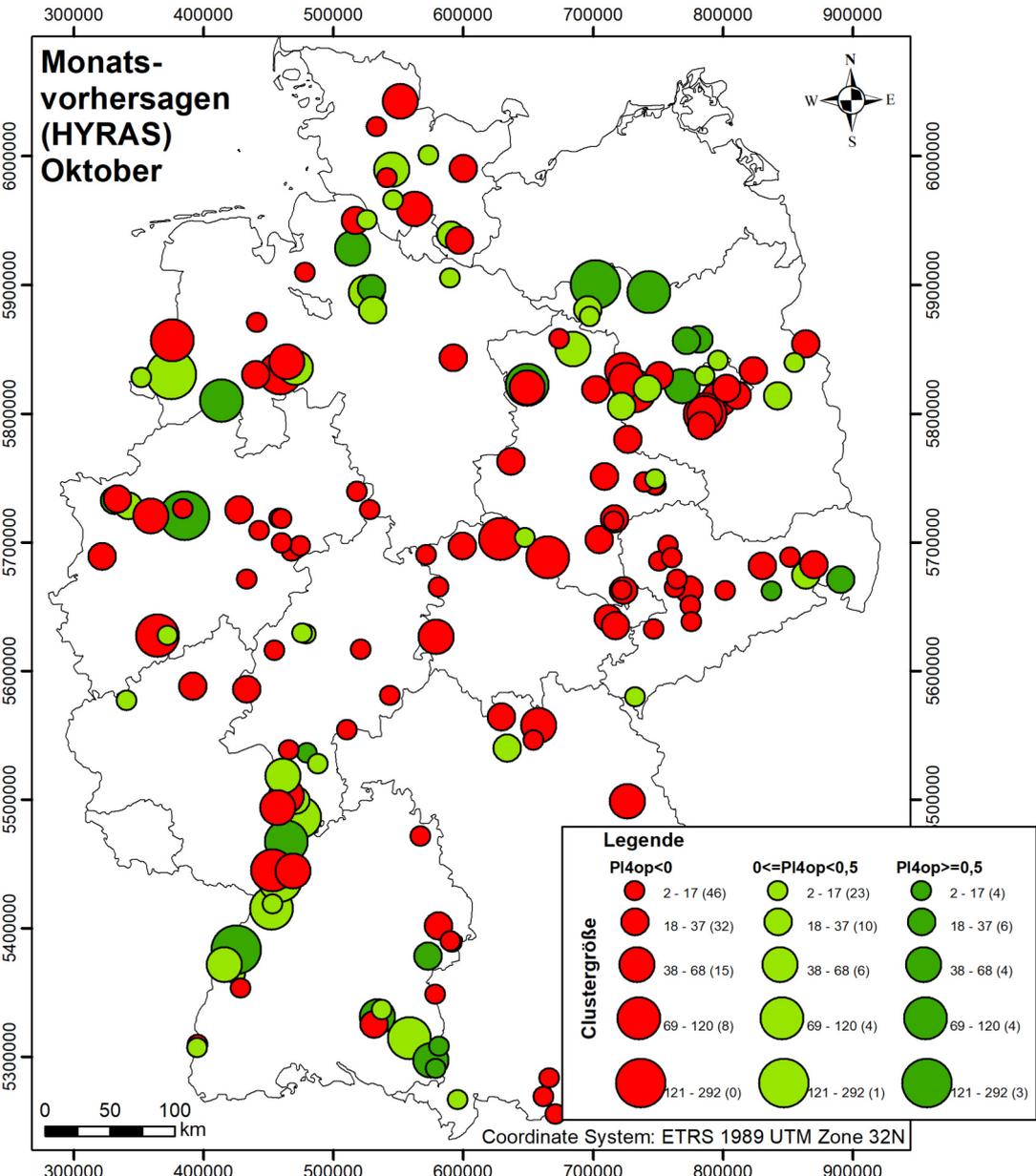
ID	HYRAS															EPIISODES														
	Wochenvorhersage					Monatsvorhersage					Saisonale Vorhersage					Saisonale Vorhersage														
	NSE	RMSE	rRMSE	Bias	rBias	PI1	NSE	R ²	RMSE	rRMSE	Bias	rBias	PI4	PI4op	NSE	R ²	RMSE	rRMSE	Bias	rBias	PI12	PI2op	NSE	R ²	RMSE	rRMSE	Bias	rBias	PI12	PI2op
NW 91163900	0.97	0.761	2.52	0.581	1.93	-5.38	0.89	0.44	1.318	4.37	0.313	1.04	-0.08	-2.36	0.63	0.48	0.285	4.95	0.037	0.65	0.75	0.59	0.43	0.31	0.378	6.56	0.085	1.48	0.67	0.30
NW 91167504	0.99	0.385	1.29	0.171	0.57	0.31	0.95	0.82	0.816	2.74	0.254	0.85	0.77	0.51	0.92	0.80	1.197	4.01	-0.309	-1.04	0.94	0.85	0.77	0.67	2.337	7.84	0.151	0.51	0.76	0.46
NW 91174909	0.97	0.148	1.77	0.104	1.24	-0.24	0.90	0.60	0.276	3.31	0.167	2.00	0.61	0.18	0.81	0.56	0.389	4.66	0.077	0.92	0.86	0.71	0.62	0.40	0.516	6.18	0.159	1.91	0.76	0.42
NW 91175604	0.98	1.332	2.34	0.484	0.85	-0.40	0.95	0.70	2.542	4.47	0.491	0.86	0.47	0.02	0.86	0.69	4.255	7.49	-0.424	-0.75	0.87	0.70	0.77	0.60	5.570	9.80	1.314	2.31	0.80	0.49
RP 2375178300	0.95	0.038	1.78	0.001	0.06	0.07	0.70	0.63	0.101	4.71	-0.009	-0.42	0.43	0.04	0.48	0.41	0.141	6.57	-0.039	-1.81	0.71	0.48	0.24	0.33	0.168	7.79	-0.028	-1.31	0.58	0.18
RP 2391187900	0.98	0.035	1.14	0.003	0.10	0.11	0.82	0.76	0.103	3.38	-0.005	-0.15	0.55	0.04	0.46	0.40	0.195	6.37	-0.011	-0.35	0.58	0.05	0.35	0.33	0.230	7.52	-0.010	-0.32	0.38	-0.26
RP 2716140900	0.98	0.069	1.51	0.029	0.63	0.42	0.87	0.84	0.201	4.37	0.085	1.84	0.72	0.41	0.64	0.56	0.346	7.53	-0.029	-0.63	0.79	0.54	0.27	0.32	0.453	9.87	-0.005	-0.11	0.67	0.30
RP 2718133900	0.91	0.104	1.81	-0.015	-0.25	-0.45	0.75	0.53	0.176	3.06	-0.036	-0.63	0.61	0.24	0.63	0.48	0.285	4.95	0.037	0.65	0.75	0.59	0.43	0.31	0.378	6.56	0.085	1.48	0.67	0.30
SH 10153089002	0.99	0.044	1.18	0.025	0.69	0.34	0.95	0.65	0.103	2.79	0.053	1.43	0.77	0.44	0.89	0.84	0.145	3.93	0.023	0.63	0.94	0.85	0.75	0.69	0.231	6.23	0.040	1.08	0.83	0.66
SH 10156010003	0.96	0.021	2.40	0.002	-0.26	-0.01	0.93	0.48	0.030	3.42	-0.003	-0.34	0.67	0.37	0.85	0.67	0.045	5.05	-0.021	-2.37	0.91	0.77	0.80	0.66	0.054	6.10	-0.024	-2.64	0.85	0.65
SH 10158064003	0.95	0.021	2.60	-0.003	-0.33	-0.04	0.90	0.50	0.033	4.09	-0.001	-0.11	0.62	0.29	0.86	0.61	0.047	5.87	-0.005	-0.68	0.87	0.67	0.82	0.60	0.052	6.48	0.002	0.31	0.83	0.57
SH 10158103001	0.92	0.044	2.74	0.011	0.66	0.20	0.82	0.60	0.073	4.51	0.014	0.85	0.72	0.50	0.78	0.61	0.113	6.95	-0.013	-0.82	0.87	0.77	0.43	0.24	0.168	10.38	0.000	0.01	0.71	0.44
SH 10158137003	0.96	0.028	3.30	-0.009	-1.10	-1.75	0.89	0.43	0.048	5.53	-0.015	-1.80	0.30	-0.96	0.72	0.47	0.068	7.94	-0.020	-2.38	0.76	0.36	0.71	0.45	0.072	8.32	-0.030	-3.51	0.72	0.25
SH 10159087003	0.98	0.019	1.71	0.005	-0.41	0.37	0.90	0.65	0.039	3.52	-0.007	-0.59	0.78	0.62	0.87	0.87	0.051	4.59	-0.019	-1.66	0.94	0.86	0.65	0.63	0.093	8.34	-0.043	-3.83	0.83	0.53
SH 10160002001	0.97	0.031	1.77	0.003	0.14	0.27	0.92	0.71	0.055	3.13	0.007	0.41	0.77	0.52	0.85	0.79	0.078	4.43	-0.008	-0.44	0.92	0.82	0.64	0.51	0.133	7.51	-0.002	-0.12	0.80	0.61
SH 10160025003	0.73	0.035	5.87	-0.005	-0.83	0.11	0.57	0.29	0.053	8.79	-0.003	-0.50	0.30	0.16	0.65	0.11	0.057	9.49	0.002	0.38	0.61	0.36	0.64	0.12	0.056	9.39	0.006	0.93	0.59	0.25
SH 10161042008	0.99	0.024	1.40	0.004	0.23	-0.16	0.94	0.70	0.049	2.88	0.011	0.65	0.63	0.21	0.83	0.63	0.092	5.41	-0.006	-0.38	0.81	0.59	0.80	0.51	0.104	6.08	-0.007	-0.43	0.72	0.43
SH 10162020008	1.00	0.017	0.82	0.008	0.38	0.58	0.97	0.85	0.041	2.00	0.024	1.15	0.85	0.66	0.88	0.86	0.096	4.64	0.006	0.28	0.93	0.77	0.80	0.80	0.121	5.84	0.002	0.11	0.87	0.69
SN 46421125	0.98	0.054	1.36	0.023	0.57	0.21	0.89	0.74	0.117	2.93	0.053	1.34	0.51	0.15	0.68	0.57	0.195	4.89	0.011	0.27	0.75	0.53	0.37	0.39	0.264	6.59	0.012	0.52	0.56	-0.07
SN 4710412	1.00	0.020	0.49	0.009	0.21	0.06	0.98	0.86	0.054	1.33	0.024	0.60	0.60	0.13	0.82	0.76	0.172	4.26	0.069	1.72	0.71	0.44	0.44	0.59	0.264	6.11	0.014	0.36	0.20	-0.34
SN 47420010	0.99	0.031	1.33	0.009	0.39	0.24	0.92	0.64	0.079	3.38	0.027	1.15	0.58	-0.06	0.85	0.58	0.123	5.25	0.000	-0.01	0.82	0.59	0.69	0.54	0.175	7.43	0.006	0.26	0.70	0.35
SN 47500596	0.99	0.025	1.02	0.015	0.61	-0.03	0.96	0.75	0.063	2.61	0.041	1.67	0.35	-0.44	0.85	0.66	0.127	5.24	0.037	1.54	0.75	0.41	0.79	0.58	0.151	6.21	0.020	0.84	0.67	0.19
SN 48421093	0.99	0.092	1.38	0.027	0.41	-0.71	0.97	0.74	0.224	3.36	0.094	1.41	-0.07	-1.26	0.84	0.71	0.395	5.92	0.084	1.26	0.84	0.61	0.79	0.68	0.523	7.84	0.088	1.32	0.75	0.27
SN 48480903	0.91	0.086	2.79	0.050	1.62	-0.15	0.84	0.54	0.123	4.00	0.042	1.35	0.52	0.30	0.66	0.53	0.201	6.53	-0.008	-0.27	0.80	0.58	0.42	0.29	0.272	8.83	-0.048	-1.57	0.63	0.31
SN 48511110	1.00	0.014	0.63	0.009	0.40	-0.32	0.98	0.73	0.043	1.90	0.029	1.28	0.34	-0.27	0.93	0.75	0.080	3.50	0.041	1.81	0.78	0.51	0.91	0.72	0.090	3.10	0.043	1.88	0.69	0.20
SN 48518085	0.98	0.071	2.28	0.023	0.73	0.05	0.88	0.42	0.141	4.53	-0.001	-0.02	0.61	0.22	0.87	0.70	0.237	7.64	-0.011	-0.35	0.83	0.67	0.76	0.56	0.279	8.99	-0.018	-0.58	0.77	0.45
SN 49420959	0.98	0.040	2.11	0.022	1.16	-0.44	0.93	0.47	0.068	3.57	0.031	1.66	0.36	-0.13	0.86	0.50	0.103	5.42	0.022	1.18	0.69	0.43	0.75	0.42	0.132	6.96	0.021	1.13	0.55	0.07
SN 49430964	0.96	0.163	3.02	0.039	0.73	-0.07	0.90	0.48	0.277	5.14	0.080	1.48	0.54	0.21	0.83	0.60	0.369	6.84	0.045	0.84	0.88	0.68	0.72	0.47	0.483	8.95	-0.014	-0.27	0.81	0.34
SN 49451041	0.97	0.062	1.89	0.021	0.63	0.19	0.91	0.63	0.123	3.73	0.026	0.78	0.51	0.30	0.70	0.49	0.195	5.93	-0.046	-1.39	0.76	0.52	0.55	0.42	0.229	6.95	-0.049	-1.49	0.65	0.29
SN 49486001	1.00	0.012	0.40	0.003	0.10	0.38	1.00	0.75	0.026	0.89	0.002	0.08	0.76	0.49	0.97	0.77	0.046	1.55	-0.006	-0.21	0.84	0.60	0.94	0.75	0.055	1.87	-0.009	-0.32	0.78	0.39
SN 49531740	0.99	0.024	0.69	0.010	0.29	0.52	0.97	0.80	0.068	1.98	0.030	0.87	0.76	0.42	0.92	0.82	0.141	4.07	0.035	1.02	0.85	0.66	0.81	0.72	0.177	5.13	0.046	1.33	0.70	0.42
SN 50430645	0.99	0.024	1.11	0.017	0.79	-0.59	0.98	0.48	0.041	1.93	0.027	1.26	0.63	0.20	0.89	0.59	0.080	3.77	0.025	1.19	0.81	0.64	0.80	0.53	0.102	4.81	0.038	1.77	0.72	0.44
SN 514350010	0.70	0.032	3.58	0.016	1.78	-0.17	0.34	0.38	0.053	6.08	0.016	1.83	0.39	0.01	0.30	0.29	0.071	8.03	-0.005	-0.60	0.69	0.37	-0.05	0.15	0.088	10.02	-0.007	1.74	0.55	0.10
SN 52411556	0.93	0.039	2.82	0.011	0.82	0.10	0.82	0.53	0.074	5.35	0.017	1.23	0.51	0.23	0.70	0.44	0.107	7.72	0.010	0.71	0.82	0.64	0.52	0.33	0.129	9.30	0.018	1.32	0.67	0.34
SN 56393708	0.88	0.066	5.52	0.047	3.92	-0.52	0.70	0.26	0.097	8.07	0.054	4.46	0.23	-0.19	0.56	0.23	0.108	9.02	0.025	2.08	0.69	0.35	0.35	0.16	0.122	10.16	0.031	2.55	0.55	0.12
ST 31350030	0.97	0.029	1.79	0.010	0.62	-0.37	0.91	0.48	0.054	3.35	0.016	0.98	0.58	-0.01	0.85	0.77	0.083	5.16	0.013	0.81	0.88	0.73	0.77	0.67	0.095	5.96	0.002	0.16	0.85	0.61
ST 32360043	0.99	0.034	1.44	0.011	0.46	0.16	0.94	0.73	0.070	2.98	0.029	1.24	0.69	0.45	0.84	0.79	0.151	6.41	0.003	0.15	0.85	0.63	0.69	0.66	0.188	8.00	-0.016	-0.69	0.79	0.44
ST 34330013	1.00	0.010	0.48	0.000	0.02	0.64	0.99	0.90	0.020	0.97	0.004	0.19	0.87	0.72	0.84	0.79	0.067	3.24	0.015	0.71	0.89	0.74	0.79	0.63	0.081	3.95	0.002	0.11	0.82	0.49
ST 34330014	0.99	0.013	0.85	0.003	0.20	0.34	0.93	0.81	0.034	2.23	0.011	0.75	0.67	0.21	0.81	0.71	0.073	4.79	0.027	1.74	0.85	0.58	0.74	0.75	0.084	5.49	0.036	2.36	0.76	0.50
ST 35370071	0.96	0.076	2.12	0.033	0.91	-0.39	0.87	0.70	0.176	4.92	0.084	2.34	0.36	-0.07	0.63	0.61	0.304	8.48	0.117	3.27	0.75	0.52	0.55	0.52	0.353	9.86	-0.027	0.74	0.72	0.34

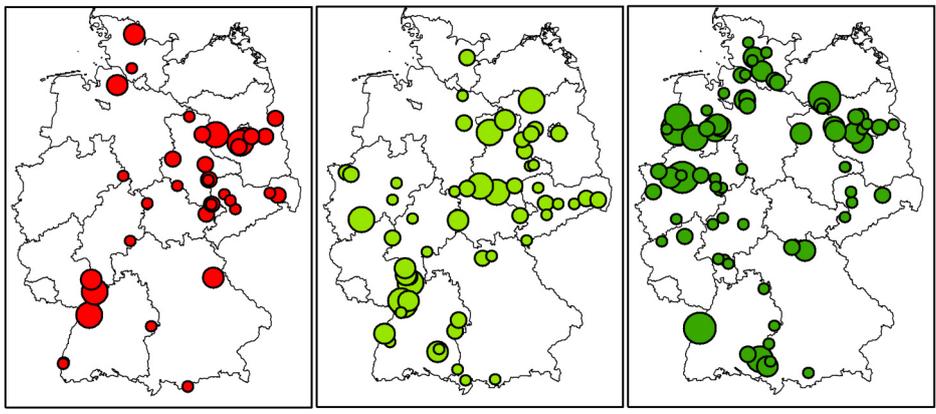
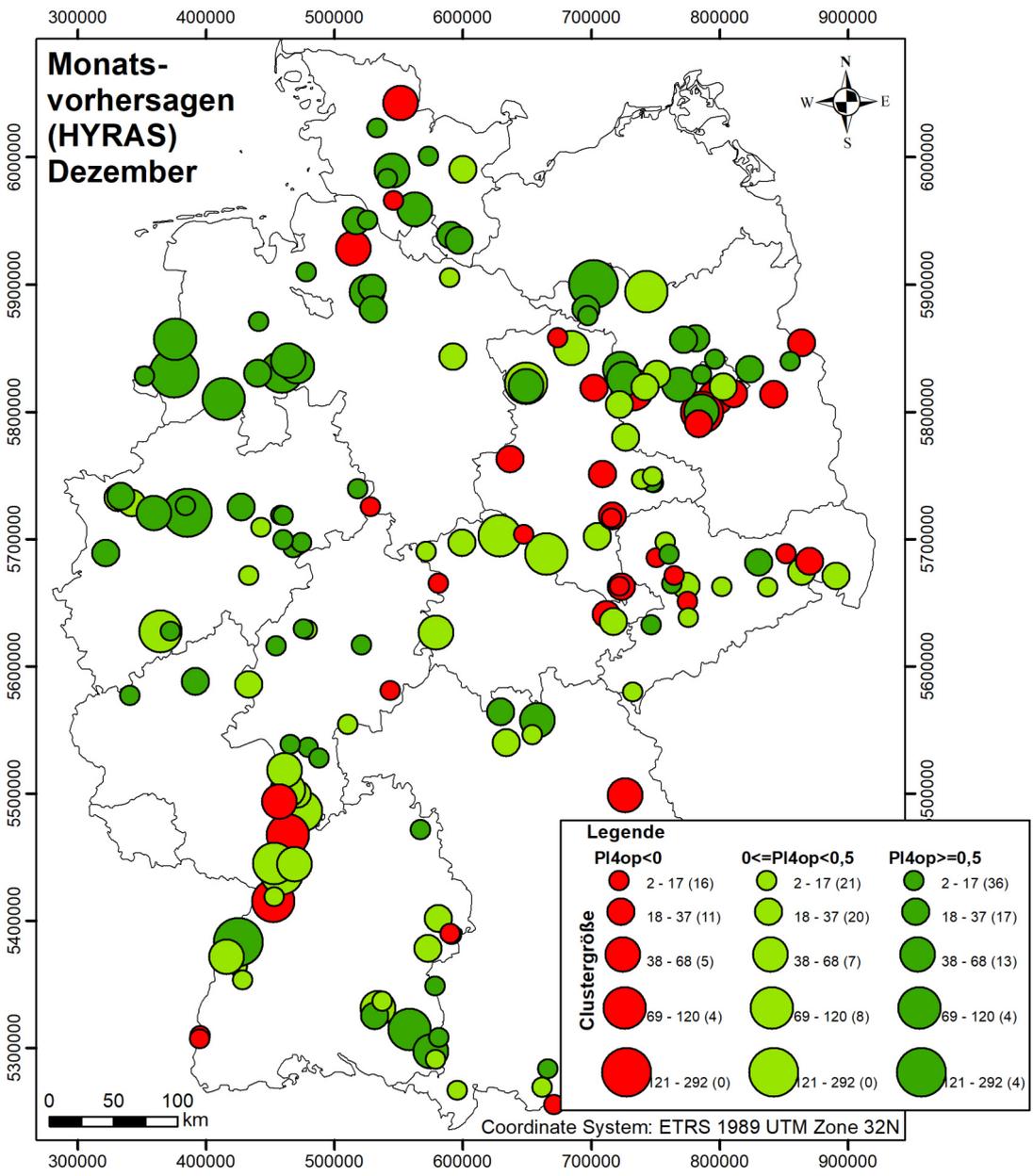
Anhang I - Übersichtskarten bez. der Vorhersagequalität einzelner (ausgewählter) Monate für alle Vorhersagehorizonte

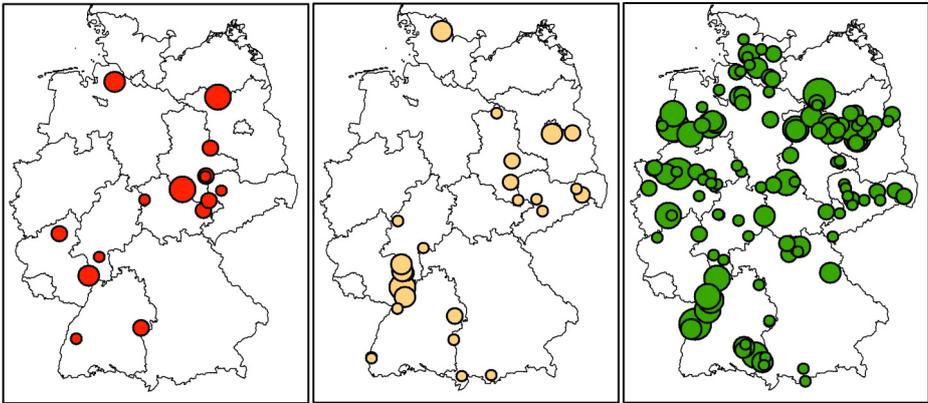
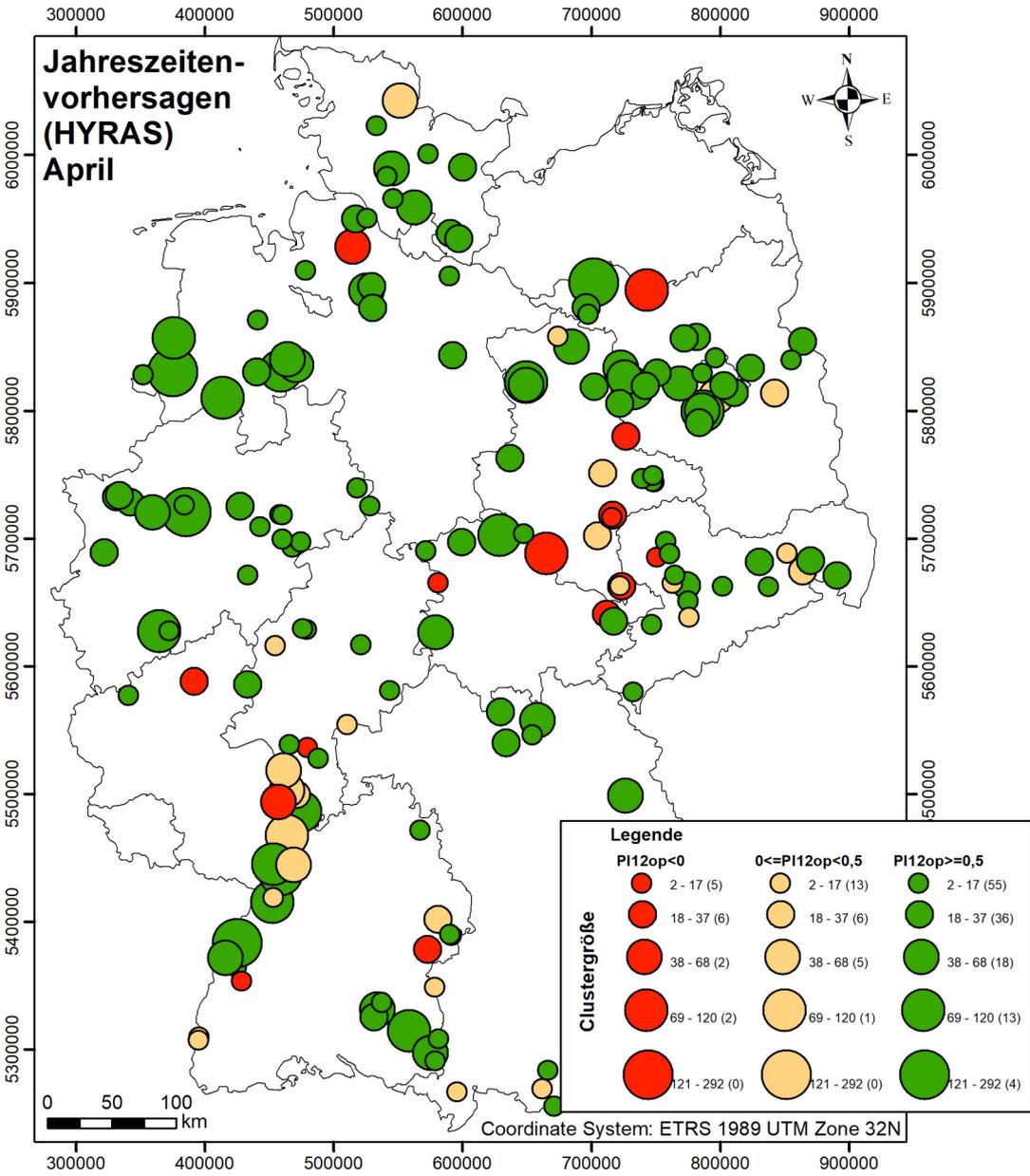


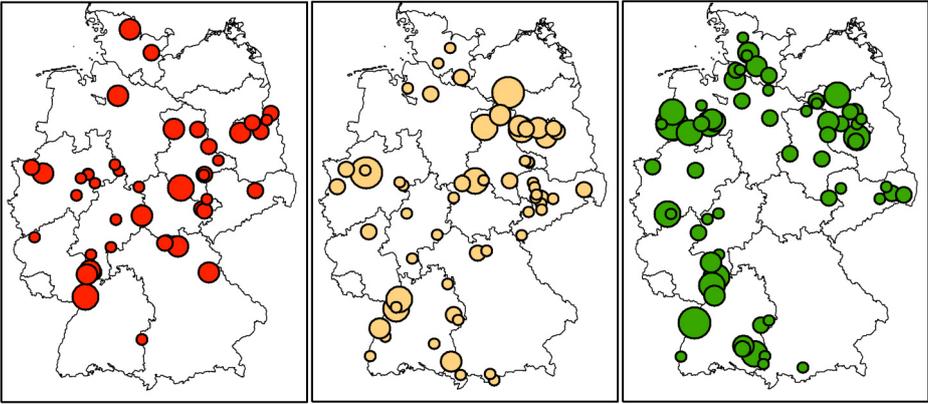
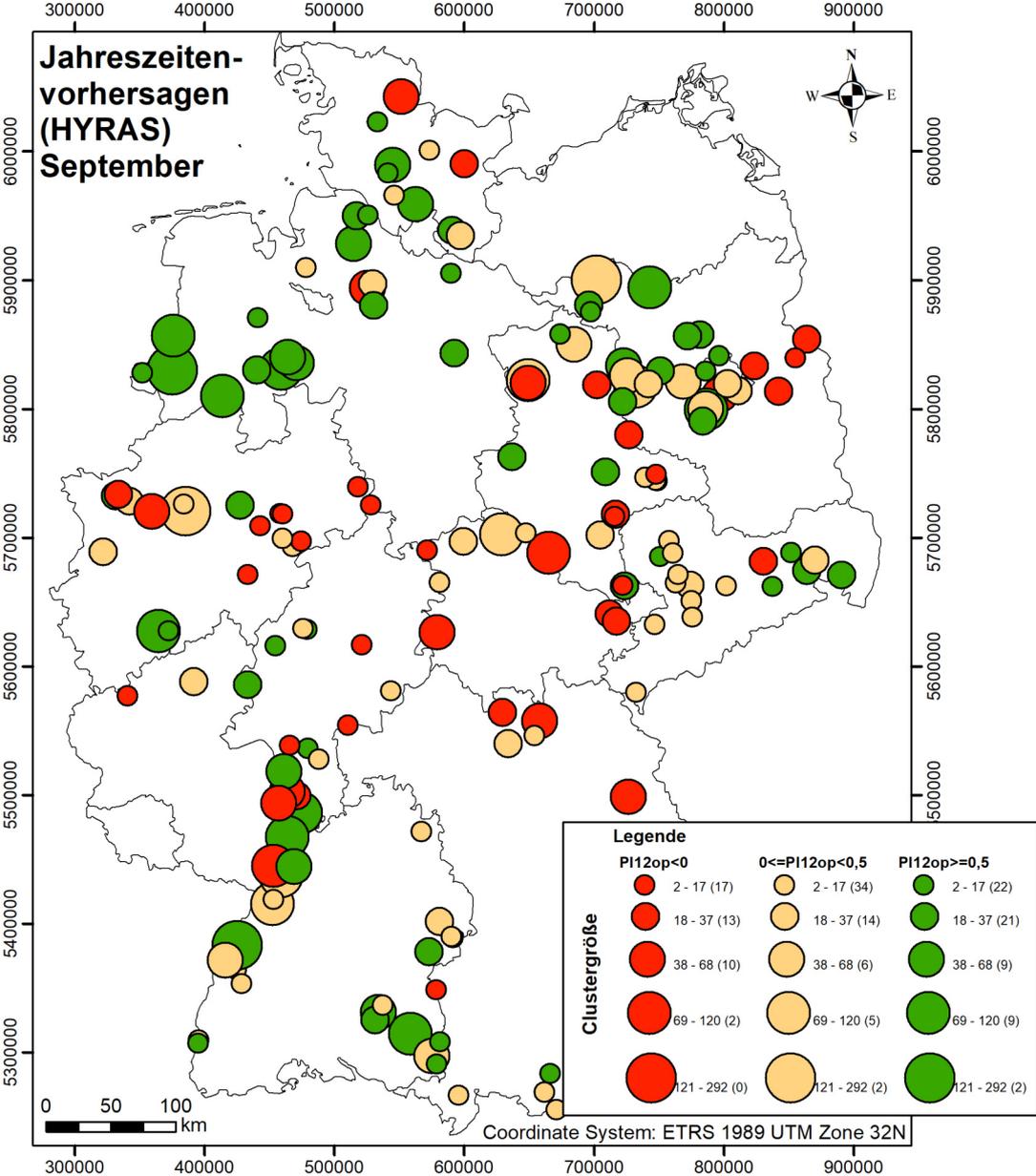












Anhang J - Liste von Messstellen mit Instabilitäten bei EPISODES-basierten Jahreszeitemvorhersagen

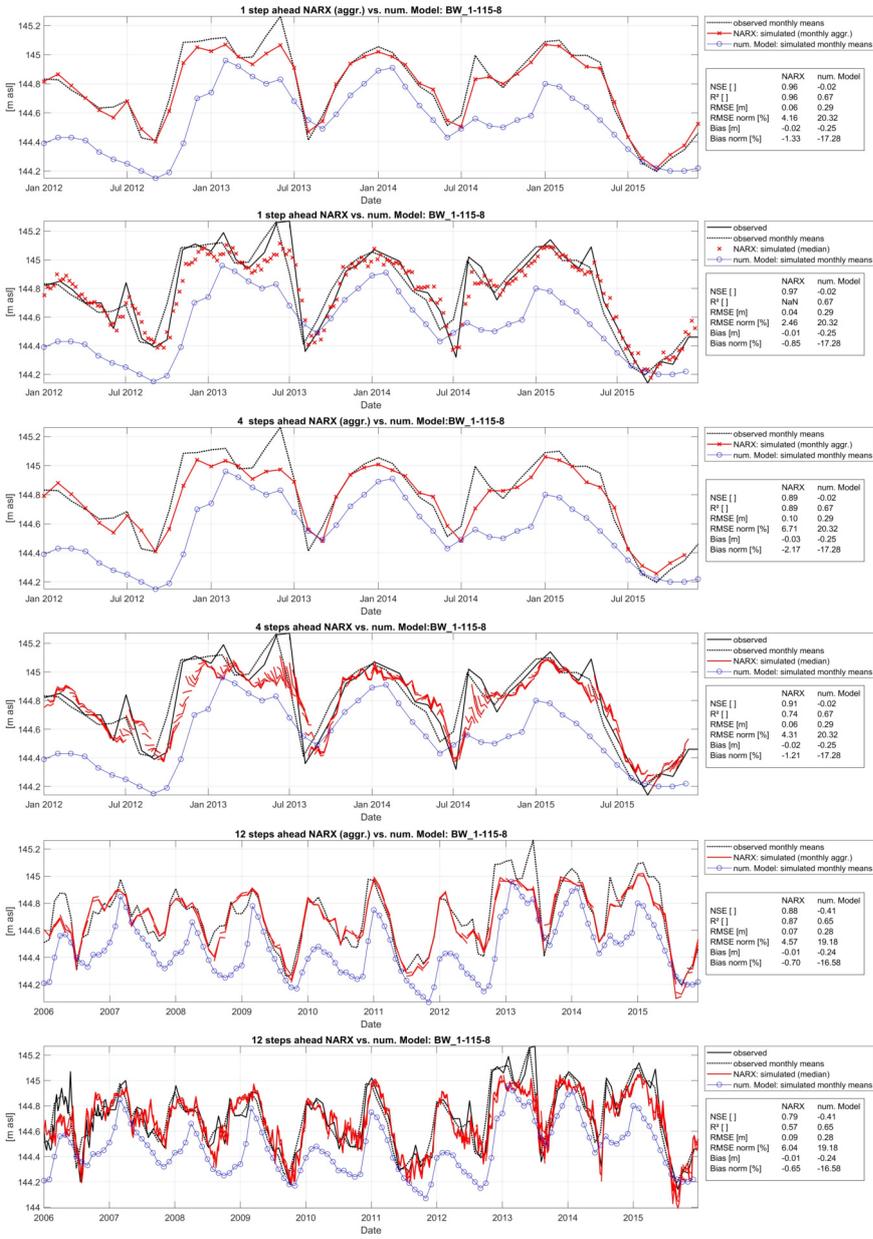
Messstellen die deutliche und häufige Instabilitäten bei EPISODES-basierten Vorhersagen aufweisen:

- BB_28410218
- BB_33465504
- BB_35406375
- BB_36465001
- BW_100-813-7
- BW_131-115-0
- BW_16-706-8
- BW_2004-569-8
- BW_765-306-3
- BY_5171
- HE_6615
- NI_200002153
- NW_91154108
- SN_47410412
- ST_41419101
- TH_5138000179

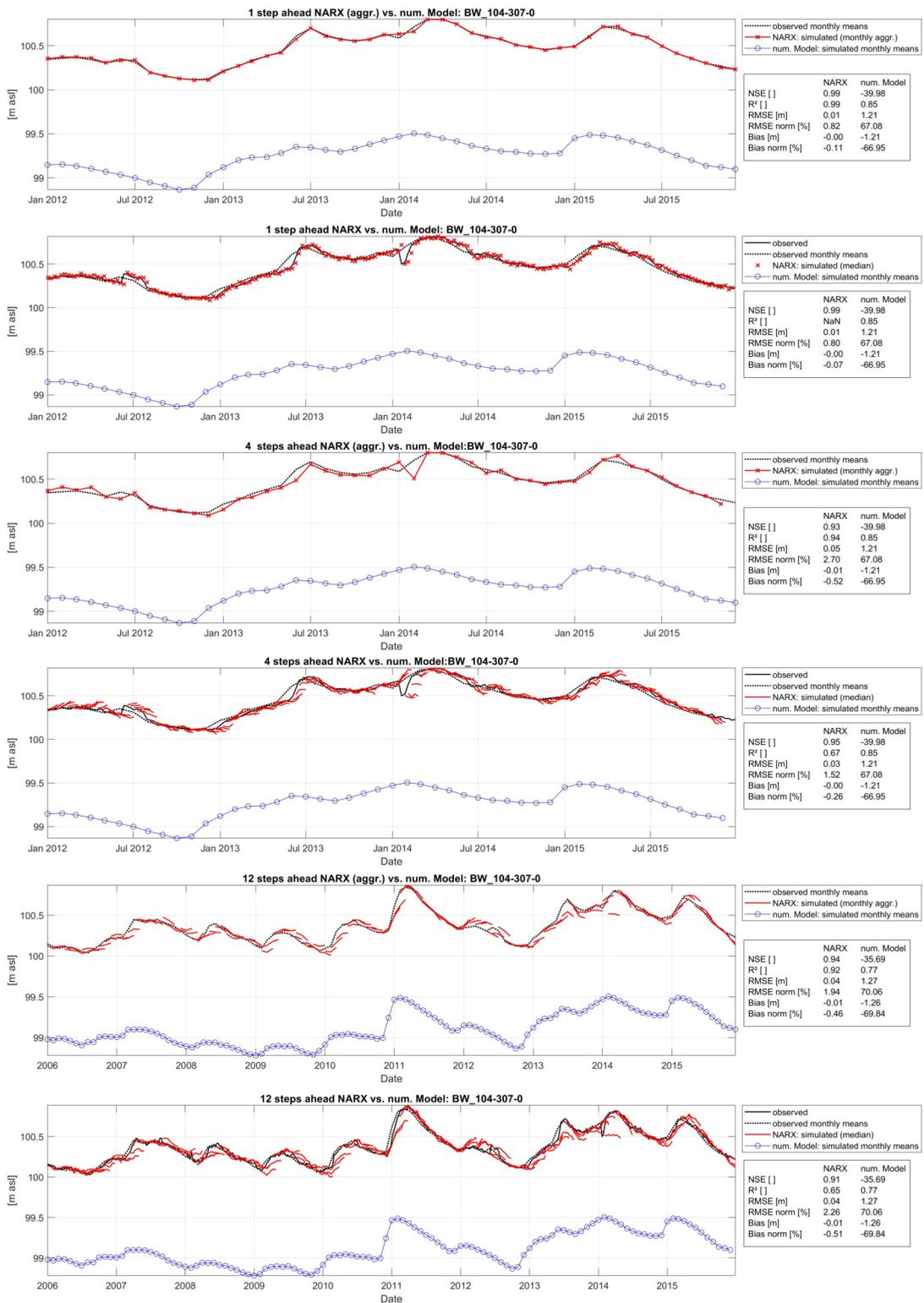
Messstellen die nur undeutlich oder in sehr geringem Umfang Instabilitäten aufweisen:

- BB_27381011
- BW_126-020-2
- HE_12447
- HE_13500
- HE_8999
- NI_500000256
- NI_9700159
- NW_40060500
- ST_44380061
- TH_4826900003

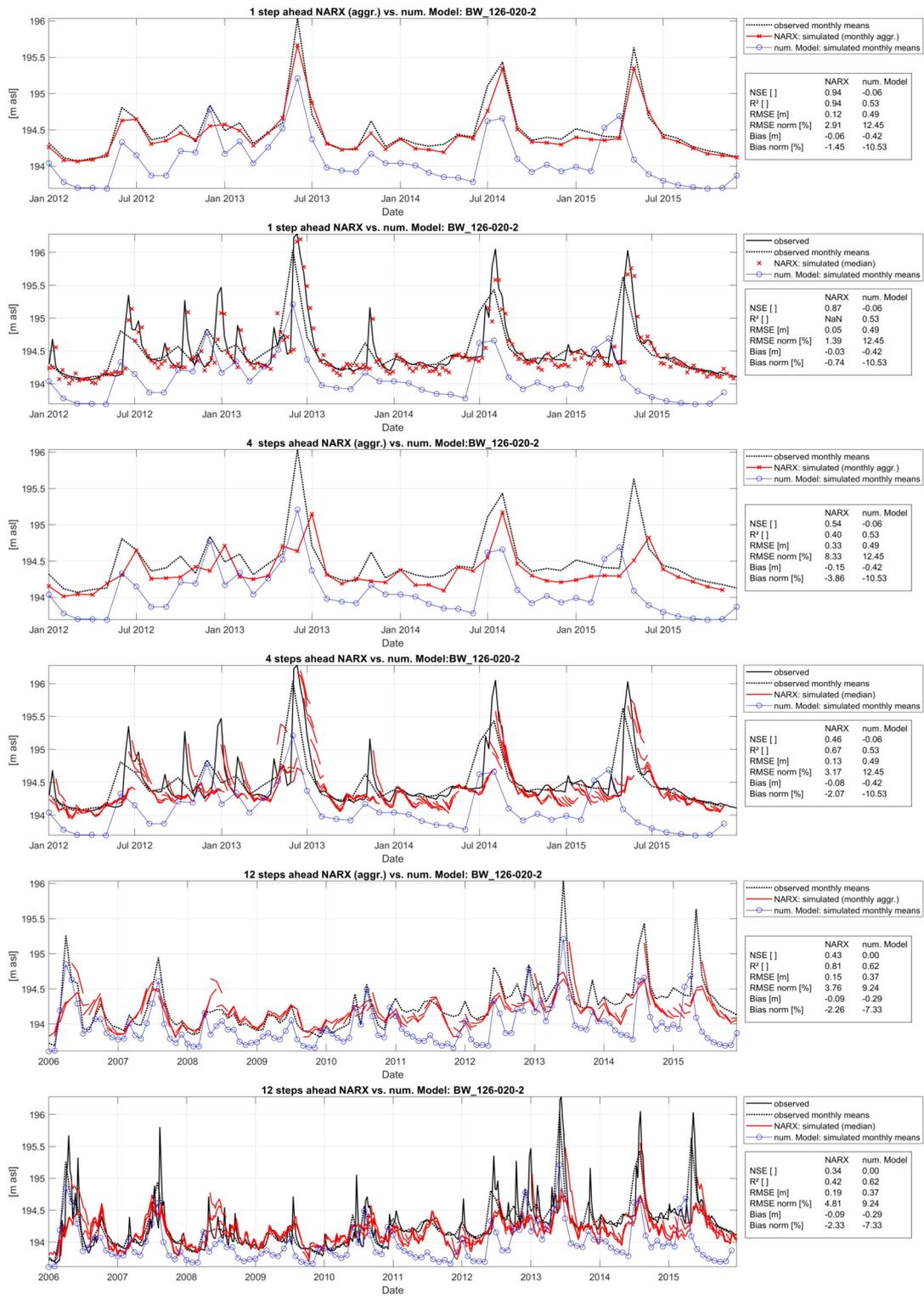
Anhang K - Modellvergleich der Vorhersagegüte an den Messstellen



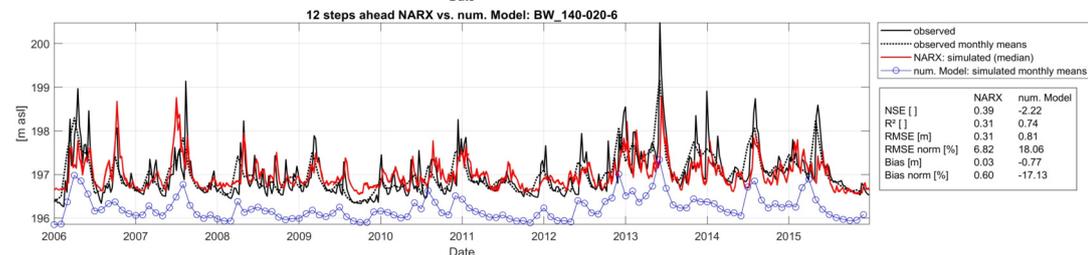
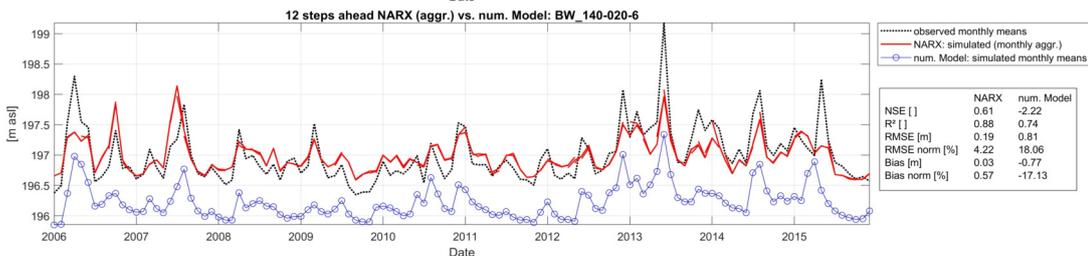
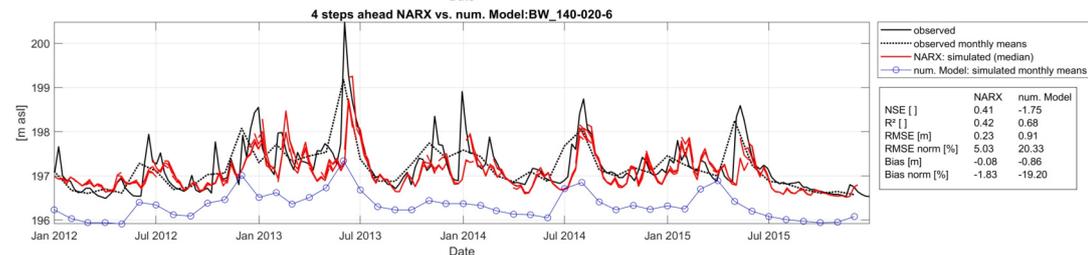
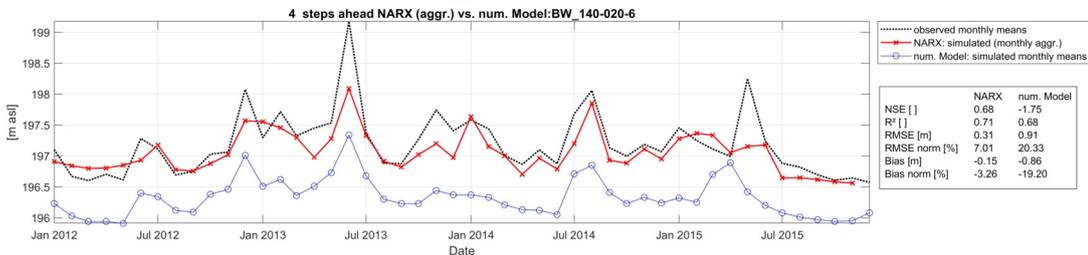
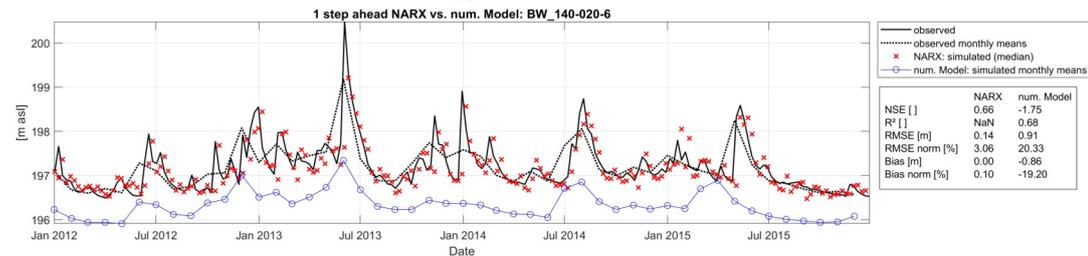
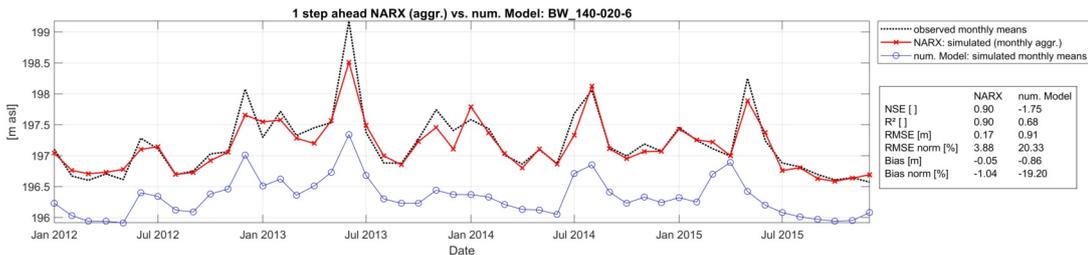
Vergleich Vorhersagegüte an Messstelle BW_1-115-8



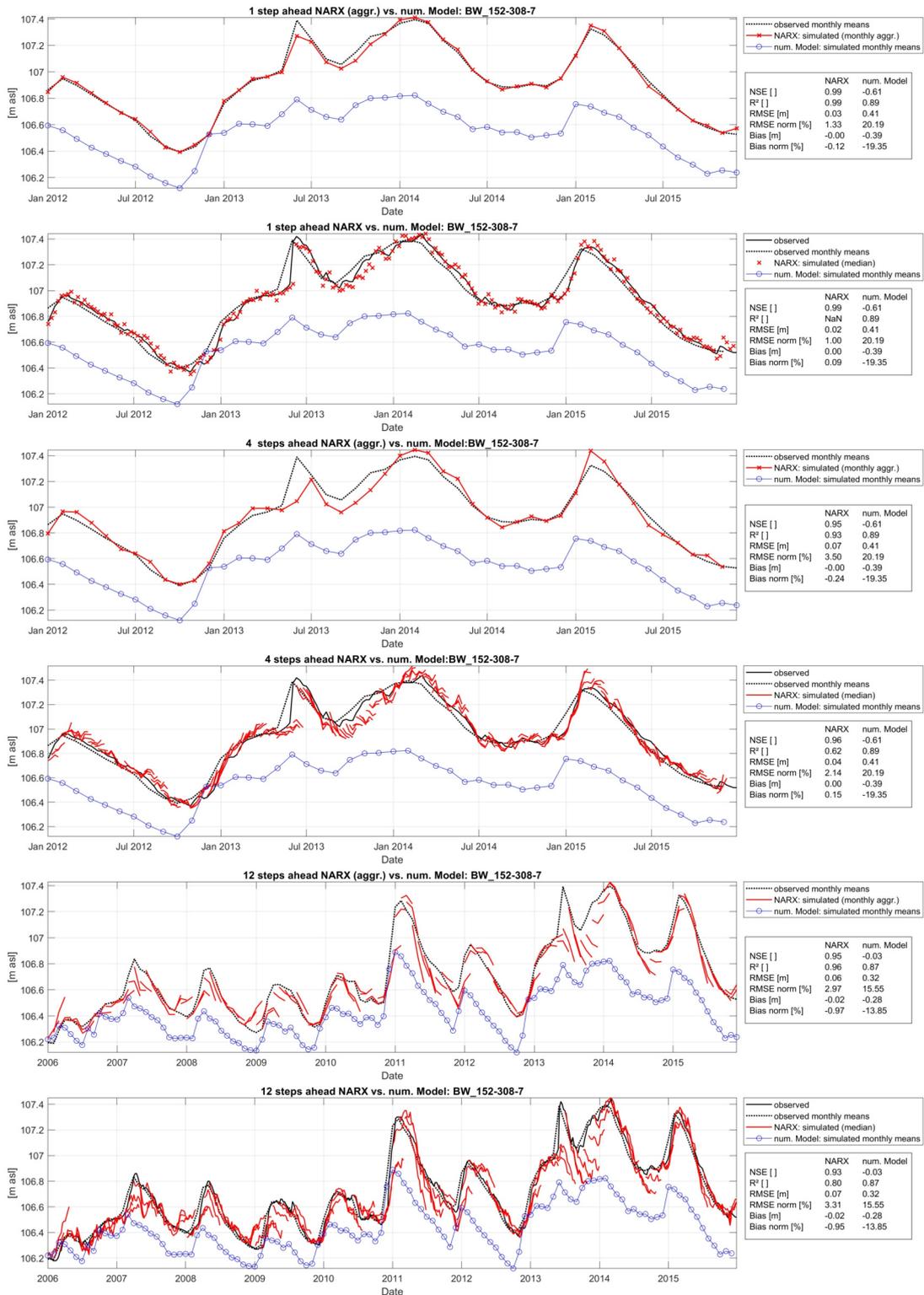
Vergleich Vorhersagegüte an Messstelle BW_104-307-0



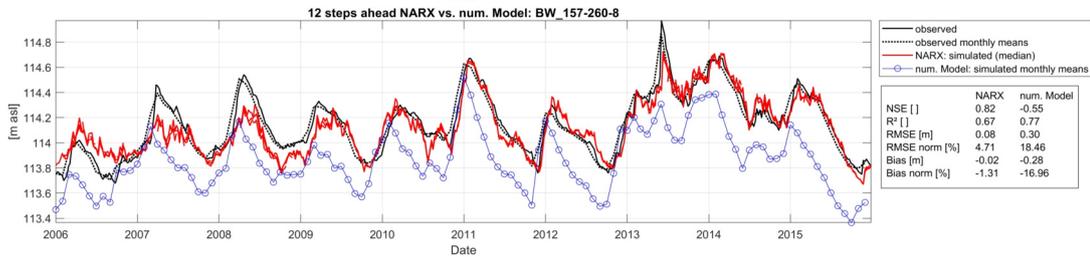
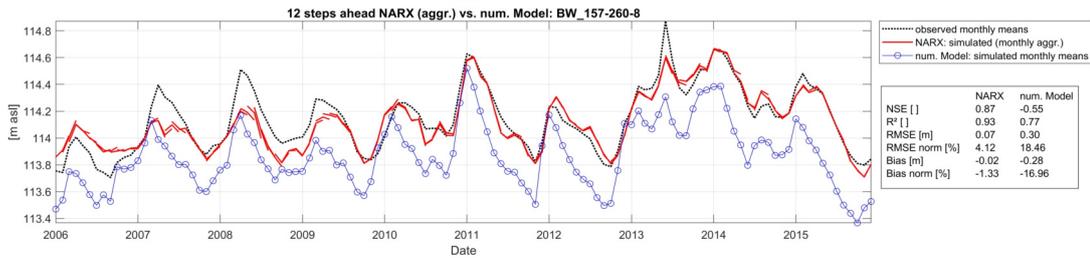
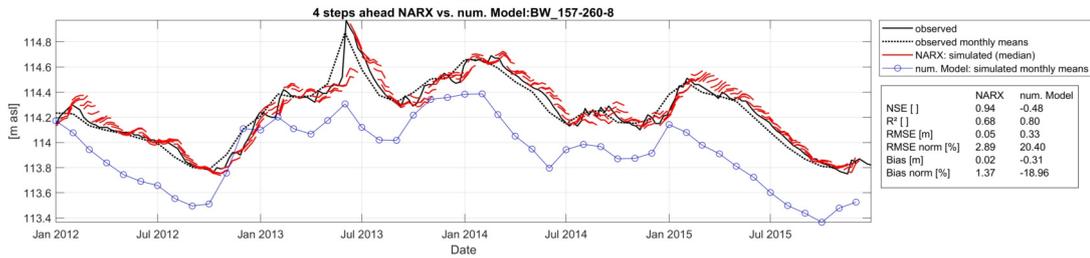
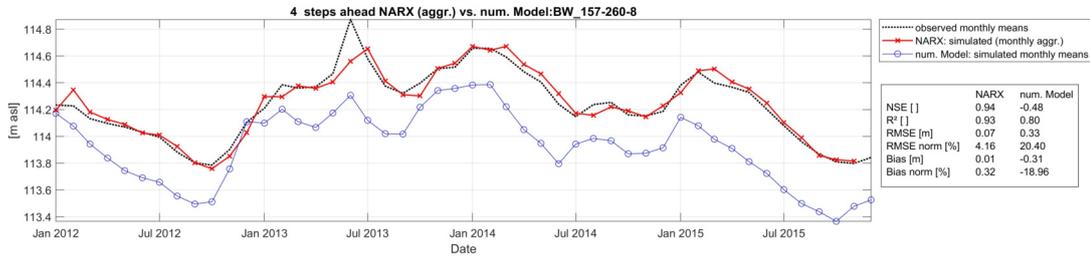
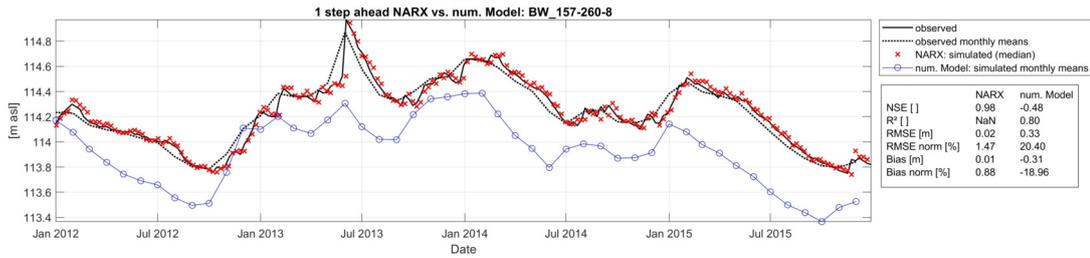
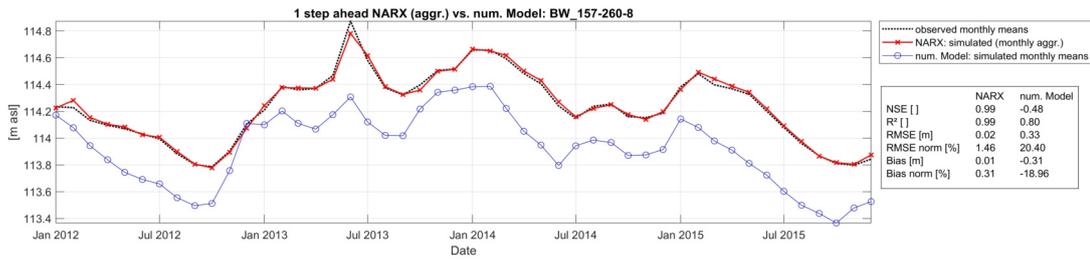
Vergleich Vorhersagegüte an Messstelle BW_126-020-2



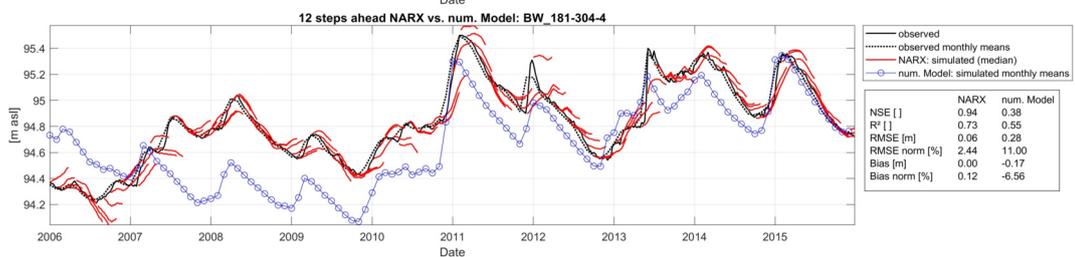
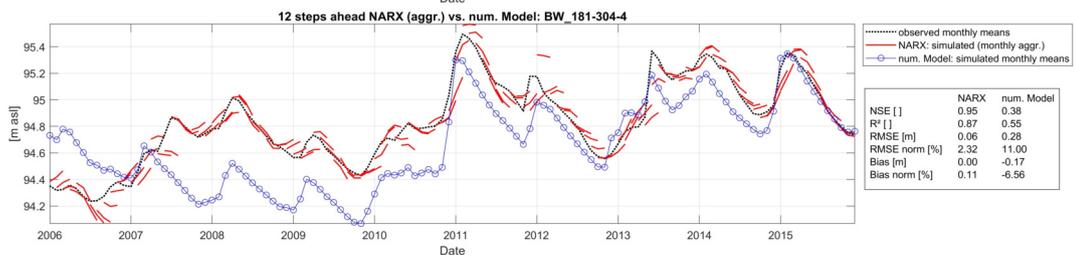
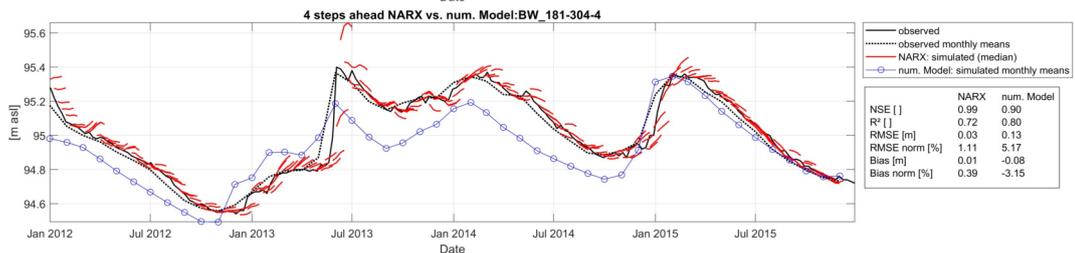
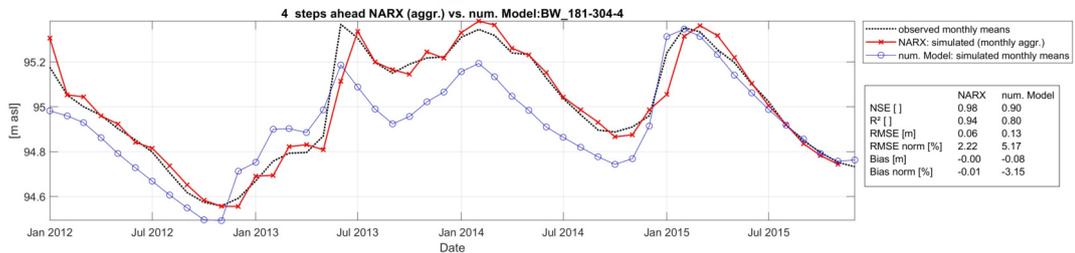
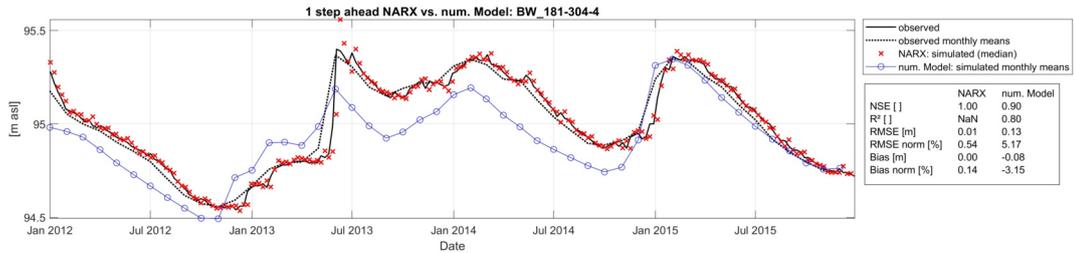
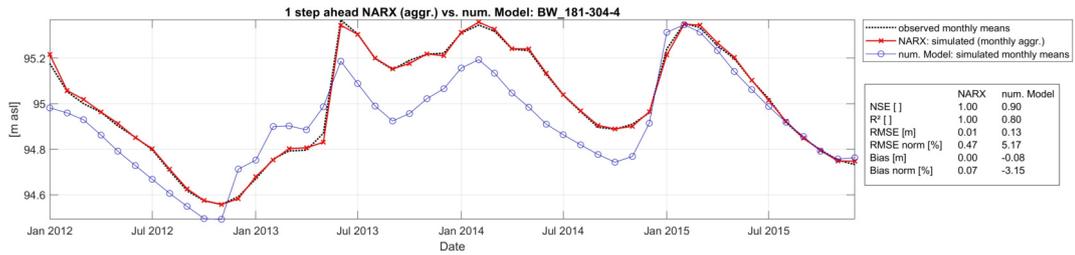
Vergleich Vorhersagegüte an Messstelle BW_140-020-6



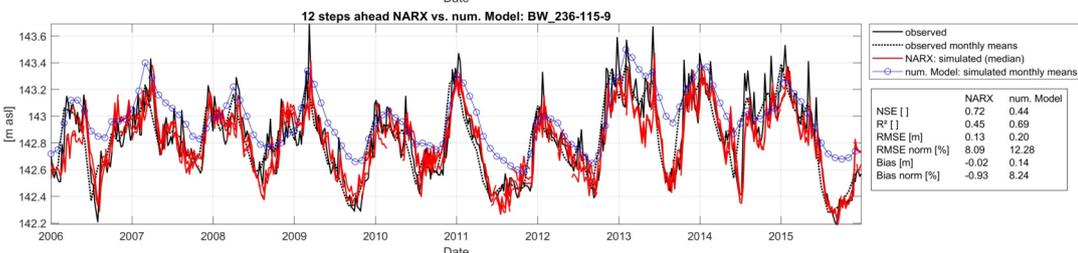
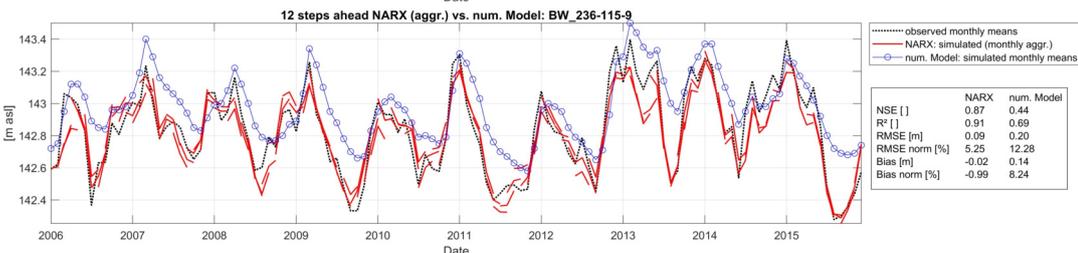
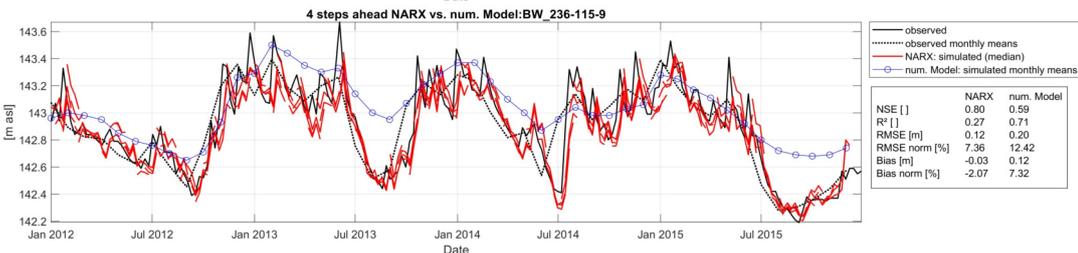
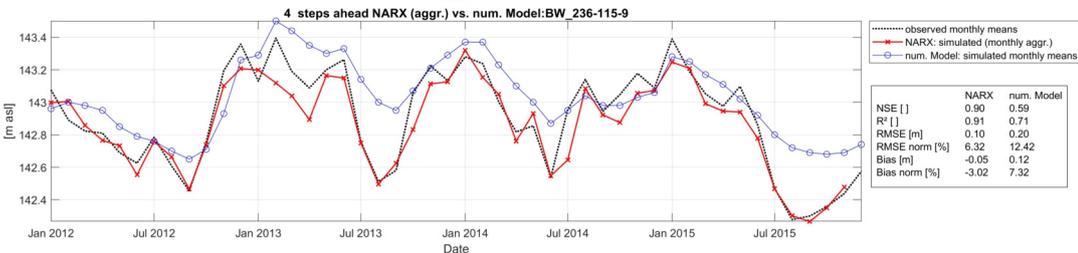
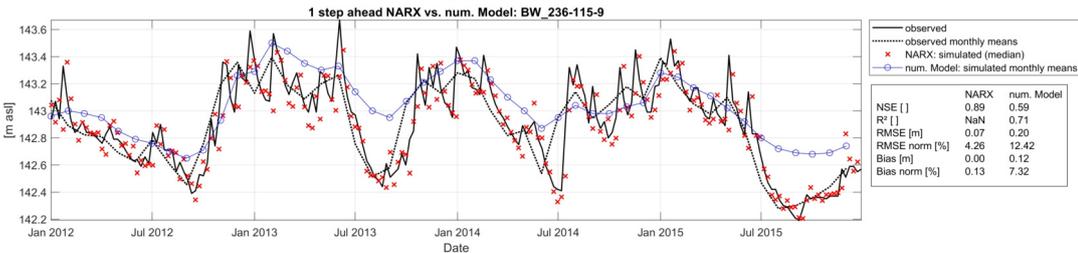
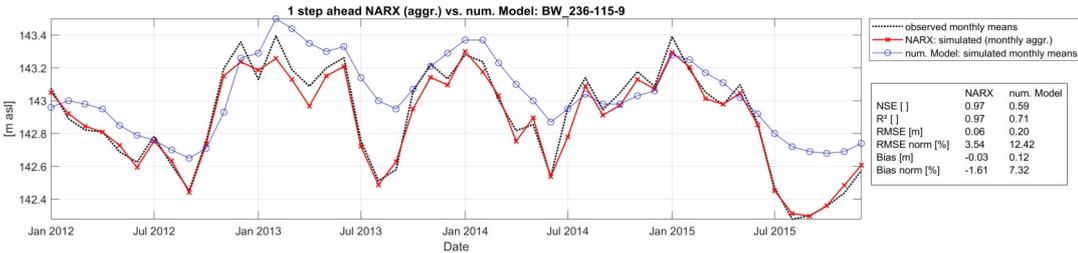
Vergleich Vorhersagegüte an Messstelle BW_152-308-7



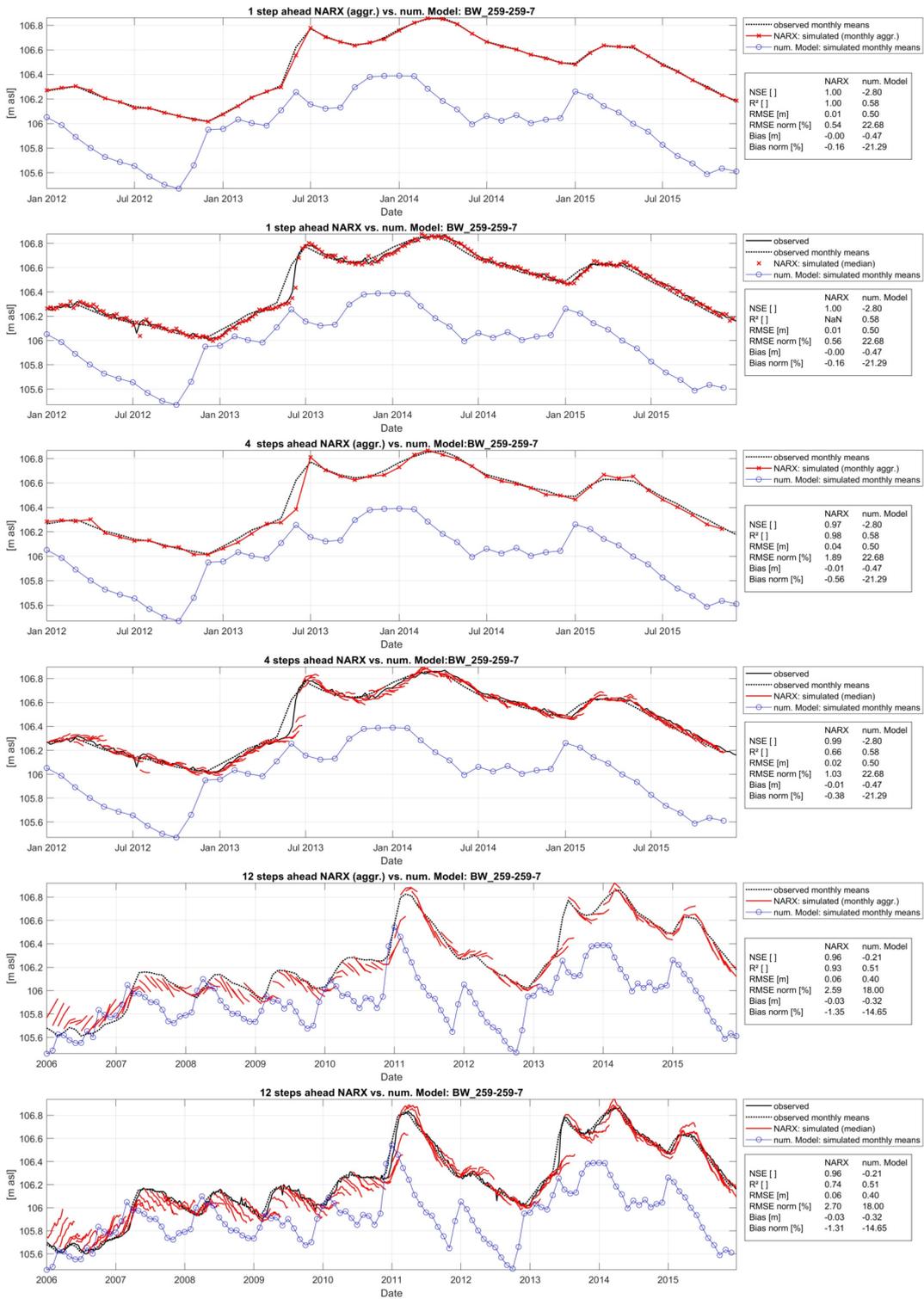
Vergleich Vorhersagegüte an Messstelle BW_157-260-8



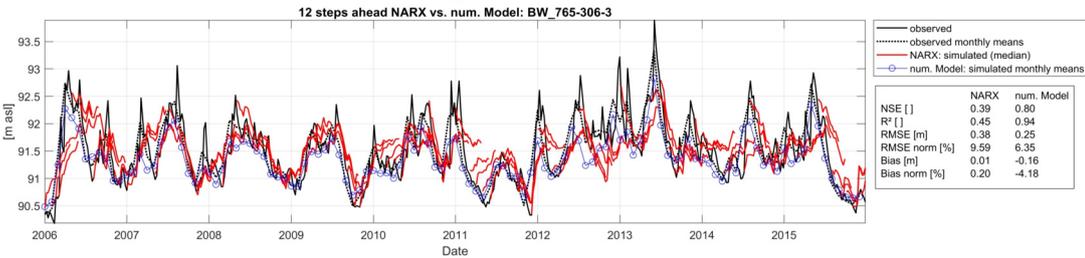
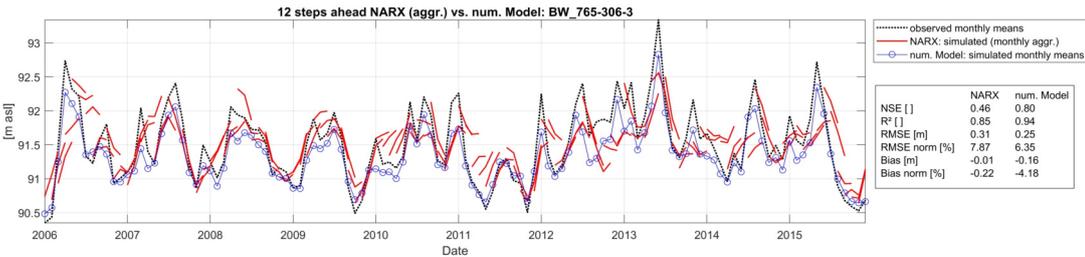
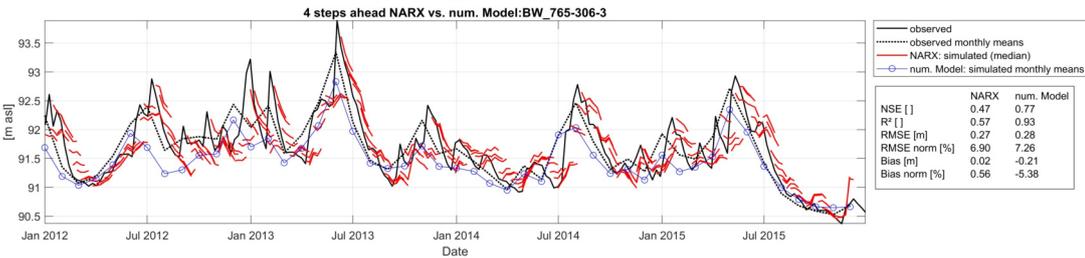
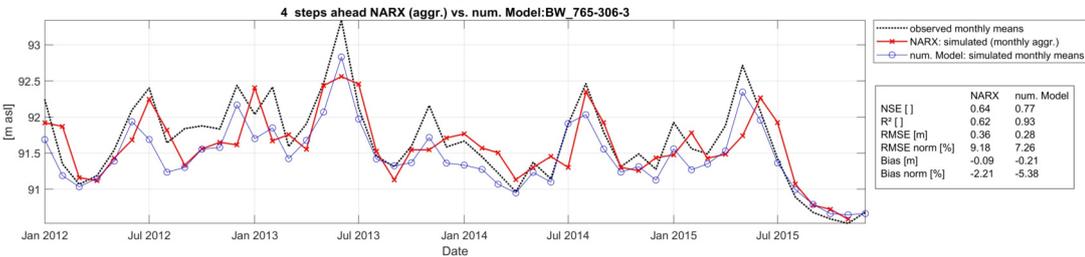
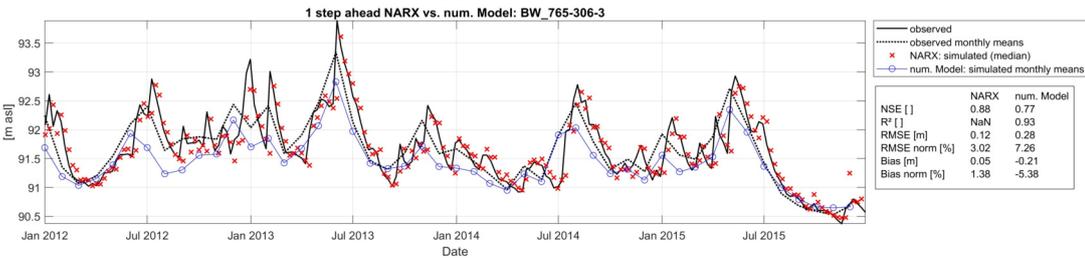
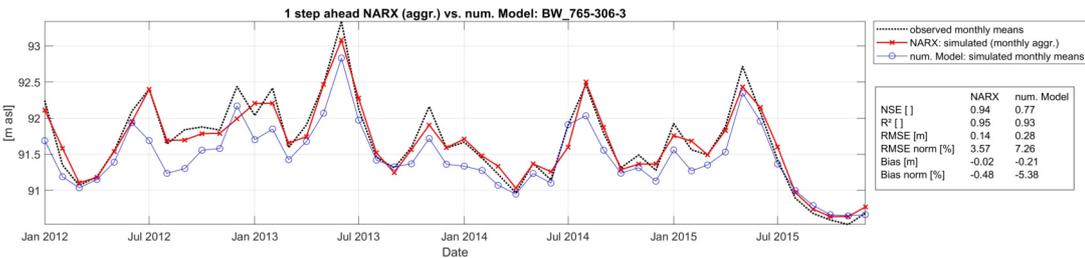
Vergleich Vorhersagegüte an Messstelle BW_181-304-4



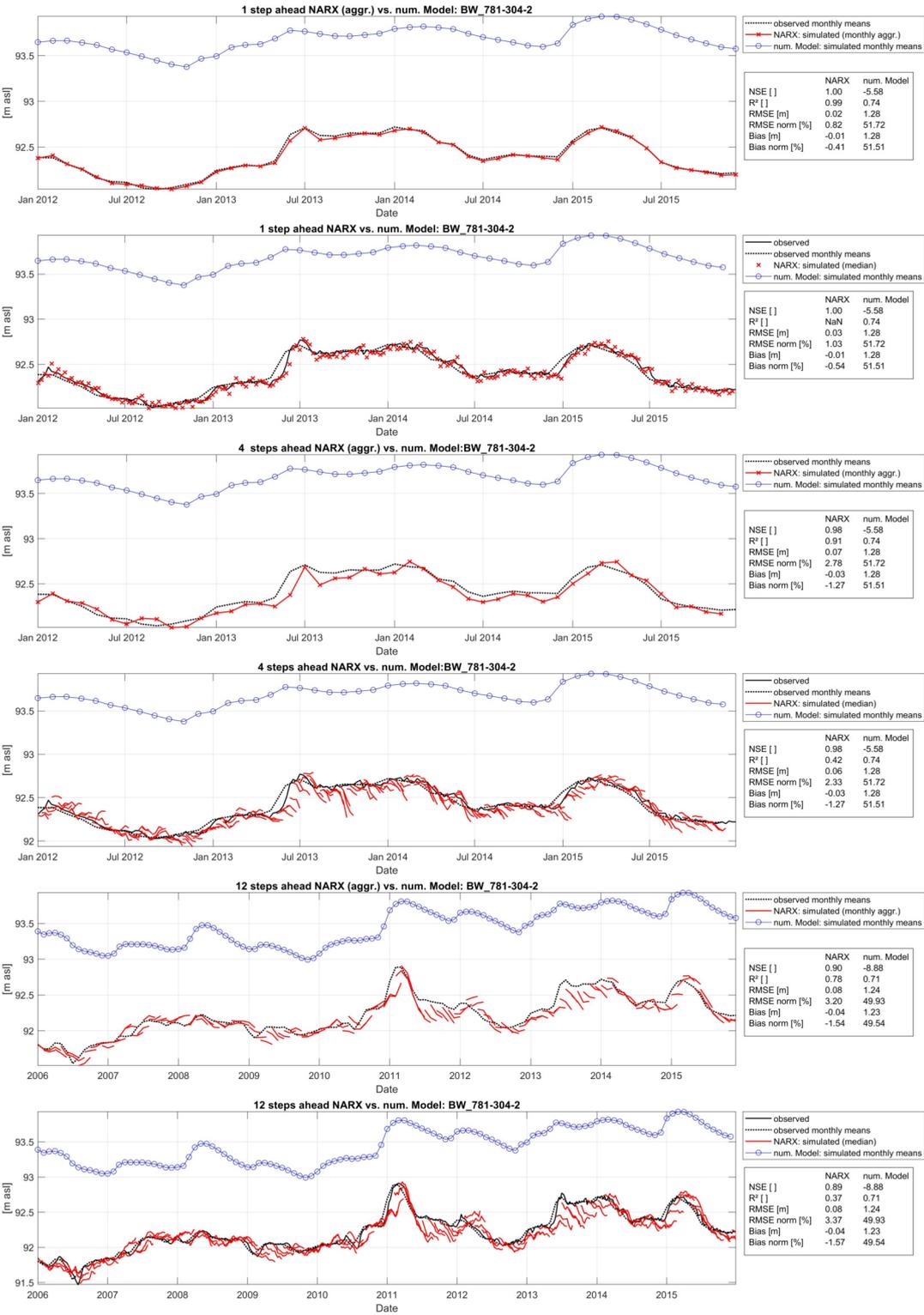
Vergleich Vorhersagegüte an Messstelle BW_236-115-9



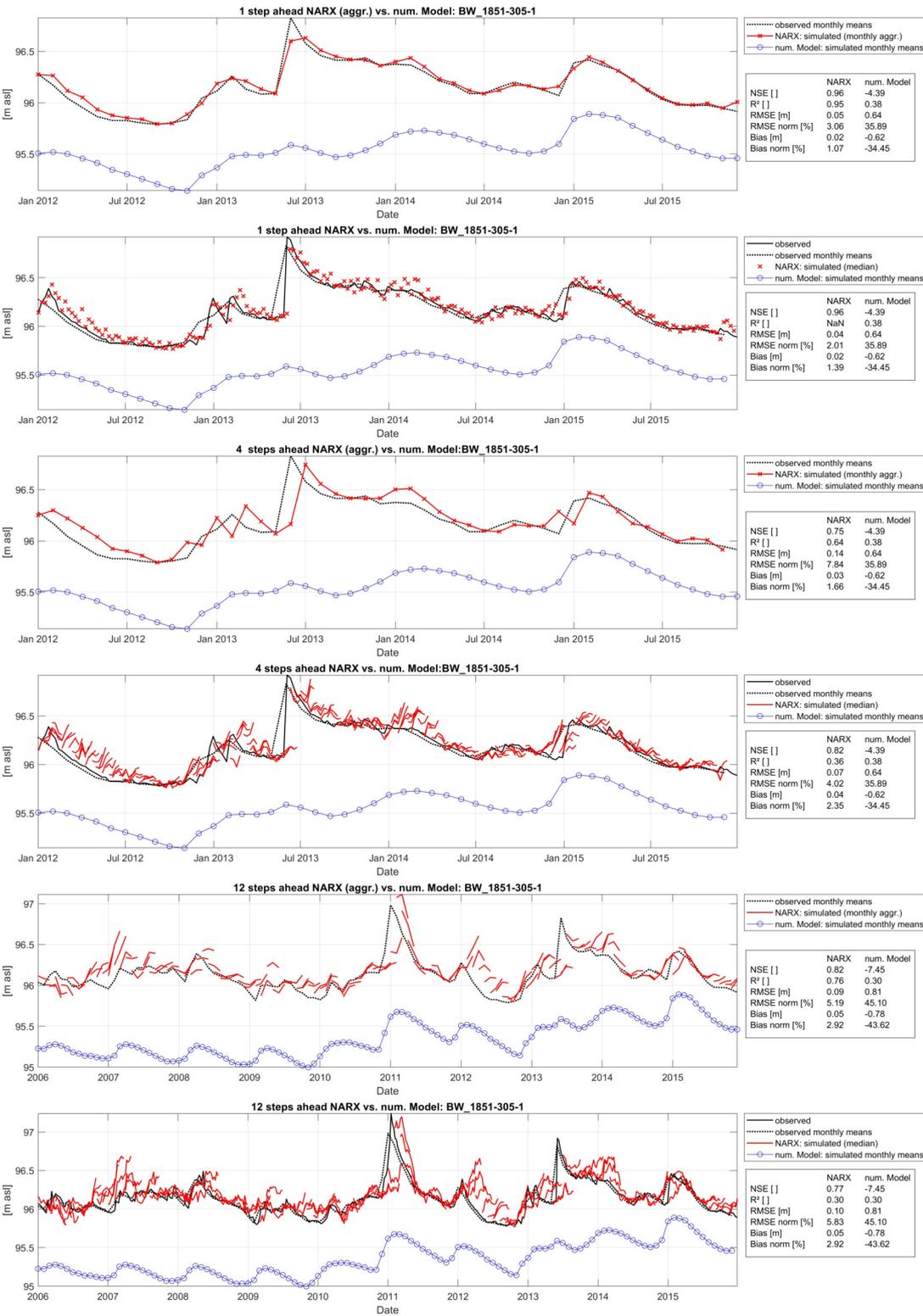
Vergleich Vorhersagegüte an Messtelle BW_259-259-7



Vergleich Vorhersagegüte an Messstelle BW_765-306-3



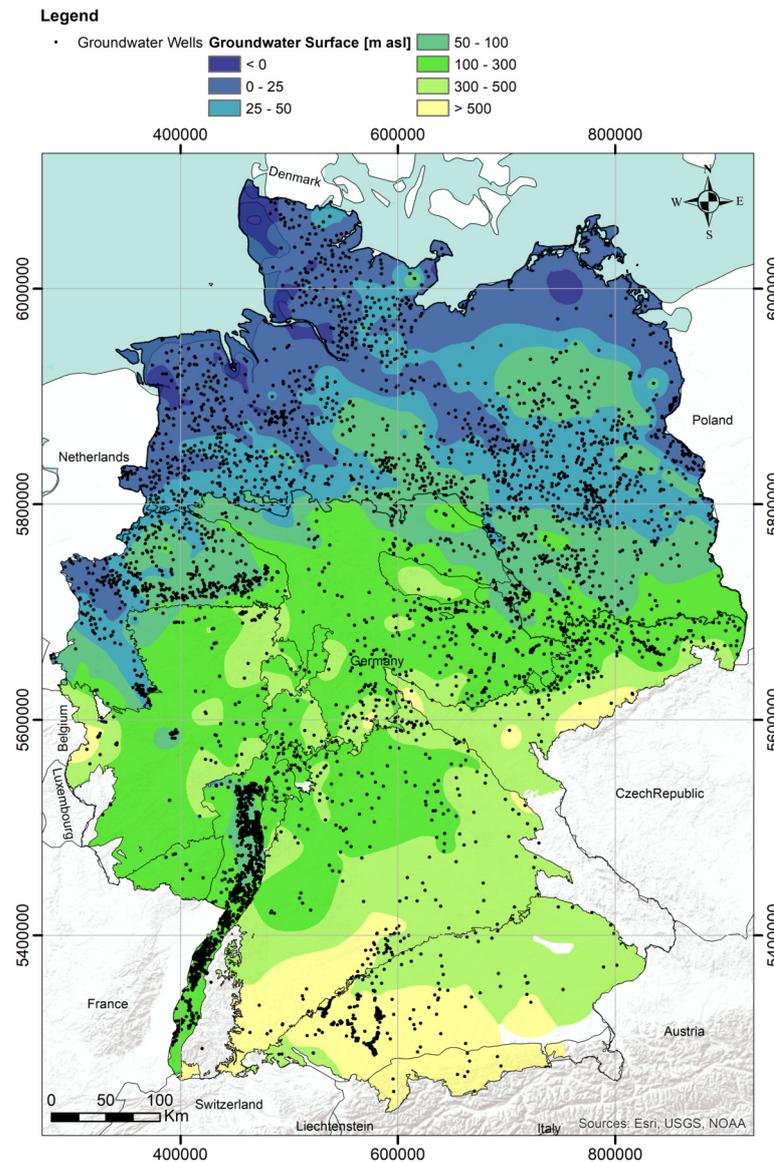
Vergleich Vorhersagegüte an Messstelle BW_781-304-2



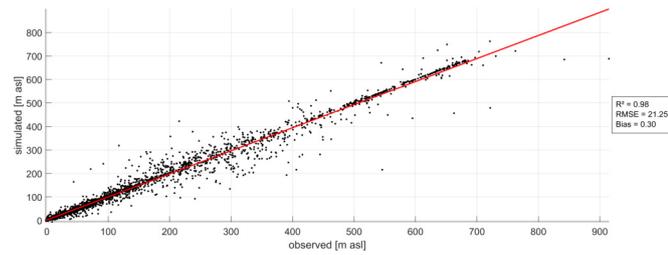
Vergleich Vorhersagegüte an Messstelle BW_1851-305-1

Anhang L - Co-Kriging Ergebnisse

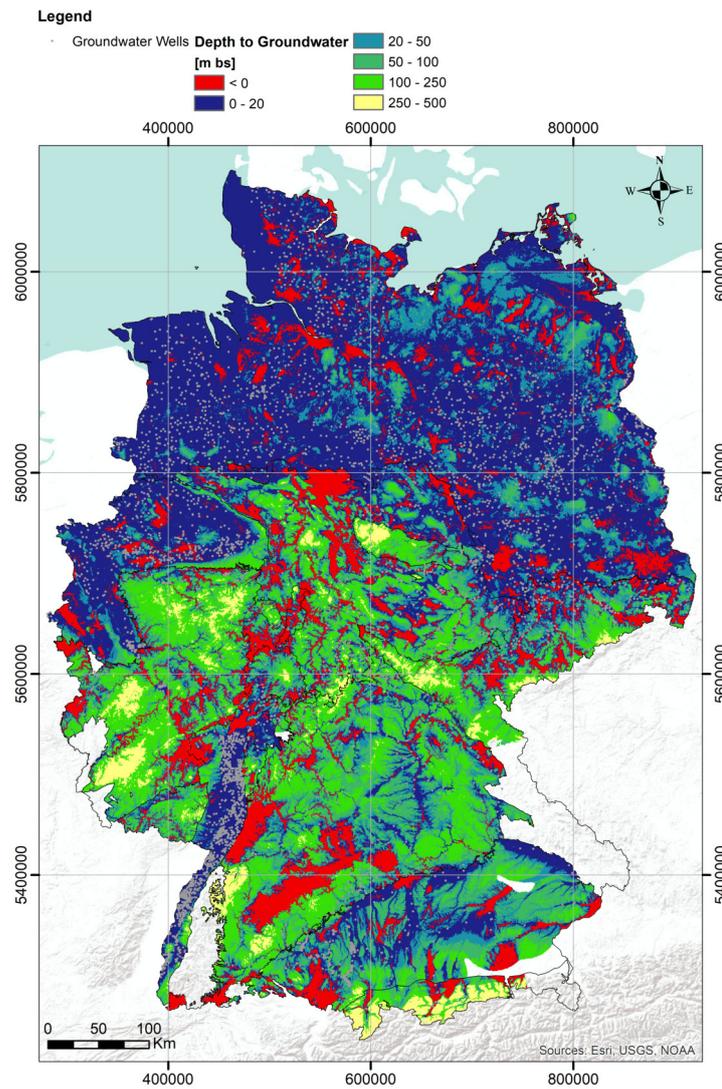
Die im Folgenden dargestellten Ergebnisse beruhen auf einer ungenügenden Datenbasis und stellen damit keine belastbare Grundlage für weitergehende Interpretationen dar. Auf eine solche sollte daher unbedingt verzichtet werden.



Grundwasseroberfläche in mNN, erstellt mittels Co-Kriging, zusammengesetzt aus den Interpolationen der Einzelgebiete. Für leere Flächen konnte wegen unzureichender Daten keine Interpolation erstellt werden.



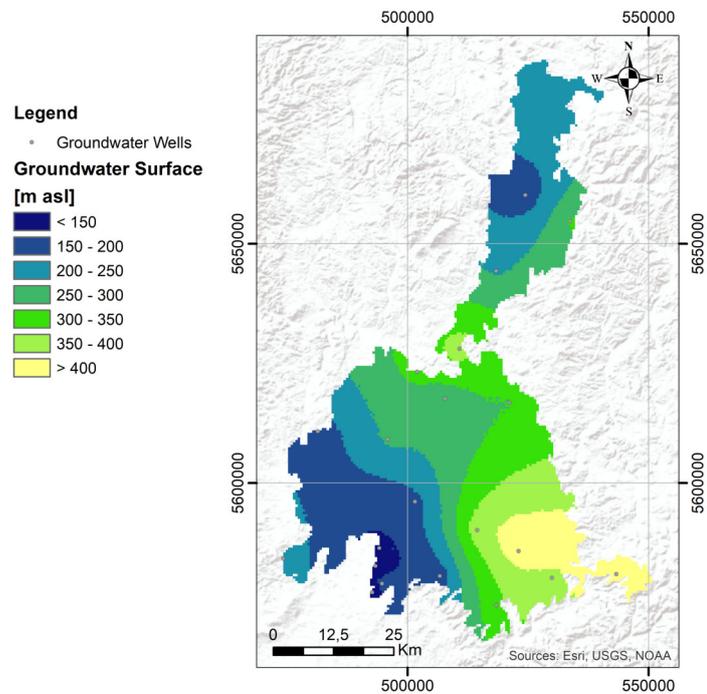
Scatterplot der Kreuzvalidierung für die mit Co-Kriging erstellte Grundwasseroberfläche von Deutschland



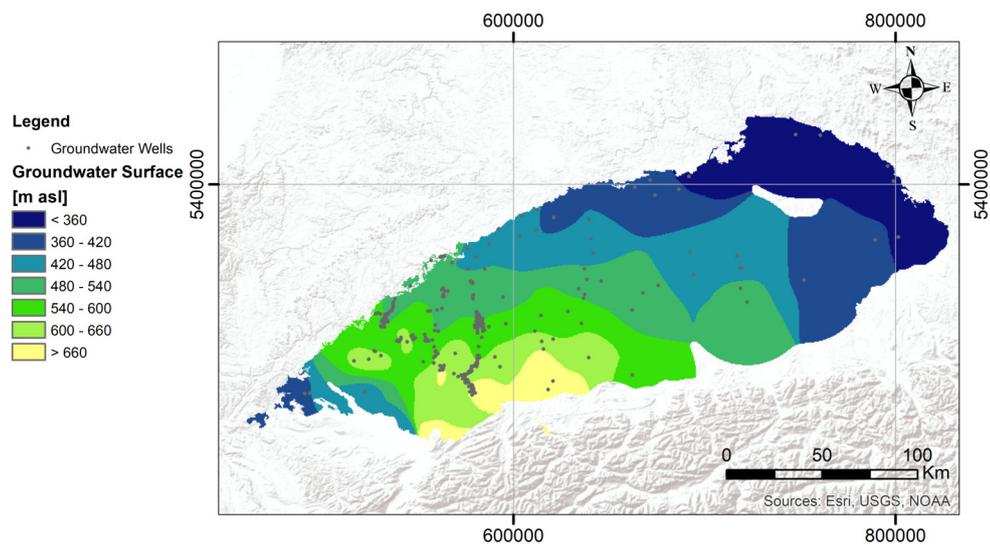
Flurabstand basierend auf SRTM-Höhenmodell und Co-Kriging GW-Oberfläche

Anhang M - Co-Kriging Ergebnisse der Einzelgebiete

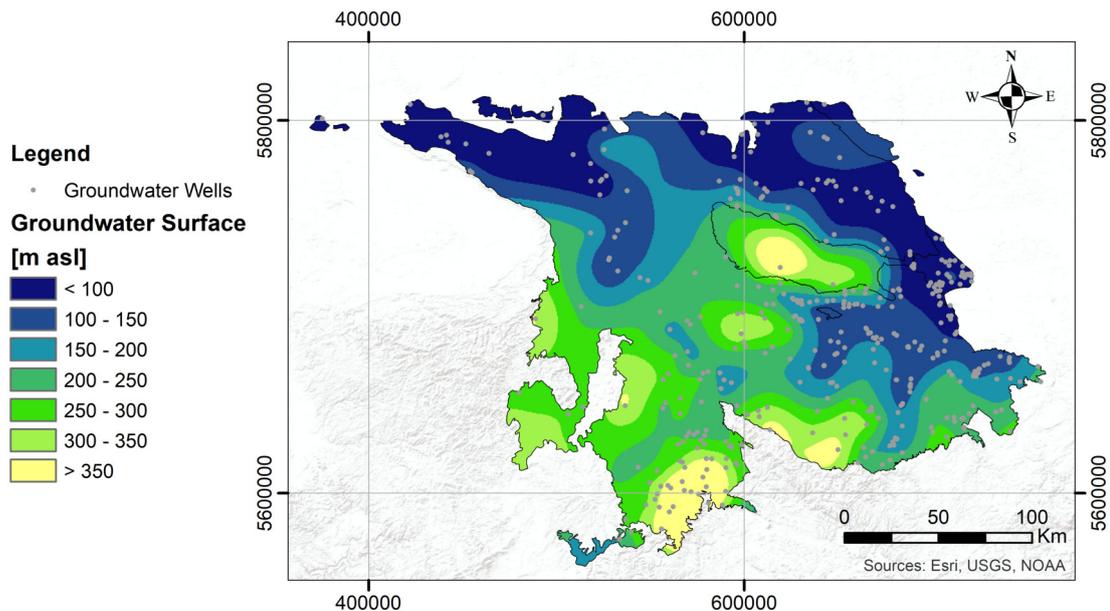
Die im Folgenden dargestellten Ergebnisse beruhen auf einer ungenügenden Datenbasis und stellen damit keine belastbare Grundlage für weitergehende Interpretationen dar. Auf eine solche sollte daher unbedingt verzichtet werden.



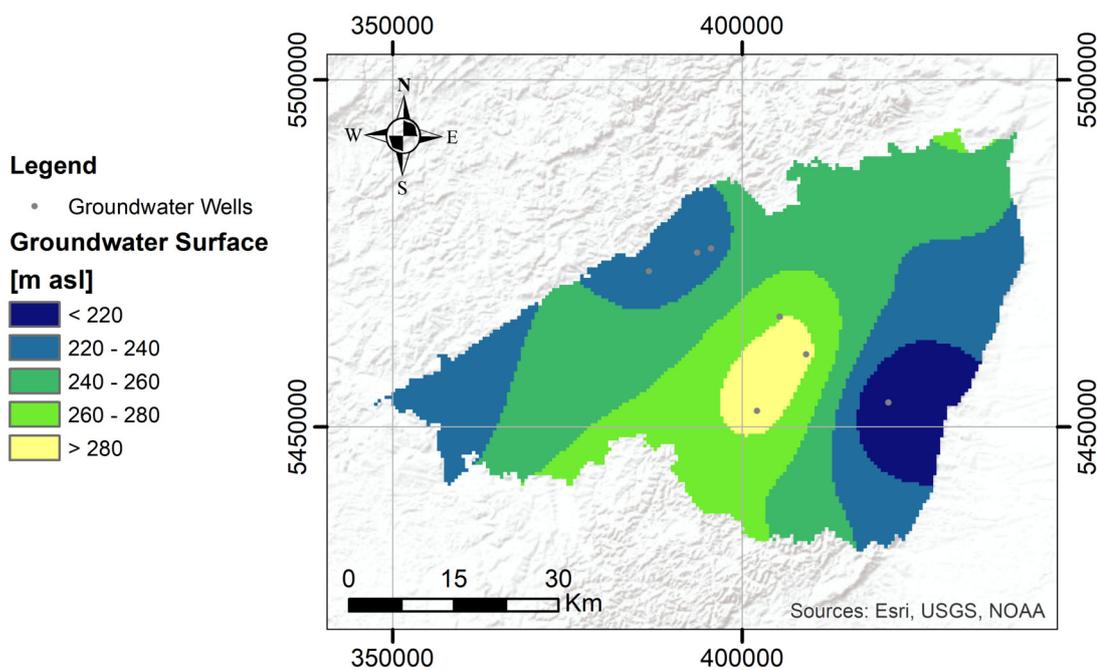
Co-Kriging Grundwasseroberfläche für Raum 33



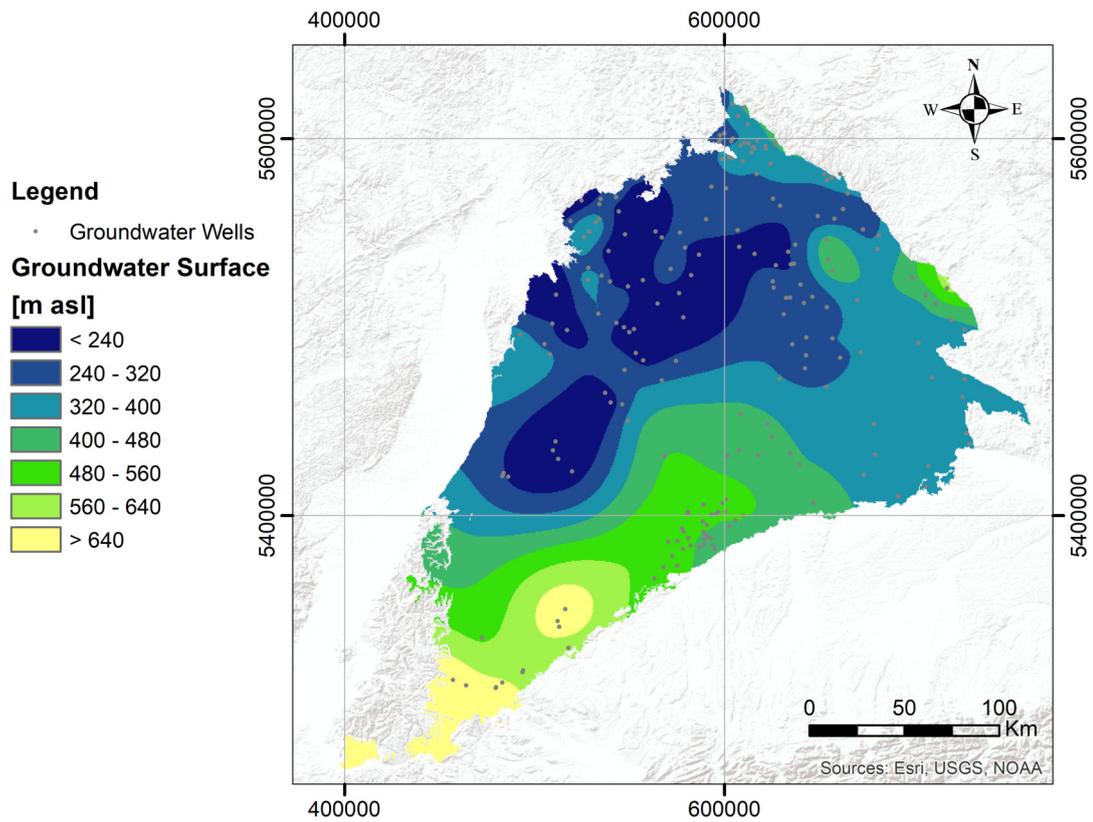
Co-Kriging Grundwasseroberfläche für Großraum 4



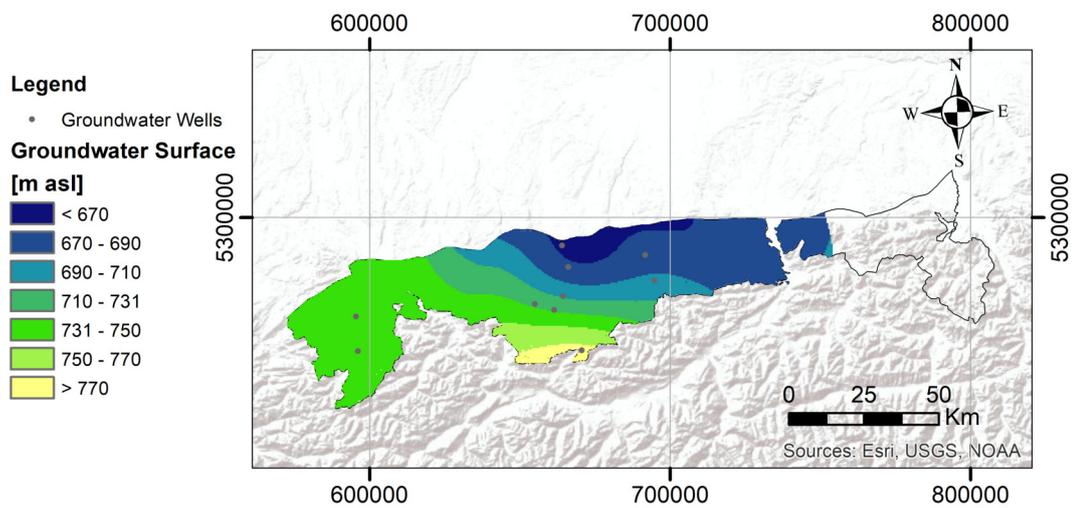
Co-Kriging Grundwasseroberfläche für Großraum 5 und Raum 83



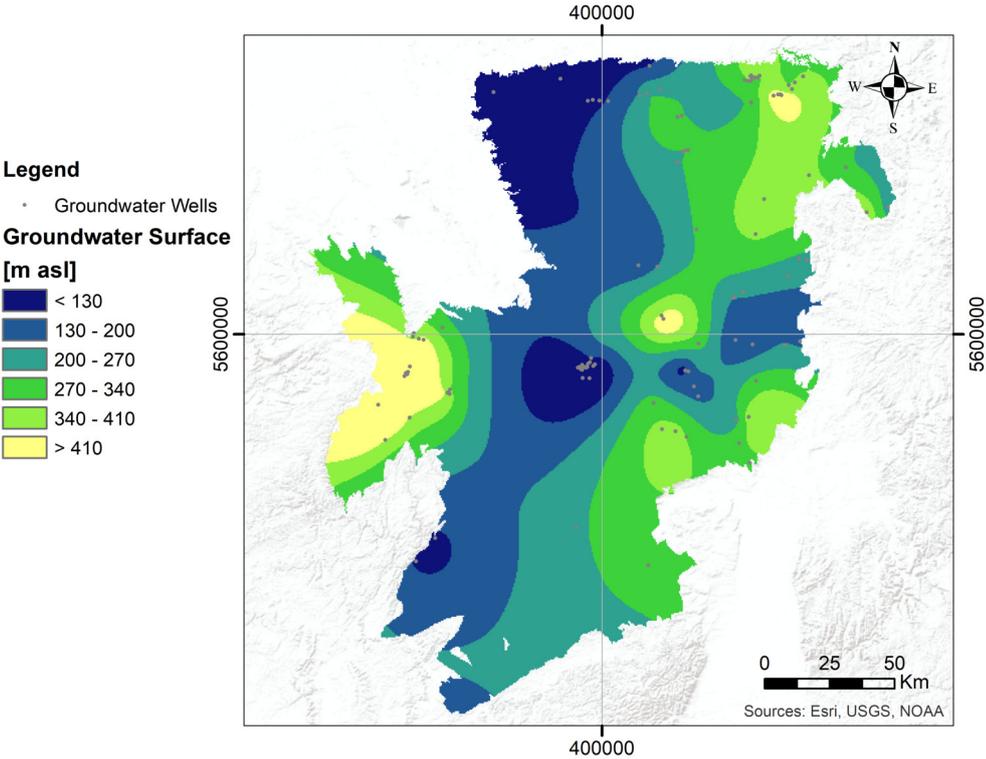
Co-Kriging Grundwasseroberfläche für Raum 61 Süd



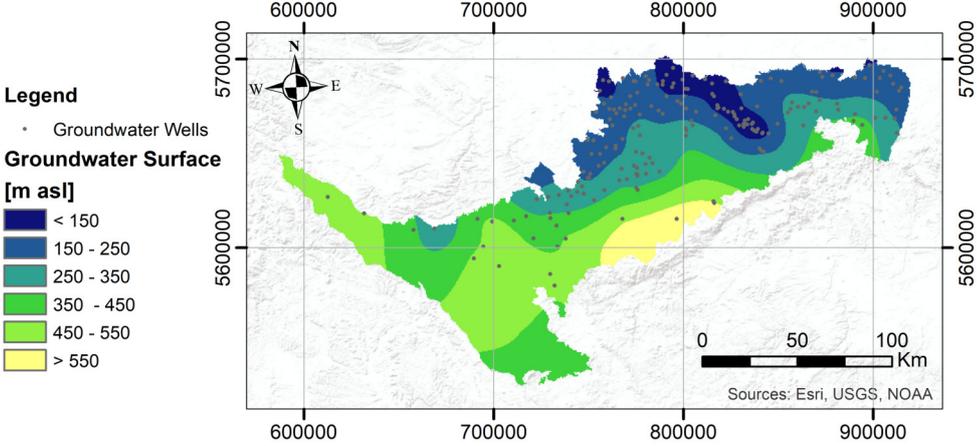
Co-Kriging Grundwasseroberfläche für Grobraum 6 ohne Raum 61



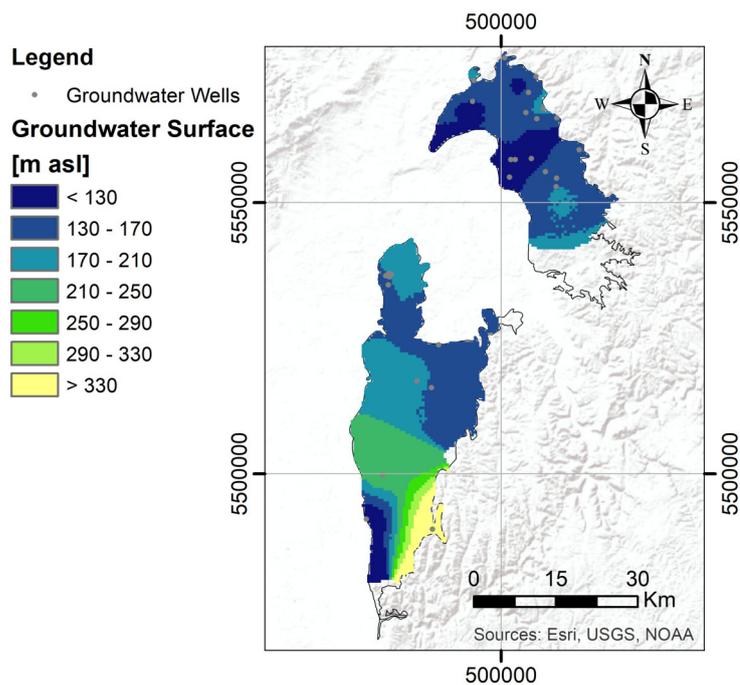
Co-Kriging Grundwasseroberfläche für Grobraum 7



Co-Kriging Grundwasseroberfläche für Großraum 8 ohne Raum 83



Co-Kriging Grundwasseroberfläche für Großraum 9 ohne Raum 95



Co-Kriging Grundwasseroberfläche für Großraum 10 Nord