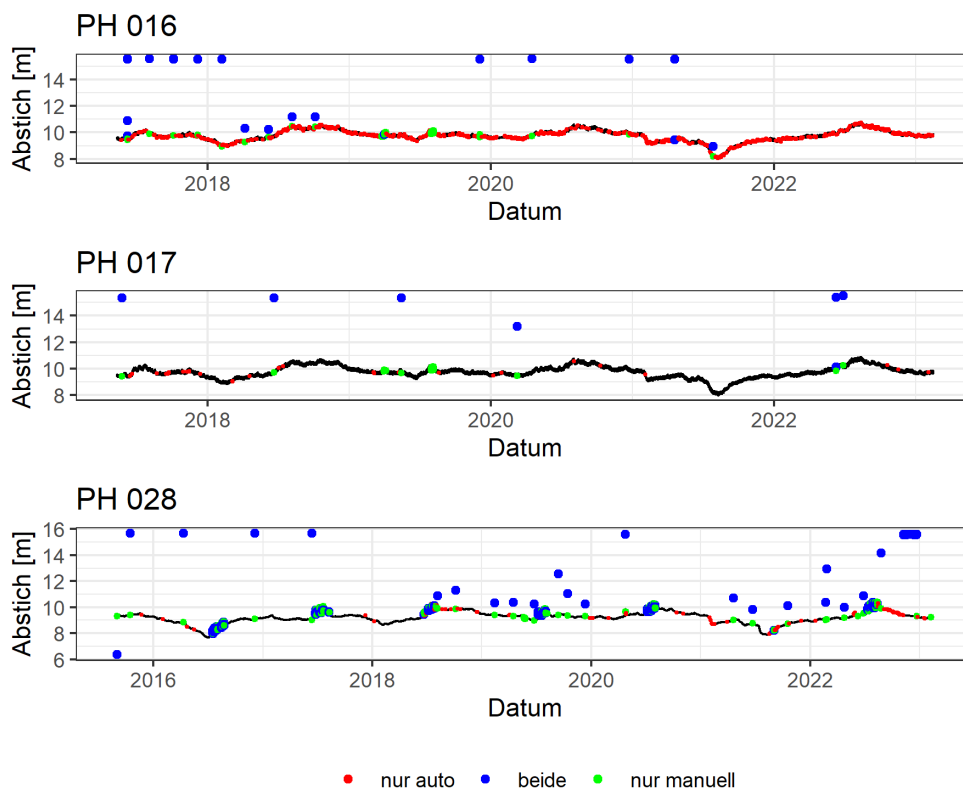


# Entwicklung und Validierung eines Workflows zur automatischen Qualitätskontrolle von Grundwasserzeitreihen

*Kira Zimmermann*



Referent:  
Prof. Dr. Markus Weiler

Korreferent:  
Dr. Jost Hellwig

Freiburg i. Br. 6. September 2023



# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>13</b>
<b>2</b>	<b>Problemstellung und Zielsetzung</b>	<b>17</b>
<b>3</b>	<b>Material, Methoden und Vorgehensweise</b>	<b>19</b>
3.1	Daten und Projektregion . . . . .	19
3.2	Anomalien in Zeitreihen . . . . .	20
3.3	Workflow Entwicklung . . . . .	23
3.3.1	Datenbereitstellung und Anwendungsschnittstelle . . . . .	23
3.3.2	Datenprozessierung und Prüfabfolge . . . . .	24
3.3.3	Testbeschreibung und Parametrisierung . . . . .	26
3.4	Anomalieerkennung mit Testparameterunsicherheit . . . . .	29
3.5	Validierung mit Experten-Qualitätskontrolle und Umweltsystem- kontext . . . . .	30
<b>4</b>	<b>Ergebnisse</b>	<b>32</b>
4.1	Qualitätskontrolle . . . . .	32
4.1.1	Duplikate . . . . .	32
4.1.2	SaQC-Testparameter . . . . .	32
4.1.3	Anomalieerkennung . . . . .	36
4.2	Anomalieerkennung mit Testparameterunsicherheit . . . . .	37
4.3	Vergleich zwischen automatischer und Experten- Qualitätskontrolle . . . . .	39
4.4	Anomalien im Umweltsystem Kontext . . . . .	47
<b>5</b>	<b>Diskussion</b>	<b>51</b>
5.1	Anomalieerkennung mit Testparameterunsicherheit . . . . .	51
5.2	Vergleich zwischen manueller und automatischer Qualitätskontrolle	52
5.3	Grenzen und Stärken des Workflows . . . . .	53
<b>6</b>	<b>Schlussfolgerungen</b>	<b>55</b>
<b>A</b>	<b>Appendix</b>	<b>62</b>

# Abbildungsverzeichnis

1	Projektregion Staufener Bucht mit den Hauptflüssen, dem Grabensystem, den Messstellen und Entnahmebrunnen . . . . .	19
2	Grundaufbau der Workflow-Struktur . . . . .	24
3	QC-Workflow unterteilt in Pre-Processing, Basic-Tests, Advanced-Test und Post-Processing . . . . .	25
4	Anteil an Duplikaten pro Station . . . . .	32
5	Zugrundeliegende Verteilungen für die Berechnung der Parameter <i>min</i> und <i>max</i> (Range (2)), <i>thresh</i> (Offset), <i>window</i> (Constant) und <i>thresh</i> (Jump), mit den berechneten und genutzten Parametern markiert. . . . .	33
6	Boxplots der ermittelten Parameter aller Stationen mit farbig markierten Expertengrenzen. . . . .	35
7	Prozentualer Anteil der als Anomalie markierten Punkte an den Duplikat-bereinigten Daten . . . . .	36
8	Anteil aller Tests an der Gesamtanzahl der Flags pro Station . . .	36
9	Anteil aller Tests an der Gesamtanzahl der Flags zusammengefasst auf alle Stationen und unterteilt in die Stationen mit 10-, bzw 60-minütiger Messfrequenz. Die Kategorie "Mix"bezieht sich auf Punkte, welche von mehr als einem Test geflagged wurden. . . . .	37
10	Geflaggte und flagbereinigte Zeitreihe der Stationen PH 016, PH 017 und PH 028 . . . . .	38
11	Ergebnisse der Monte-Carlo-Analyse pro Parameter, mit der Unsicherheit (Prozentuale Abweichung von dem im Workflow genutzten Parameter) auf der x-Achse und dem Verhältnis der Flaganzahl der aktuellen Simulation zu der Flaganzahl der Simulation mit den meisten Flags auf der y-Achse. Der übergeordnete lineare Trend ist in rot dargestellt. . . . .	40
12	Anomalien, welche durch mindestens einen Test bei allen MC-Simulationen als Anomalie markiert wurden (jeweils oben) und die maximale Anzahl an Flags bei 100 Monte-Carlo-Simulationen durch einen Test (jeweils unten). Bei mehreren Tests mit der gleichen Anzahl an Anomaliemarkierungen wird folgendermaßen priorisiert: Range (2), Offset, LOF, Constant, Jump. . . . .	41
13	Expertenflags für alle drei duplikatbereinigten Validierungszeitreihen, farbig markiert je nach Anzahl der Experten, die den Punkt markiert haben . . . . .	42
14	Validierungszeitreihen mit den Datenpunkten farbig markiert, welche nur durch das automatische, nur durch das manuelle oder durch beide Verfahren erkannt wurden . . . . .	43
15	Test-Ursprung bei Punkten, welche nur durch die automatischen Tests bzw. durch diese und die Experten erkannt wurden. . . . .	45
16	Teilzeitreihen der Validierungsstationen mit Anomaliemarkierungen der Experten und des Workflows zur Analyse einzelner Events.	47
17	Zeitreihe der Rohdaten der Station PH 028 mit Markierungen durch die Experten und den manuellen Workflow und Niederschlagssummen der Monate April bis Oktober für die Jahre 2016 bis 2022	48

18	Manuelle Abstichmessungen im Vergleich zu den automatischen Loggermessungen für die Stationen PH 006 und PH 083 . . . . .	50
19	Verteilung der Abstichwerte . . . . .	63
20	Verteilung der Differenzen zwischen zwei nacheinanderfolgenden Werten . . . . .	64
21	Verteilung der Perioden mit unverändertem Grundwasserstand . .	65
22	Verteilung der Differenzen zwischen zwei nacheinanderfolgenden Tagesmittelwerten . . . . .	66
23	Geflaggte und flagbereinigte Zeitreihe der Station PH 001 . . . . .	67
24	Geflaggte und flagbereinigte Zeitreihe der Station PH 005 . . . . .	67
25	Geflaggte und flagbereinigte Zeitreihe der Station PH 006 . . . . .	68
26	Geflaggte und flagbereinigte Zeitreihe der Station PH 015 . . . . .	68
27	Geflaggte und flagbereinigte Zeitreihe der Station PH 018 . . . . .	69
28	Geflaggte und flagbereinigte Zeitreihe der Station PH 019 . . . . .	69
29	Geflaggte und flagbereinigte Zeitreihe der Station PH 023 . . . . .	70
30	Geflaggte und flagbereinigte Zeitreihe der Station PH 025 . . . . .	70
31	Geflaggte und flagbereinigte Zeitreihe der Station PH 027 . . . . .	71
32	Geflaggte und flagbereinigte Zeitreihe der Station PH 028 . . . . .	71
33	Geflaggte und flagbereinigte Zeitreihe der Station PH 034 . . . . .	72
34	Geflaggte und flagbereinigte Zeitreihe der Station PH 036 . . . . .	72
35	Geflaggte und flagbereinigte Zeitreihe der Station PH 056 . . . . .	73
36	Geflaggte und flagbereinigte Zeitreihe der Station PH 059 . . . . .	73
37	Geflaggte und flagbereinigte Zeitreihe der Station PH 078 . . . . .	74
38	Geflaggte und flagbereinigte Zeitreihe der Station PH 083 . . . . .	74
39	Geflaggte und flagbereinigte Zeitreihe der Station PH 105 . . . . .	75
40	Geflaggte und flagbereinigte Zeitreihe der Station PH 119 . . . . .	75

## Tabellenverzeichnis

1	Metadaten der Stationen, der darin installierten Sensoren zur Grundwasseraufzeichnung und Distanzen zu den nächstgelegenen Entnahmehbrunnen . . . . .	21
2	Erläuterung der Anomlien, die im Workflow berücksichtigt werden	22
3	Beschreibung der Testparameter und ihre Ermittlung . . . . .	27
4	Beispielhafte Darstellung einer Confusionmatrix . . . . .	30
5	Der Varianzunterschied der statistisch ermittelten Parameter unterteilt in Stationen mit 10 <i>min</i> und 60 <i>min</i> zeitlicher Auflösung .	34
6	Anzahl der Stationen (von insgesamt 20), welche bei Änderung eines Parameters einen signifikanten Trend aufweisen. . . . .	39
7	Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 016 . . . . .	44
8	Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 017 . . . . .	44
9	Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 028 . . . . .	44
10	Multiclass-Confusionmatrix der Validierungszeitreihe PH 016 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden. . . . .	46
11	Multiclass-Confusionmatrix der Validierungszeitreihe PH 017 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden. . . . .	46
12	Multiclass-Confusionmatrix der Validierungszeitreihe PH 028 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden. . . . .	46
13	Die berechneten Parameter für die Tests Range (2), Offset, Constant, Jump . . . . .	62
14	Die Kennzahlen Recall, Präzesion, Spezifität und F-Score abhängig von der Definition der wahren Klasse. Ein Datenpunkt wird hier als Anomalie gewertet, wenn 1. mindestens ein Experte ihn markiert oder 2. die Experten mehrheitlich für eine Anomalie stimmen. . .	62

## Formelverzeichnis

Symbol	Einheit	Beschreibung
$max$	[m]	max Parameter des Range-Tests
$min$	[m]	min Parameter des Range-Tests
$\mu$	–	Mittelwert
$n$	[m]	thresh Parameter des LOF-Tests
$p$	–	Parameter
$\sigma$	–	Standardabweichung
$thresh_{jump}$	[m]	thresh Parameter des Jump-Tests
$thresh_{LOF}$	[m]	thresh Parameter des LOF-Tests
$thresh_{offset}$	[m]	thresh Parameter des Offset-Tests
$tolerance$	[m]	tolerance Parameter des Offset-Tests
$tolerance$	[m]	tolerance Parameter des Offset-Tests
$window_{constant}$	[Minuten]	window Parameter des Constant-Tests
$window_{jump}$	[Minuten]	window Parameter des Jump-Tests
$x$	–	Anzahl der Standardabweichungen

## Abkürzungsverzeichniss

<b>DWA</b>	Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall
<b>FB</b>	False Bad
<b>FG</b>	False Good
<b>IDC</b>	International Data Corporation
<b>ISO</b>	International Organization for Standardisation
<b>KI</b>	Künstliche Intelligenz
<b>ML</b>	Maschinelles Lernen
<b>QC</b>	Qualitätskontrolle
<b>QA</b>	Qualitätssicherung
<b>SaQC</b>	System of automatic Quality Control
<b>TG</b>	True Good
<b>TB</b>	True Bad
<b>WMO</b>	World Meteorological Organization





# Zusammenfassung

In der fortschreitenden Digitalisierung der Wasserwirtschaft liefern Sensordaten viele Daten, deren Qualitätskontrolle (QC) manuell oder automatisch erfolgt. Im Vergleich ist die manuelle QC zeitintensiv und wegen subjektiven Entscheidungen häufig nicht reproduzierbar, ist aber wegen Unkenntnissen über die Unsicherheiten in der automatischen QC gegenwärtig das Standard-QC-Verfahren. Diese Masterarbeit entwickelt einen Workflow zur automatischen Qualitätskontrolle von Grundwasserzeitreihen unter Einbeziehung von 20 Sensor-Stationen der BadenoVA AG, um (1) die Robustheit der automatischen Anomalieerkennung gegenüber Unsicherheiten in der statistischen Testparametrisierung und (2) den Unterschied zwischen manueller und automatischer Qualitätskontrolle zu untersuchen. Inbegriffen sind 5 QC-Tests (Range, Offset, Constant, LOF, Jump) des Systems of Automatic Quality Control (SaQC), die jeweils für 17 Sensor-Stationen statistisch parametrisiert werden. Die Validierung erfolgt an 3 Sensor-Stationen mit 3 manuellen Expertenkontrollen, und die Analyse zur Robustheit der Anomalieerkennung durch 100 Monte-Carlo-Simulationen im Unsicherheitsbereich von  $\pm 20\%$  des ermittelten Testparameters.

Allgemein unterscheidet sich die Anzahl der Anomalieerkennungen, wie auch die berechneten Testparameter, zwischen den jeweils pro Station vorkommenden Messintervallen 10min (8 Stationen) und 60min (12 Stationen). Insgesamt zeigen 80% der Stationen im Unsicherheitsbereich der Testparameter *max* (Range), *window* (Constant) und *thresh* (LOF) eine signifikante Änderung der Anomalieerkennung. Die relative Anomalieerkennung (3,2%) ist im Mittel bei 60min (4,8%) höher als bei 10min (1,6%). Bei 60min erkennt der Constant-Test (90%) die meisten Anomalien, wohingegen bei 10min Offset (48,9%) und LOF (54,2%) dominieren. Bei 10min wird der statistisch ermittelte window (Constant) durch einen im Workflow gewählten Expertengrenzwert korrigiert, sodass Constant bei 10min (3,3%) signifikant weniger Anomalien als bei 60min (90%) ohne Grenzwertkorrektur erkennt. Durchschnittlich werden 12,4% der markierten Anomalien durch mehrere Tests erkannt.

Allgemein ist die Erkennung der automatischen QC von den Datenpunkte ohne anormales Verhalten im Vergleich zur manuellen QC hoch (Mittelwert Spezifität = 0,99). Die Präzision ist jedoch niedrig (Mittelwert = 0,32), da die manuelle QC insgesamt mehr Anomalien erkennt, wodurch die Anzahl der *FALSE BAD* hoch ist. Zu berücksichtigen ist, dass alle drei Experten nur in 10% der Anomalieerkennung übereinstimmen und 89% der Anomalieerkennungen auf einen Experten zurückzuführen sind. Werden nur manuelle Anomalieerkennungen genommen, bei denen die Experten mehrheitlich übereinstimmen, werden 98% dieser Datenpunkte auch durch die automatische QC erkannt. Während in dieser Betrachtung nur noch wenig Datenpunkte als *FALSE GOOD* kategorisiert werden, bleiben noch immer viele Datenpunkte in der Kategorie *FALSE BAD* (Mittelwert Präzision = 0,27). Weder die manuelle noch die automatische Qualitätskontrolle konnten Zeiträume, in denen ein Einfluss durch nahegelegene Grundwasserentnahme vorliegt, eindeutig identifizieren.

Zusammenfassend zeigen sowohl die manuelle als auch die automatische QC relevante Unsicherheiten. Da eine weitere Optimierung der automatischen Testparametrisierung mehr manuelle QC-Ergebnisse verlangt, muss im ersten Schritt die

gegenwärtige manuelle QC standardisiert und umfanglich dokumentiert werden. Auch im Sinne der nationalen Wasserstrategie, um bis 2030 ein qualitätsgesichertes Echtzeit-Monitoring im Grundwasserressourcenmanagement zu etablieren.

**Keywords:** Anomalien, automatisierte Qualitätskontrolle (QC), Grundwasserzeitreihen, Monte-Carlo-Simulation, Workflowentwicklung

## Danksagung

Zu Beginn möchte ich mich herzlich beim Team von HydrosConsult bedanken. Ein besonderer Dank geht an Karuna Jutglar. Deine unermüdlichen Anregungen und der stetige Zuspruch während der gesamten Forschungs- und Schreibphase dieser Arbeit waren für mich eine große Stütze. Der Austausch mit dir war nicht nur hilfreich, sondern hat meine Arbeit auch maßgeblich bereichert. Stephen Schrempp hat mir die Möglichkeit gegeben, diese Arbeit durchzuführen und war immer mit einem offenen Ohr für mich da - dafür bin ich sehr dankbar. Mein Dank geht auch an Joscha Schellhorn. Deine Expertise und dein sorgfältiger Blick bei der Validierung der Zeitreihen waren unverzichtbar.

An der Universität möchte ich mich besonders bei Prof. Dr. Markus Weiler bedanken. Unsere regelmäßigen Treffen und Diskussionen waren nicht nur lehrreich, sondern haben mich auch dazu angeregt, stets über den Tellerrand zu blicken und neue Perspektiven einzunehmen. Ihre konstruktive Kritik und die Diskussionen über meine Methodik waren entscheidend für die Qualität meiner Arbeit.

Ein herzlicher Dank geht auch an die Badenova AG, ohne deren Datenbasis diese Arbeit nicht möglich gewesen wäre. Simon Brenner, deine Hilfsbereitschaft und Schnelligkeit bei der Bereitstellung von Informationen waren für mich von unschätzbarem Wert. Dein Engagement und deine Unterstützung haben viele meiner Herausforderungen erleichtert.

Zuletzt möchte ich mich beim UFZ, insbesondere bei Lennart Schmidt und David Schäfer, bedanken. Das gesamte Entwicklungsteam des SaQC war während dieser Zeit eine zuverlässige Säule. Ihr habt nicht nur die Grundlage für meine Arbeit gelegt, sondern mit eurem Fachwissen und eurer Expertise auch maßgeblich zu ihrem Erfolg beigetragen.



# 1 Einleitung

Die Aufzeichnung des Grundwasserstandes liefert Zeitreihen, die es ermöglichen, das Verständnis der Grundwasserdynamik zu vertiefen, Ressourcen effektiv zu verwalten und informierte Entscheidungen zum Schutz des Grundwassers zu treffen (Bannick et al., 2008). Bekesi et al. (2008), Mirzavand und Ghazavi (2014) und Halder et al. (2020) analysieren Zeitreihen, die an verschiedenen Messorten (zwischen 14-653 Messstationen) über einen Zeitraum von 20 bis ca. 40 Jahren aufgenommen wurden. Diese Grundwasserzeitreihen dienen nicht nur als Schlüsselkomponente für zuverlässige Vorhersagen (Mirzavand und Ghazavi, 2014, Lin et al., 2022), sondern sind auch für ein effektives Grundwassermanagement unerlässlich, um Trends zu erkennen und informierte Entscheidungen zu treffen (Halder et al., 2020). Dies spielt wiederum eine zentrale Rolle bei Aufgaben wie der Trinkwasserversorgung und der landwirtschaftlichen Bewässerung (Bekesi et al., 2008). Insbesondere in dicht besiedelten oder besonders trockenen Gebieten ist die Quantifizierung der Grundwasserressourcen essenziell, um die Wasserversorgung sowohl aktuell als auch zukünftig sicherzustellen (Zamani et al., 2022, Yang et al., 2016). Diese Wichtigkeit erkennen auch die EU-Wasserrahmenrichtlinien an, indem sie die Mitgliedsstaaten seit dem Jahr 2000 dazu verpflichten, den mengenmäßigen Zustand des Grundwassers zu überwachen (WRRL, 2000). Diese Forderung wird auch von der nationalen Wasserstrategie vom 15.03.2023 unterstützt, welche ein Echtzeitmonitoring von Grundwasser fordert (Bundesministerium für Umwelt, Nukleare Sicherheit und Verbraucherschutz, 2023). Vor dem Hintergrund des Klimawandels und einer wachsenden Bevölkerung wird die Bedeutung verlässlicher Grundwasserdaten auch in Zukunft weiter zunehmen (Patle et al., 2015).

Mit dem Aufkommen von Technologien zur effizienteren Datenerhebung entsteht eine stetig wachsende Menge an Daten. Laut der Digital Universe Prognose der International Data Corporation (IDC) (Gantz und Reinsel, 2012) wurden bis 2020 40 Zettabyte Daten erzeugt. Diese Entwicklung zeigt auch der folgende Vergleich zwischen vier wissenschaftlichen Studien vor 1990 und drei Studien nach 2019. Jackson (1974), Law (1974), Adamowski und Hamory (1983) und Nevulis et al. (1989) verwenden Datensätze, die zwischen einer und 84 Messstellen umfassen, wobei die Messintervalle von täglich bis monatlich variieren. Im Gegensatz dazu zeigen neuere Studien von Rinderer et al. (2019), Haaf et al. (2020) und Erdbrügger et al. (2022), eine deutliche Intensivierung der Datenerfassung: Messungen werden mit einer Frequenz zwischen fünf Minuten und 24 Stunden dokumentiert, wobei umfangreichere Datensätze mit 24 bis 341 Stationen verwendet werden. Während Law (1974) die Hydrographen für seine Studie noch vom Magnetband ausliest, sprechen Jeong et al. (2021) von Datenbanken, auf denen große Mengen an Daten gespeichert werden. In Lettland stieg die Anzahl an Beobachtungsbrunnen zwischen 1959 und 2019 von 15 auf 612 an, begleitet von einer steigenden Frequenz und Digitalisierung der Messungen (Retike et al., 2022). Diese Entwicklung unterstreicht einen klaren Trend zur Erfassung detaillierterer und umfangreicherer Daten in der Grundwasserforschung und dem Ressourcenmanagement, welcher die Fortschritte in der Überwachungstechnik und die zunehmende Bedeutung umfassender, hochauflösender Datensätze für hydrologische Studien widerspiegelt.

Probleme in Sensordaten sind in wissenschaftlichen Studien ein immer wiederkehrendes Thema (Porter et al., 2012). Durch die Automatisierung der Datenerhebung wird diese einerseits effizienter, jedoch sind die erhobenen Daten auch anfällig für Fehler (SADC-GMI, 2019). Softwaredefekte, Übertragungsfehler, fehlende Daten oder menschliche Fehler können die Datenqualität beeinflussen, was unbehandelt zu Folgefehlern in der Weiterverarbeitung führen kann (Campbell et al., 2013). In zahlreichen Studien ist die Erkennung und Korrektur solcher Anomalien der Schritt vor der weiteren Verarbeitung oder Analyse (Jeong et al., 2021; Bakker und Schaars, 2019).

Dabei können Anomalien in die Hauptgruppen Punkt- und Teilsequenzanomalien (Erhan et al., 2020) und weiter in Ausreißer (unwahrscheinlich und unmöglich), Sprünge und konstant bleibende Werte unterteilt werden (Panagopoulos et al., 2021; Jeong et al., 2021). Weitere Anomalien sind stetige Verschiebungen (Drift) und Rauschen in den Daten (Erhan et al., 2020). Ein weiterer Schritt in der Vorverarbeitung kann die Untersuchung der Zeitstempel auf Duplikate sein (Faybishenko et al., 2022). Allgemein richtet sich die Datenqualität und der daraus folgende Umgang mit den Daten immer nach den Anforderungen und der Datennutzung (International Organization for Standardization, 2023). So erwähnt Patle et al. (2015) kein Pre-Processing, während Erdbrügger et al. (2022) die Daten vor der weiteren Verarbeitung auf Ausreißer und Sprünge untersuchen, welche sie mit Wartungsarbeiten und Schneeschmelze in Verbindung bringen. Die Entscheidung über den genauen Umfang der Qualitätskontrolle (QC) der Daten wird aktuell für jedes Paper und jede Studie aufs Neue gefällt und unterliegt der persönlichen Einschätzung der Wissenschaftler und Experten (Sturtevant et al., 2021).

Richtlinien: In den letzten Jahren wurden neue Richtlinien und Leitfäden in den Bereichen der Wasserwirtschaft, Meteorologie und den Umweltwissenschaften veröffentlicht, welche sich mit dem Datenmanagement und der Datenqualität beschäftigen (World Meteorological Organization, 2021; International Organization for Standardization, 2023, Woolf et al., 2023; Jousma et al., 2006). Darin werden Ansätze vorgestellt, durch die die Abfolge bestimmter Prozesse den Umgang mit Messdaten vereinheitlichen soll (World Meteorological Organization, 2021, International Organization for Standardization, 2023). Es kann zwischen der Qualitätssicherung (QA) und der Qualitätskontrolle (QC) unterschieden werden. Bei der QA wird das gesamte System der Datenerhebung betrachtet und sichergestellt, dass die produzierten Daten die für die weitere Verarbeitung geforderte Qualität besitzen. Die QC analysiert und bearbeitet spezifisch die bereits erhobenen Daten, um die Datenqualität zu optimieren und die Integrität für unterschiedliche Plattformen zu gewährleisten (Sturtevant et al., 2021). Da der in dieser Arbeit vorgestellte Workflow innerhalb der QC angesiedelt ist, wird die QA im Weiteren nicht besprochen. Neben Formatierungsvorschlägen werden in allen Richtlinien und Leitfäden Prozesse und verschiedene Tests beschrieben, welche die Zeitreihen auf Fehler und Unstimmigkeiten untersuchen. Dabei sollen Anomalien in den Daten erkannt und markiert werden, um diese in der weiteren Datennutzung zu berücksichtigen. Es bleibt jedoch offen, wie genau dieser Prozess ablaufen soll. Die ISO erklärt, dass die spezifischen Methoden zur Vorverarbeitung von Daten entsprechend den Datencharakteristika und Anwendungsszenarien definiert werden

sollten (International Organization for Standardization, 2023). Einerseits werden in allen Richt- und Leitlinien die Vorteile von automatisierten Abläufen hervorgehoben, gleichzeitig wird betont, dass der Blick eines geschulten Hydrologen bei der Überprüfung der Markierungen weiterhin eine große Rolle spielt (World Meteorological Organization, 2021; Woolf et al., 2023; Clemens-Meyer et al., 2021). Für den Fachbereich der Siedlungs- und Ingenieurhydrologie werden auch konkretere Prozesse und Testabläufe gefordert: Die DWA wird 2023 den Gelbprint DWA-A-181 veröffentlichen, welcher eine hohe Datenqualität von Messdaten von Entwässerungssystemen zum aktuellen Standard erklärt (Hollenberg et al., 2011). Clemens-Meyer et al. (2021) zeigen für Daten aus Abwassersystemen mögliche Tests und diskutieren deren Stärken und Limitationen. Solche konkreten und allgemein anerkannten Regelwerke können ein gemeinsames Verständnis für die Erwartungen an die Datenqualität schaffen und als Leitfäden für die Implementierung von Qualitätsmanagementmethoden in der Praxis dienen. Eine wesentliche Weiterentwicklung im Bereich der Qualitätskontrolle war die Betonung der standardisierten Dokumentation des Prüfprozesses durch Metadaten und der Speicherung der unversehrten Rohdaten. Ein solches Vorgehen stellt sicher, dass der Prozess reproduzierbar und nachvollziehbar ist (Woolf et al., 2023; Jousma et al., 2006). Allgemeiner Konsens herrscht in der Dringlichkeit einer guten Dokumentation und der Speicherung der unversehrten Rohdaten. Die FAIR-Prinzipien unterstreichen die Bedeutung der Findbarkeit (findable), Zugänglichkeit (accessible), Interoperabilität (interoperable) und Wiederverwendbarkeit (reusable) von Daten. Die Einhaltung dieser Prinzipien trägt dazu bei, die Effizienz der Datenverarbeitung und -analyse zu steigern, die Zusammenarbeit zwischen verschiedenen Disziplinen und Institutionen zu erleichtern und die Wiederverwendung von Daten für zukünftige Forschungsprojekte zu ermöglichen. Durch die Verfolgung der FAIR-Prinzipien werden Daten für künstliche Intelligenz (KI) und maschinelles Lernen (ML) nutzbar (Wilkinson et al., 2016). Das ist wichtig, da die Nutzung KI- und ML-basierter Modelle in den letzten Jahren einen exponentiellen Anstieg erfahren hat (Tao et al., 2022).

Zur Umsetzung dieser Leitfäden haben viele wissenschaftliche Disziplinen bereits diverse Methoden und Tools zur automatisierten Qualitätskontrolle implementiert. Dabei kommen sowohl statistische Verfahren als auch maschinelles Lernen (ML) und künstliche Intelligenz (KI) zum Einsatz (Erhan et al., 2020; Fiebrich et al., 2010). Diese Techniken erlauben die Erkennung von Anomalien, Ausreißern und Fehlern in den Daten durch die Identifikation von Mustern und Zusammenhängen innerhalb der Zeitreihen (Clemens-Meyer et al., 2021). Dennoch gibt es Raum für Verbesserungen bei der Zusammenführung dieser Verfahren in einen schlüssigen und effizienten Arbeitsablauf. Einigkeit herrscht darin, dass die Qualitätskontrolle von Zeitreihen wichtig ist, nicht jedoch darin, wie diese Qualitätskontrolle genau ablaufen soll (Porter et al., 2012). Es wurden einige Tools entwickelt, welche die Qualitätssicherung von Zeitreihen vereinfachen sollen. Darunter *TimeCleanser* (Gschwandtner et al., 2014), *Visplause* (Arbesser et al., 2016) und *Know your Enemy* (Gschwandtner und Erhart, 2018), die zwar automatische Checks zur Verfügung stellen, sich aber auf die visuelle Kontrolle durch Experten spezialisieren. Gschwandtner und Erhart (2018) sehen die Notwendigkeit, das "Human in the Loop" Prinzip zu beachten. Sie argumentieren, dass zur Differen-

zierung zwischen fehlerhaften und ungewöhnlichen Daten Hintergrundwissen nötig ist, welches nur durch einen geschulten Experten bereitgestellt werden kann. Auch in Lettland setzt der QC-Ablauf, der für die Grundwasserstands-Datenbank entwickelt wurde, auf das Vier-Augen-Prinzip durch einen *Controller* und einen *Corrector* (Retike et al., 2022). Dieser Experteneinsatz ist zwar eine etablierte Methode der Qualitätssicherung, ist jedoch auch zeitaufwendig, geprägt durch einen starken persönlichen Bias und nicht immer skalierbar (Campbell et al., 2013). Demgegenüber stehen Ansätze, die sich vermehrt auf die Automatisierung konzentrieren. Mit *AutoQC4Env* entwickeln Kaffashzadeh et al. (2019) aktuell ein Tool, bei dem automatische Tests und deren flexibler Einsatz im Vordergrund stehen. Die menschliche Interaktion kann hier zwar reduziert werden, jedoch müssen Testparameter noch immer manuell bestimmt werden. In den Qualitätskontrollansätzen von Taylor und Loescher (2013), Panagopoulos et al. (2021) und Jeong et al. (2021) für Umweltdaten, bezogen auf Meteorologie oder Oberflächengewässer, wird der Bedarf an menschlicher Interaktion durch eine zunehmend statistische Parametrisierung der Tests weiter minimiert. Obwohl diese Studien wichtige Erkenntnisse und Methoden zur automatisierten Qualitätskontrolle von Umweltdaten liefern, zeigen sie auch, dass es noch keinen spezifischen Workflow für die Überprüfung und Analyse von Umweltdaten, insbesondere von Grundwasserzeitreihen, gibt. Diese Forschungslücke in der Hydrologie unterstreicht die Notwendigkeit, weiterhin an der Entwicklung eines automatisierten Workflows für die Qualitätskontrolle von Grundwasserzeitreihen zu arbeiten.



## 2 Problemstellung und Zielsetzung

Hydrologische Daten können beispielsweise aufgrund von Softwaredefekten, Übertragungsfehlern und menschlichen Fehlern beeinträchtigt sein, wodurch ihre Verwendbarkeit beeinflusst wird. Obwohl die Datenqualität als Erfolgskriterium für Wissenschaftler, Ingenieure und Entscheidungsträger gleichermaßen wichtig ist, wird diese vor allem im Umgang mit Grundwasserdaten oft nur unzureichend erfüllt. Die traditionelle manuelle Qualitätssicherung stützt sich auf die Fachkenntnisse und Erfahrung von Experten. Diese Expertenbewertungen sind zwar in vielen Fällen unerlässlich, doch gibt es auch Bedenken hinsichtlich der Konsistenz, der Geschwindigkeit und der Skalierbarkeit solcher manueller Ansätze. Subjektive Entscheidungen liefern oft keine einheitlichen und reproduzierbaren Ergebnisse, und oft wird die Dokumentation der Qualitätskontrolle nicht nachvollziehbar gestaltet. Mit dem Fortschritt der Technologie wird daher die Möglichkeit einer automatischen Qualitätssicherung immer attraktiver. Der Einsatz automatisierter Tests für die automatische Qualitätssicherung erfordert eine Reihe von Testparametern. Diese Parameter werden statistisch ermittelt und unterliegen einer Unsicherheit. Die Genauigkeit und Robustheit dieser Parameter können sich auf die Effizienz und Genauigkeit des gesamten Systems auswirken. Außerdem stellt sich die Frage, ob ein solcher automatischer Workflow das Fachwissen von Experten wirklich abbilden kann. Vor diesem Hintergrund können zwei konkrete Forschungsfragen identifiziert werden:

1. **Wie robust ist die automatische Anomalieerkennung gegenüber Unsicherheiten in der statistischen Testparameterisierung?**
2. **Kann der automatische Workflow die manuelle Qualitätssicherung von Experten abbilden und verbessern?**

In dieser Masterarbeit wird ein automatisierter Workflow zur Qualitätssicherung von Grundwasserzeitreihen entwickelt, der auf modernen Datenverarbeitungstechniken und aktuellen wissenschaftlichen Erkenntnissen und Anforderungen basiert. Der vorgeschlagene Workflow kombiniert verschiedene QC-Verfahren, um eine umfassende und effiziente Qualitätskontrolle zu gewährleisten, und ermöglicht die Identifikation von Datenqualitätsproblemen. Der Workflow zur automatisierten Qualitätssicherung von Grundwasserzeitreihen orientiert sich an den genannten Richtlinien und existierenden Ansätzen und legt den Fokus insbesondere auf die Reduzierung menschlicher Interaktion. Der Workflow wird mit dem System of Automatic Quality Control (SaQC) umgesetzt. SaQC ist eine Toolbox des *Research Data Management Teams* des Helmholtz-Instituts für Umweltforschung, die Algorithmen zur Qualitätssicherung von Zeitreihen zur Verfügung stellt. Die Parametrisierung der Tests erfolgt statistisch aus Metadaten und der statistischen Beschaffenheit der Rohdaten. Als Datengrundlage dienen 20 Abtich-Zeitreihen von Grundwassermessstellen der Badenova AG in der Staufener Bucht. Da die Parameterbestimmung von QC-Tests stets mit einer Unsicherheit verbunden ist, wird eine Monte-Carlo-Simulation durchgeführt, um die Robustheit der Ergebnisse gegenüber Änderungen der Testparameter zu analysieren. Die Validierung des Workflows erfolgt anhand von drei Zeitreihen, die jeweils von drei unabhängigen Experten manuell geprüft werden. Im Anschluss an die manuelle Prüfung werden

dieselben Zeitreihen durch den entwickelten Workflow automatisch gekennzeichnet. Die Ergebnisse beider Methoden werden verglichen, um die Effizienz des automatisierten Workflows zu bestimmen und zu untersuchen, ob eine automatische QC die manuelle Prüfung teilweise ersetzen kann. Übergeordnetes Ziel des Workflows ist es, die Datenqualität zu verbessern, manuelle Eingriffe zu reduzieren und Prozesse zu vereinheitlichen, um eine Vergleichbarkeit zu gewährleisten. Darüber hinaus ist die Dokumentation der Qualitätskontrolle ein zentraler Aspekt für gute Nachvollziehbarkeit und Reproduzierbarkeit.

### 3 Material, Methoden und Vorgehensweise

#### 3.1 Daten und Projektregion

Die Daten, die dieser Arbeit zugrunde liegen, wurden in der Projektregion Staufener Bucht im Südwesten Baden-Württembergs erhoben. Die Staufener Bucht ist Teil des Wassergewinnungsgebietes des Hausener Wasserwerks an der Möhlin und liefert, zusätzlich zum östlich von Freiburg gelegenen Zartener Becken, Trinkwasser für die Stadt Freiburg im Breisgau. Im Wasserwerk Hausen werden jährlich 9 Millionen Kubikmeter Wasser von der Badenova AG zur Trinkwasserversorgung des westlichen Teils von Freiburg gefördert (BadenovaNetze, 2023). Der poröse Grundwasserleiter in der Staufener Bucht besteht aus kristallinen Schottern aus dem Schwarzwald und alpinen karbonatischen Materialien. Beide Arten von Schotter wurden während der letzten Eiszeit im Jungpleistozän abgelagert. Als Hauptgrundwasserleiter dienen die oberflächennahen, jungquartären Schotter. Auch die altquartären Ablagerungen sind grundwassergefüllt, erfüllen jedoch eine geringere Relevanz für die wasserwirtschaftliche Nutzung (Betting et al., 2006).

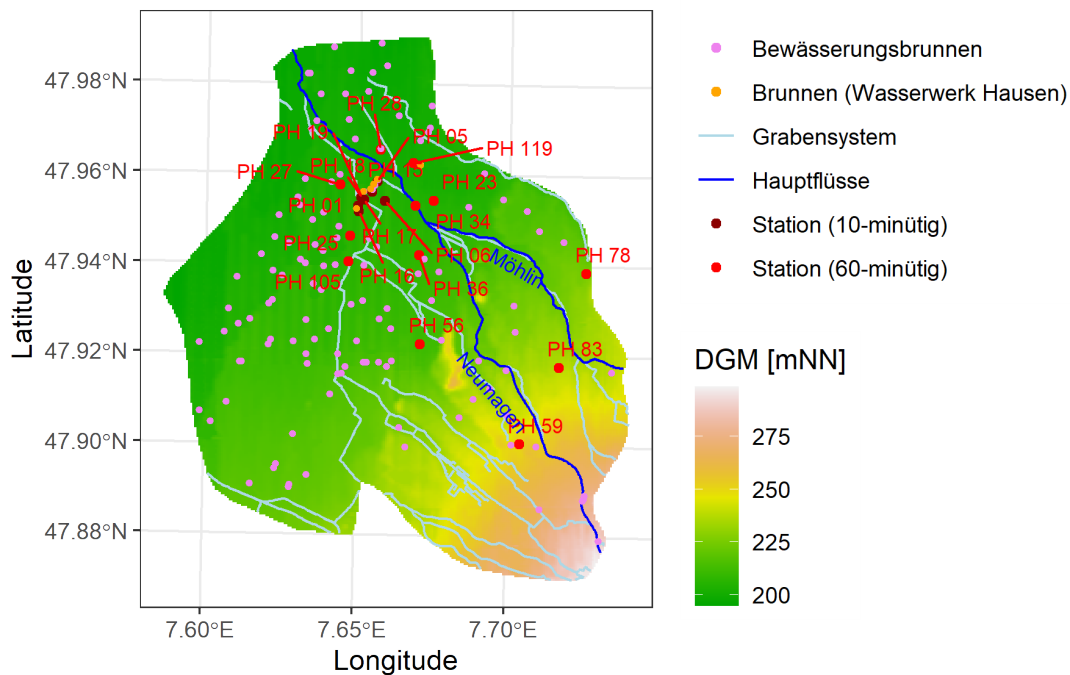


Abbildung 1: Projektregion Staufener Bucht mit den Hauptflüssen, dem Grabensystem, den Messstellen und Entnahmebrunnen

Der Jahresniederschlag liegt im Bereich von 650 bis 1050 mm, wodurch die Grundwasserneubildungsrate zwischen 100 und 300 mm variiert (Betting et al., 2006). Neben dem Niederschlag ist die Gewässerinfiltration die dominierende Inputgröße für das Grundwassersystem. Mit 1020 l/s macht sie laut Junker et al., 1977 einen Anteil von 58 Prozent am Gesamtinput aus. Die Hauptflüsse der Staufener Bucht sind Möhlin und Neumagen. Durch die hohe Durchlässigkeit der jungquartären Kies- und Schotterablagerungen besteht eine hohe Infiltration von Oberflächenge-

wässer in das Grundwasser. Diese Gewässerinfiltration ist von großer Bedeutung für die Grundwasserqualität. Die aus dem Schwarzwald fließenden Oberflächengewässer sind wenig mit Nitrat und Pestiziden belastet. Dadurch haben sie einen großen Verdünnungseffekt auf die flächenhafte Grundwasserneubildung aus dem Niederschlag, welche durch landwirtschaftliche Flächen mit Pestiziden und Nitrat angereichert wird (Peter, 1998). Nippes und Hettich (1988) haben die Infiltration der Möhlin untersucht und herausgefunden, dass infiltrierende Verhältnisse stark schwanken und sich teilweise auch umkehren. Die Grundwasserdynamik ist mit 2-3 m jahreszeitlichen Schwankungen gering, wobei oberflächengewässernahe Pegel durch die dort herrschenden Abflussverhältnisse geprägt sind.

Die Datenerhebung erfolgte an 20 Beobachtungsbrunnen, die von der Badenova AG betrieben werden. Ausgestattet sind diese mit automatischen Drucksonden der Firma Terratransfere, welche in 10- bzw. 60-minütigen Intervallen den Abstich aufzeichnen. Auffällig ist, dass einige Zeitstempel in den Rohdaten mehrfach auftauchen. Dies könnte möglicherweise durch einen Übertragungsfehler bei der Abspeicherung oder beim Auslesen von der Datenbank entstanden sein. Eine 10-minütige Messfrequenz wurde für Stationen gewählt, bei denen starke Fluktuationen des Grundwasserstandes aufgrund der Brunnennähe zu erwarten sind. Phasenweise veränderte Messfrequenzen treten jedoch bei allen Zeitreihen auf. Dabei weichen die Lücken zwischen zwei Messungen meist nur um wenige Sekunden oder Minuten von der üblichen Messfrequenz ab. Für die Station PH 034, welche eine Standard-Messfrequenz von 60 Minuten aufweist, werden ca. 19 % der Daten in einem 12-minütigen Intervall gemessen. Sechs Stationen weisen Lücken von 10 bis 69 Tagen auf (PH 034, PH 025, PH 059, PH 015, PH 036, PH 056). Bei Station PH 023 liegt eine Lücke von 15 Jahren vor. Insgesamt variiert die Zeitreihenlänge zwischen sechs Monaten bis neun Jahre (ohne Datenlücken).

Abbildung 1 zeigt die beschriebene Projektregion mit den Messstandorten, Entnahmebrunnen und den Oberflächengewässern. Die Distanzen zwischen den Messstationen und dem nächstgelegenen Entnahmebrunnen sind aus Tabelle 1 zu entnehmen. Bei der Validierung werden außerdem Zeitreihen von manuellen Handmessungen genutzt, welche ebenfalls von der Badenova AG bereitgestellt werden. Eine Niederschlagszeitreihe mit einer täglichen Messfrequenz wird von der nächstgelegenen Messstation in Schallstadt-Mengen (Stations-ID: 4419) des DWD bezogen (Deutscher Wetterdienst (2023)).

## 3.2 Anomalien in Zeitreihen

Durch die Automatisierung der Datenerfassung wird zwar die Effizienz erhöht, aber es können auch vermehrt Anomalien in den Zeitreihen auftreten (SADC-GMI, 2019). Diese Anomalien können als Datenpunkte definiert werden, die nicht das erwartete Verhalten zeigen (Blázquez-García et al., 2020). Anomalien spiegeln nicht immer Fehler oder unerwünschte Daten wider, sondern können auch auf reale und ungewöhnliche Ereignisse hinweisen. Grundsätzlich unterscheiden Blázquez-García et al. (2020) zwischen ungewollten Daten und interessanten Ereignissen.

In der Literatur zum Thema Anomalien in Zeitreihen werden verschiedene Arten von Anomalien beschrieben und diskutiert (Campbell et al., 2013; Horsburgh et al. (2015)). Jede dieser Anomalien besitzt eigene Merkmale und hat unterschiedli-

Tabelle 1: Metadaten der Stationen, der darin installierten Sensoren zur Grundwasseraufzeichnung und Distanzen zu den nächstgelegenen Entnahmeverbrunnen

Station	Sensor	Kabellänge [m]	Anzahl Datenpunkte	Distanz Station - Entnahmeverbrunnen [m]	Messfrequenz [min]	Messzeitraum
PH 001	900798	30	316234	67	10	2017-05-11 - 2023-02-16
PH 005	900799	30	165197	40	10	2020-01-01 - 2023-02-16
PH 006	900603	15	30608	430	10	2022-07-19 - 2023-02-16
PH 015	900712	15	301655	51	10	2017-02-02 - 2023-02-16
PH 016	900781	15	439517	100	10	2017-05-11 - 2023-02-16
PH 017	900778	15	442102	198	10	2017-05-11 - 2023-02-16
PH 018	021207	15	136961	160	10	2017-01-18 - 2019-08-01
PH 019	021202	15	608402	153	10	2017-01-18 - 2023-02-16
PH 023	021126	15	160980	873	60	1998-01-01 - 2023-02-16
PH 025	902140	14	24690	335	60	2020-03-04 - 2023-02-16
PH 027	900604	15	4694	226	60	2022-08-04 - 2023-02-16
PH 028	021392	15	157917	21	60	2015-09-01 - 2023-02-16
PH 034	902277	30	7812	1024	60	2020-03-06 - 2023-02-16
PH 036	210306	35	79452	146	60	2013-12-09 - 2023-02-16
PH 056	021207	15	75315	539	60	2020-01-27 - 2023-02-16
PH 059	902030	10	26125	211	60	2020-01-15 - 2023-02-16
PH 078	901220	15	55434	961	60	2018-03-20 - 2023-02-16
PH 083	900684	12	88888	1303	60	2017-02-23 - 2023-02-16
PH 105	902282	20	26974	333	60	2020-01-29 - 2023-02-16
PH 119	902229	10	46274	166	60	2020-03-04 - 2023-02-16

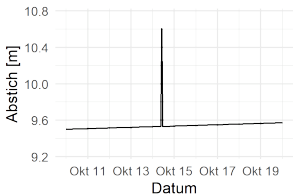
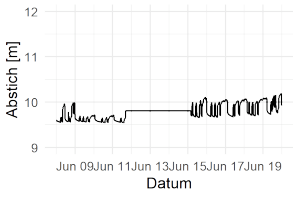
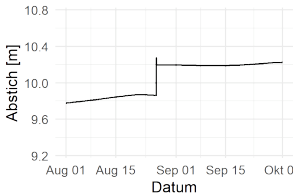
che Auswirkungen auf die Dateninterpretation. Der hier vorgestellte Workflow konzentriert sich auf die vier in Tabelle 2 genannten Anomalien, die in der wissenschaftlichen Literatur häufig und wiederkehrend hervorgehoben werden (Horsburgh et al., 2015; Durre et al., 2010). Diese vier Anomalien wurden wegen ihrer besonderen Relevanz im Kontext von Grundwasserzeitreihen und der spezifischen Analysemethoden, die in dieser Arbeit angewendet werden, ausgewählt.

Blázquez-García et al. (2020) unterscheiden zwischen Punktanomalien und Teilsequenzanomalien. Zu ersteren zählen einzelne Datenpunkte, die im Vergleich zu den benachbarten Punkten (lokale Anomalie) oder zu anderen Punkten in der gesamten Zeitreihe ein unerwartetes Verhalten zeigen. Beispiele für Punktanomalien sind Ausreißer, die laut DWA-M-181 Hinweise auf messtechnische Fehlfunktionen sein können, z. B. infolge von Spannungsspitzen oder darauf, dass das Mess-

verfahren über seine Grenzen hinaus eingesetzt wurde (Hollenberg et al., 2011). Allerdings können auch besondere geologische Bedingungen oder seltene hydrologische Ereignisse, wie außergewöhnliche Niederschlags- oder Abflussverhältnisse, zu ungewöhnlich hohen oder niedrigen Grundwasserständen führen.

Teilsequenzanomalien beziehen sich auf aufeinanderfolgende Datenpunkte, deren gemeinsames Verhalten ungewöhnlich ist, obwohl jede Beobachtung für sich genommen nicht unbedingt ein Punkt-Ausreißer ist (Blázquez-García et al., 2020). Sprünge in Zeitreihen werden von Blázquez-García et al. (2020) zu Teilsequenzanomalien gezählt. Häufig ist die Ursache hierfür die Anpassung eines neuen Messpunktes oder Referenzwerts. Neben dieser Ursache können jedoch auch andere Faktoren zu sprunghaften Veränderungen in der Datenreihe führen. Dazu gehören beispielsweise abrupte Veränderungen der hydrologischen Bedingungen, technische Eingriffe wie Reparaturen oder Gerätewechsel, oder extreme Wetterereignisse, die zu einem plötzlichen Anstieg oder Abfall des Grundwasserspiegels führen.

Tabelle 2: Erläuterung der Anomalien, die im Workflow berücksichtigt werden

Anomalien	Erläuterung	Beispiel														
Duplikate	Doppelte oder mehrfache gleiche Zeitstempel	<table border="1"> <thead> <tr> <th>time</th> <th>value</th> </tr> </thead> <tbody> <tr> <td>2017-02-23 14:21:14</td> <td>1.13188</td> </tr> <tr> <td>2017-02-23 14:21:14</td> <td>1.13188</td> </tr> <tr> <td>2017-02-23 15:00:00</td> <td>9.47102</td> </tr> <tr> <td>2017-02-23 15:00:00</td> <td>9.47102</td> </tr> <tr> <td>2017-02-23 16:00:00</td> <td>9.47000</td> </tr> <tr> <td>2017-02-23 16:00:00</td> <td>9.47000</td> </tr> </tbody> </table>	time	value	2017-02-23 14:21:14	1.13188	2017-02-23 14:21:14	1.13188	2017-02-23 15:00:00	9.47102	2017-02-23 15:00:00	9.47102	2017-02-23 16:00:00	9.47000	2017-02-23 16:00:00	9.47000
time	value															
2017-02-23 14:21:14	1.13188															
2017-02-23 14:21:14	1.13188															
2017-02-23 15:00:00	9.47102															
2017-02-23 15:00:00	9.47102															
2017-02-23 16:00:00	9.47000															
2017-02-23 16:00:00	9.47000															
Ausreißer	Die Über- oder Unterschreitung der technischen, physikalischen oder logischen Grenzen															
Flat-Line	Der Wert verändert sich über eine bestimmte Dauer nicht oder die Veränderung liegt unterhalb eines bestimmten Grenzwertes															
Sprung	Die Änderung des Mittelwertes zweier hintereinanderliegender Fenster liegt über einem bestimmten Wert															

Eine weitere typische Teilsequenzanomalie, die in Grundwasserzeitreihen auftreten kann, ist die sogenannte *flatline* oder *stuck value*-Anomalie. Diese Art von Anomalie tritt auf, wenn über einen längeren Zeitraum hinweg keine Veränderungen in den Messwerten beobachtet werden. Solch ein Muster ist ungewöhnlich, da natürliche Prozesse selten einen vollkommen konstanten Grundwasserstand

hervorbringen. Die Ursachen für diese Art von Anomalie können vielfältig sein. Oftmals weisen sie auf ein Problem mit dem Messgerät hin. Dies könnte ein mechanischer Defekt sein, der das Gerät daran hindert, Veränderungen im Grundwasserstand zu erfassen, oder ein Softwareproblem, das dazu führt, dass die gleichen Messwerte wiederholt ausgegeben werden. Diese Anomalie kann auch dann auftreten, wenn der Grundwasserspiegel so stark absinkt, dass der Sensor nicht mehr mit Wasser in Kontakt steht. Trotzdem kann ein unveränderter Grundwasserstand auch ein echtes Phänomen widerspiegeln. In Regionen mit sehr stabilen hydrogeologischen Bedingungen oder während Zeiten mit wenig Niederschlag und geringer Verdunstung könnte der Grundwasserstand über längere Zeiträume relativ konstant bleiben (Panagopoulos et al., 2021; Taylor und Loescher, 2013; Branisavljevic et al., 2011).

Schließlich können auch mehrfache Zeitstempel oder Lücken in den Zeitreihen Fehler darstellen. Solche Fehler können durch Softwareprobleme oder technische Probleme bei der Datenerfassung verursacht werden.

### 3.3 Workflow Entwicklung

Die aktuelle Qualitätssicherung der Daten erfolgt bei der Badenova AG ausschließlich manuell. Alle 3-4 Monate wird ein neues Datenpaket in die Datenbank von Terratransfer geladen. Mittels visueller Prüfung durch einen Experten, Simon Brenner, werden Ausreißer erkannt und durch einen vom Experten bestimmten Wert ersetzt. Diese Auswahl und Korrektur wird nicht dokumentiert und beruht ausschließlich auf der persönlichen Erfahrung des Experten. Der Workflow soll diese manuelle Kontrolle ersetzen, indem automatische Tests durchgeführt werden, die die Zeitreihe auf Anomalien überprüfen und markieren. Um menschliche Interaktion zu minimieren, sollen die Testparameter statistisch berechnet werden. Statistische Kennwerte der Roh- und Metadaten dienen hierfür als Berechnungsgrundlage. Dieses Vorgehen reduziert zum einen einen möglicherweise vorhandenen persönlichen Bias und zum anderen die Dauer des QC-Prozesses. Um den Workflow nachvollziehbar und reproduzierbar zu gestalten, werden alle verwendeten Tests und eingesetzten Parameter dokumentiert. Die statistischen Tests zur Qualitätskontrolle stammen aus der Toolbox des *System of automatic Quality Control* (SaQC). Das SaQC wurde vom *Research Data Management*-Team des Helmholtz-Instituts für Umweltforschung - UFZ entwickelt. Das SaQC entstand durch die Erfahrungen und Probleme, welche im Zuge der Einrichtung und Handhabung von automatisierten Qualitätskontrollsystemen für Umweltsensordaten aufgetreten sind (Schäfer et al., 2023, Schmidt et al., 2023). Die Auswahl und Parametrisierung der Tests orientieren sich an den Empfehlungen des DWA-M-181 (Hollenberg et al., 2011) sowie an relevanter wissenschaftlicher Literatur (Hubbard et al., 2005; Shulski et al., 2014; Taylor und Loescher, 2013).

#### 3.3.1 Datenbereitstellung und Anwendungsschnittstelle

Abbildung 2 zeigt die Architektur des Workflows. Für die Qualitätskontrolle werden ausschließlich die Abstichmessungen der Rohdaten genutzt, die in csv-Dateien gespeichert sind. Anhand des Dateinamens können alle Zeitreihen einem Sensor und einer Messstelle zugeordnet werden. Flags, die während der Qualitätskontrolle generiert werden, sowie mögliche Korrekturen, werden in zusätzlichen csv-

Dateien ausgegeben. Auch diese Dateien können über den Dateinamen einem Sensor und einer Station zugeordnet werden. Die Flags werden auf der Ebene des Zeitstempels in zusätzlichen Spalten pro Test gespeichert. Eine weitere Ausgabedatei beinhaltet die Parameter der verschiedenen Tests. In dieser Datei werden auch das Anfangs- und Enddatum der Zeitreihe gespeichert, um die Reproduzierbarkeit der Berechnungsgrundlage sicherzustellen. Metadaten und Zeitreihen von manuellen Handmessungen der Brunnen werden in der Smallworld GIS-Datenbank von GIT HydrosConsult gespeichert und zur Nutzung als csv-Dateien bereitgestellt. Der Workflow wird in R-Studio programmiert (RStudio Team, 2020). Da die Schnittstelle zu SaQC jedoch in Python realisiert wurde, kombiniert eine R-Markdown-Datei sowohl Python-Code als auch R-Code-Abschnitte.

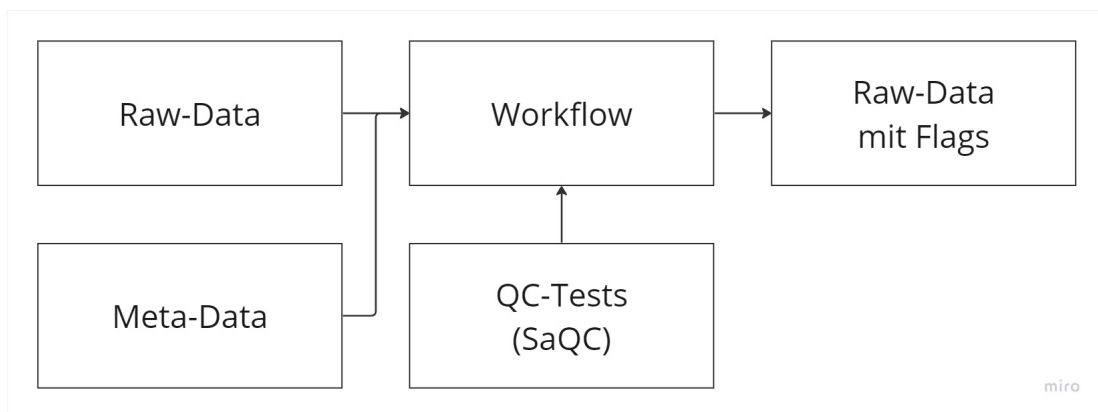


Abbildung 2: Grundaufbau der Workflow-Struktur

### 3.3.2 Datenprozessierung und Prüfabfolge

Der Workflow setzt sich aus den in Abbildung 3 gezeigten Abschnitten zusammen. Im Folgenden werden diese Schritte genauer erklärt.

**Preprocessing** Ziel des Preprocessing ist es, die Rohdaten in ein, für die weiteren Schritte geeignetes und einheitliches Format transformiert. Der Fokus liegt dabei in der Formatierung von Datums- und Zeitstempeln sowie dem Umgang mit doppelten Zeitstempeln. Das System of automatic Quality Control (SaQC) erfordert eine bestimmte Struktur als Inputformat. Dabei muss der Zeitstempel in Python als Index im Format *YYYY-MM-DD hh:mm:ss* vorliegen. Für SaQC muss die Zeitreihe monoton steigend oder fallend sein, für einige Algorithmen ist eine streng monoton steigende oder fallende Zeitreihe Voraussetzung. Dies bedeutet, dass die Zeitreihe chronologisch sortiert sein muss und doppelte Zeitstempel teilweise nicht verarbeitet werden können. Um Konflikte in den Algorithmen zu vermeiden, werden alle Daten vor dem Laden in das SaQC sortiert und doppelte Zeitstempel bereinigt. Bei der Bereinigung doppelter Zeitstempel wird zwischen zwei Fällen unterschieden: (1) Wenn doppelte Zeitstempel mit den gleichen Messwerten vorliegen, wird ein Datenpunkt behalten und alle anderen als *Duplikat* markiert. Dies folgt der Annahme, dass der Messwert an sich valide ist und lediglich doppelt gespeichert wurde. Sollte ein solcher Datenpunkt in den *Basic*- bzw. *Advanced*-Tests geflaggt werden, so werden auch die Duplikate mit dem



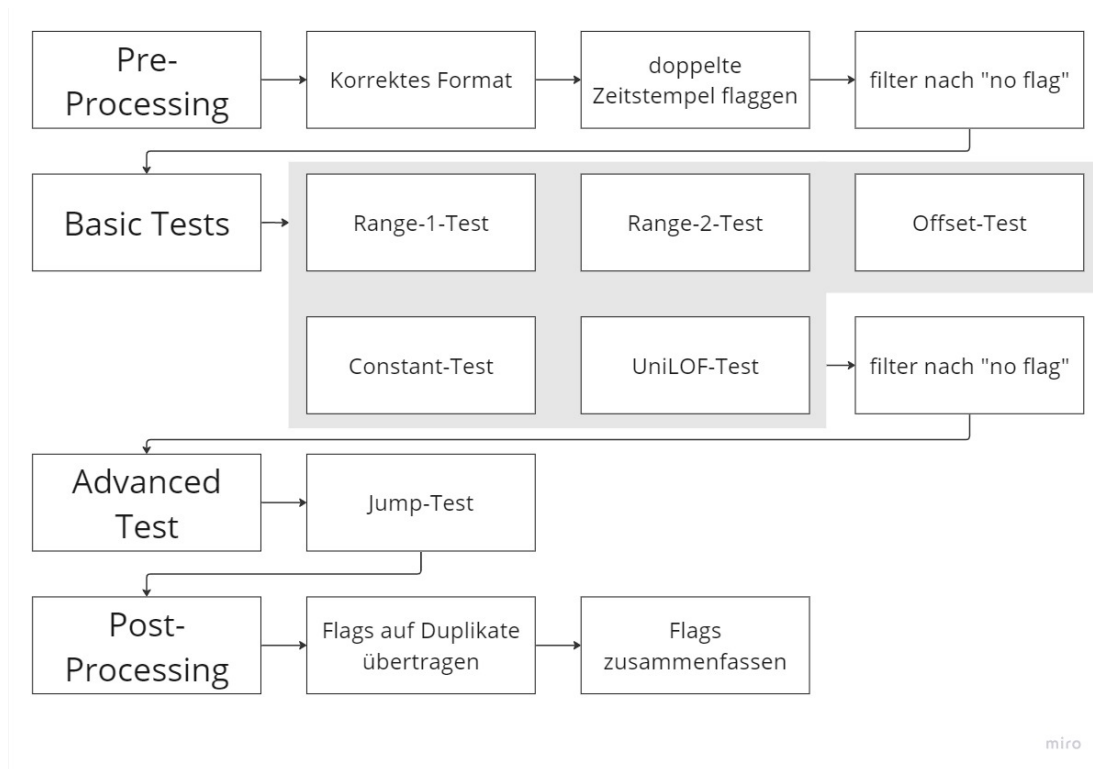


Abbildung 3: QC-Workflow unterteilt in Pre-Processing, Basic-Tests, Advanced-Test und Post-Processing

gleichen Flag versehen. (2) Liegen doppelte Zeitstempel mit unterschiedlichen Messwerten vor, werden alle Datenpunkte als *Duplikat* markiert. Grund hierfür ist, dass während des Pre-Processing nicht eindeutig bestimmt werden kann, welcher der Werte zu diesem Zeitpunkt valide bzw. fehlerhaft ist. Der Ausschluss sämtlicher Zeitstempel dient somit in diesem Fall (2) einer möglichst hohen Datenqualität. In den nachfolgenden Schritten werden nur Daten bearbeitet, welche im Pre-Processing nicht als *Duplikat* markiert wurden.

**Identifikation von Anomalien** Nach dem Preprocessing der Daten werden sogenannte Basic- und Advanced-Tests durchgeführt, um die in Tabelle 2 dargestellten Anomalien zu identifizieren. Jeder Basic-Test wird stets auf den gleichen, Duplikat-bereinigten Rohdaten durchgeführt, um eine gegenseitige Beeinflussung der Tests zu verhindern. Pro Test wird daher eine neue Spalte für die Markierungen der Anomalien erstellt. Zur Identifizierung von Anomalien werden sechs spezifische Tests eingesetzt. Range (1), Range (2), Offset und LOF wurden ausgewählt, um einzelne Punkt-Ausreißer zu identifizieren. Der **Offset-Test** bietet darüber hinaus die Möglichkeit, Gruppen von Ausreißern zu erkennen. Zur Erkennung der *flat-line*-Anomalie wird der **Constant-Test** verwendet. Sprünge in der Zeitreihe werden mittels des **Jump-Tests** identifiziert. Dieser letzte Test wird in die Gruppe der *Advanced-Tests* eingeordnet, da eine Korrektur durch einfaches Löschen des anomalen Datenpunktes und anschließendes Interpolieren hier nicht möglich ist. Die *Advanced-Tests* sowie deren Parametrisierung werden auf den von den *Basic-Tests* bereinigten Daten durchgeführt, nicht auf den Duplikat-bereinigten Rohdaten. Kapitel 3.3.3 beschreibt die einzelnen Tests sowie deren

Parametrisierung näher.

**Post-Processing** Nachdem Anomalien durch die automatisierten Tests markiert wurden, werden diese Flags auf die zuvor entfernten Duplikate übertragen. Anschließend werden alle Ergebnisse zusammengetragen, indem eine neue Spalte erstellt wird, in der der Datenpunkt gekennzeichnet wird, sobald mindestens ein Test diesen als Anomalie erkannt hat. Um Nachvollziehbarkeit und Reproduzierbarkeit zu gewährleisten, werden die Ergebnisse der Tests auf Zeitstempel-Basis abgespeichert und die genutzte Testabfolge inklusive ihrer Parameter dokumentiert. Es werden keine Korrekturen der Rohdaten vorgenommen.

### 3.3.3 Testbeschreibung und Parametrisierung

Die Algorithmen-Toolbox von SaQC stellt eine Vielzahl von statistischen Tests zur Qualitätskontrolle zur Verfügung. Alle verwendeten Algorithmen des SaQC erfordern die Angabe bestimmter Parameter. Tabelle 3 zeigt die genutzten Tests, die Ermittlung der Testparameter und eine jeweilige Beschreibung. Wie auch bei Taylor und Loescher (2013) erfolgt die Ermittlung der Parameter *min* und *max* (Range(2)), *thresh* und *tolerance* (Offset), *window* (Constant) und *thresh* (Jump) datengesteuert. Die statistische Berechnung basiert auf der Annahme einer Normalverteilung der Daten und folgender Formel, wie sie auch von Shulski et al. (2014), Hubbard et al. (2005) und Taylor und Loescher (2013) angewandt wird:

$$p = \mu \pm x \cdot \sigma \quad (1)$$

Hier wird der gesuchte Parameter ( $p$ ) durch den Mittelwert ( $\mu$ ) plus bzw. minus einer bestimmten Anzahl ( $x$ ) multipliziert mit der Standardabweichung ( $\sigma$ ) berechnet. Die zugrundeliegende Verteilung für die Berechnung von Mittelwert und Standardabweichung sowie die Wahl von  $x$  hängen von dem zu bestimmenden Parameter ab.

**Range (1)-Test** Der Range (1) markiert Werte, die über oder unter physikalisch unmöglichen Messgrenzen liegen. Die Parameter des Range (1) werden aus den Metadaten bestimmt. Für *max* wird die Kabellänge des Sensors verwendet. Wird ein Wasserstand gemessen, der tiefer liegt als der Sensor, kann dieser Wert nicht als zuverlässige Messung betrachtet werden. Der *min*-Parameter wird für alle Stationen auf 0 gesetzt. Ein Wert  $\leq 0$  kann ebenfalls als Fehler betrachtet werden, da der Sensor keinen Grundwasserstand oberhalb der Messoberkante messen kann.

Für die Ermittlung der Parameter *min* und *max* (Range (2)), *thresh* (Offset) und des *emphwindow* (Constant) werden die Zeitreihen bereinigt, indem alle durch den Range (1) als *BAD* markierten Punkte aus der Zeitreihe entfernt werden. Der Grund dafür ist, dass grobe Anomalien, die mit großer Sicherheit als Fehler gewertet werden können, aus den nachfolgenden Parameterberechnungen ausgeschlossen werden. Da nicht alle Messpunkte der Standard-Messfrequenz der jeweiligen Zeitreihe entsprechen, werden für die Berechnung der Parameter alle Datenpunkte mithilfe der in SaQC zur Verfügung gestellten *shift*-Funktion auf einheit-

liche Messfrequenzen verschoben. Bei einer 10-minütig aufgenommenen Zeitreihe kann es beispielsweise vorkommen, dass einige Datenpunkte 10 Minuten und eine Sekunde auseinanderliegen. Dafür wurde die *nshift*-Methode genutzt. Diese verschiebt die Messpunkte zum nächstgelegenen Messfrequenz-Zeitpunkt, sofern dieser nicht mehr als eine halbe Messfrequenz entfernt ist.

Tabelle 3: Beschreibung der Testparameter und ihre Ermittlung

Test	Parameter	Beschreibung	Ermittlung (Berechnungsgrundlage)
Range (1)	<i>min</i>	Untergrenze für Daten	Messoberkante (0 m)
	<i>max</i>	Obergrenze für Daten	Sensor-Kabellänge
Range (2)	<i>min</i>	Untergrenze	Formel [2] (Abstich-Verteilung)
	<i>max</i>	Obergrenze	Formel [3] (Abstich-Verteilung)
Offset	<i>window</i>	<i>max.</i> Dauer einer Ausreißergruppe	10 * Messfrequenz
	<i>thresh</i>	Mindest-Differenz für Anomalieerkennung	Formel [4] (Abstichdifferenzen-Verteilung)
	<i>tolerance</i>	Zulässige Differenz zum Ursprungswert vor/nach Offset	Formel [5] (Abstichdifferenzen-Verteilung)
Constant	<i>window</i>	Fensterlänge ohne signifikante Änderungen	Formel [6] (Verteilung der Perioden ohne Änderung)
	<i>thresh</i>	<i>max.</i> Änderung pro Fenster für Anomalieerkennung	0 m
LOF	n	Anzahl der Perioden für LOF-Berechnung	20 (SaQC Standardwert)
	<i>thresh</i>	Schwellenwert für Anomalieerkennung	1.5 (SaQC Standardwert)
Jump	<i>window</i>	Größe zweier Fenster für Mittelwertberechnung	24 Stunden
	<i>thresh</i>	Mindest-Differenz der Fenstermittelwerte zur Anomalieerkennung	Formel [7] (Verteilung der Differenzen der Tagesmittelwerte)

**Range (2)-Test** Dieser Test prüft unwahrscheinliche Werte, die über oder unter einer gegebenen Grenze liegen. Zur Berechnung dieser Grenzen ( $min$ ,  $max$ ) wird die Abtich-Verteilung der durch den Range (1) bereinigten Zeitreihe für die folgenden Formeln genutzt:

$$min = \mu - 4 \cdot \sigma \quad (2)$$

$$max = \mu + 4 \cdot \sigma \quad (3)$$

Bei einer Normalverteilung kann davon ausgegangen werden, dass etwa 99,99 % der Daten innerhalb des Bereichs  $mean \pm 4 \cdot std$  liegen (Shulski et al., 2014). Die Wahl von vier Standardabweichungen wird von Shulski et al. (2014) und Hubbard et al. (2005) nahegelegt und stellt einen Kompromiss zwischen der Identifikation von echten Ausreißern und der Verringerung der Wahrscheinlichkeit dar, dass normale Werte als fehlerhaft markiert werden.

**Offset-Test** Der Offset-Test benötigt die Parameter  $thresh$ ,  $window$  und  $tolerance$ . Der  $thresh$ -Parameter bestimmt die maximale Differenz, um die sich zwei benachbarte Punkte unterscheiden dürfen. Wird dieser Wert überschritten, werden die folgenden Werte so lange als Anomalie markiert, bis der Wert wieder auf den Ursprungswert,  $\pm$  die gegebene  $tolerance$ , abfällt oder ansteigt. Falls der Verlauf der Offset-Werte länger andauert, als der gegebene  $window$ -Parameter erlaubt, findet keine Markierung statt. Der  $emphwindow$  wird in Rücksprache mit der Badenova AG auf das Zehnfache der Messfrequenz festgelegt. Die Ermittlung von  $thresh$  und  $tolerance$  erfolgt über die Verteilung aller Differenzen zwischen benachbarten Datenpunkten der Range (1)-bereinigten und auf equidistante Messintervalle verschobenen Daten und wird über folgende Formeln berechnet:

$$thresh_{offset} = \frac{(\mu + 4 \cdot \sigma) + |(\mu - 4 \cdot \sigma)|}{2} \quad (4)$$

$$tolerance = \frac{(\mu + 2 \cdot \sigma) + |(\mu - 2 \cdot \sigma)|}{2} \quad (5)$$

Um die Parametrisierung des  $thresh$  nicht vollkommen datengesteuert zu belassen, können hier manuell Parametergrenzen durch Experten bestimmt werden. Liegen die berechneten Parameterwerte außerhalb dieser Limits, wird der Parameter auf das nächste Limit angehoben oder herabgesetzt.

**Constant-Test** Der Constant-Test benötigt die Parameter  $thresh$  und  $window$ . Der Test markiert Datenpunkte als Anomalien, wenn sich der Wert für mindestens die Dauer von  $window$  um nicht mehr als  $thresh$  ändert. Für den  $thresh$ -Parameter wird der Wert 0 für alle Stationen verwendet. Der  $window$ -Parameter wird aus der Verteilung aller Längen der Perioden ohne Änderung durch Formel [6] berechnet. Die Verteilung unterscheidet sich von einer Normalverteilung, indem sie nur positive Werte aufweist, einen Mittelwert nahe Null besitzt und eine ausgeprägte Abnahme zu höheren Werten zeigt. Deshalb wird hier ausschließlich die Formel verwendet, bei der die Standardabweichungen addiert werden (Shulski et al., 2014).

$$window_{constant} = \mu + 6 \cdot \sigma \quad (6)$$

In Anlehnung an die Methode von Shulski et al. (2014) wurde auch in dieser Studie der Wert  $x$  auf 6 festgelegt. Wie auch für den *Offset-thresh*-Parameter können auch hier manuelle Grenzen bestimmt werden. Liegen die berechneten Werte des *window*-Parameters außerhalb dieser Limits, wird der Parameter auf das nächste Limit angehoben oder herabgesetzt.

**LOF-Test** Der LOF-Test wählt zunächst einen bestimmten Datenpunkt aus und identifiziert dessen Nachbarn. Bei den Nachbarn handelt es sich um andere Datenpunkte, die dem ausgewählten Punkt aufgrund ihrer Werte am nächsten liegen, nicht basierend auf ihrer Position. Der LOF-Test schätzt die lokale Dichte des ausgewählten Datenpunkts, indem er die Abstände zu seinen Nachbarn betrachtet. Es wird berechnet, wie dicht die Nachbarn basierend auf ihren Werten um den ausgewählten Punkt gruppiert sind. Nachdem die lokale Dichte des ausgewählten Datenpunkts bestimmt wurde, vergleicht der LOF-Test diese mit den Dichten seiner Nachbarn. Wenn der ausgewählte Punkt eine geringere Dichte als seine Nachbarn aufweist, deutet dies darauf hin, dass der Punkt im Vergleich zu den umliegenden Punkten relativ isoliert ist und ein Ausreißer sein könnte. Der LOF-Test weist dem ausgewählten Datenpunkt einen Anomalie-Score zu, der auf dem Unterschied zwischen seiner Dichte und den Dichten seiner Nachbarn basiert. Punkte, deren Dichte deutlich geringer ist als die ihrer Nachbarn, erhalten einen höheren Anomalie-Score, was darauf hindeutet, dass sie mit größerer Wahrscheinlichkeit Ausreißer sind (Breunig et al., 2000). Die Werte für die Parameter  $n$  (Anzahl der betrachteten nächsten Nachbarn) und *thresh* (Schwellenwert für die Kennzeichnung als Anomalie) werden von den default-Werten des SaQC übernommen.

**Jump-Test** Der Jump-Test markiert Datenpunkte, bei denen sich der Mittelwert der Daten signifikant ändert. Die Änderung wird durch den Vergleich der Mittelwerte zweier benachbarter Fenster mit der jeweiligen Länge *window* ermittelt. Wenn die Mittelwerte der Fenster sich um mehr als einen vorher definierten Schwellenwert (*thresh*) unterscheiden, wird der Wert zwischen den Fenstern als Sprung gekennzeichnet. Der *window*-Parameter wird einheitlich auf 24 Stunden festgelegt. Zur Bestimmung des *thresh*-Parameters wird die durch den Basic-Test bereinigte Zeitreihe zunächst auf Tagesmittelwerte aggregiert. Anschließend werden die Differenzen zwischen aufeinanderfolgenden Tagen ermittelt. Aus dieser Verteilung wird mithilfe der folgenden Formel der *thresh*-Parameter berechnet:

$$thresh_{jump} = \frac{(\mu + 4 \cdot \sigma) + |(\mu - 4 \cdot \sigma)|}{2} \quad (7)$$

### 3.4 Anomalieerkennung mit Testparameterunsicherheit

Um zu untersuchen, wie sich der Output eines Tests bei Änderung eines Parameters verändert, werden für alle Parameter aus Tabelle 3 Unsicherheitsbereiche festgelegt. Mithilfe des Monte-Carlo-Verfahrens werden 100 Simulationen mit zufällig gezogenen Parameterkombinationen aus einer uniformen Verteilung innerhalb dieser Unsicherheitsbereiche durchgeführt (Harrison, 2010). Anschließend wird die Anzahl aller markierten Datenpunkte verglichen. Die 100 Simulationen werden jeweils separat für die Tests Range (2), Offset, Constant, LOF und Jump

durchgeführt. Dieses Verfahren untersucht die Robustheit der Ergebnisse gegenüber Änderungen und Unsicherheiten der Eingabeparameter. Es kann auch das Verständnis der Beziehungen zwischen Input- und Outputvariablen in einem System oder Modell verbessern. Darüber hinaus ermöglicht es, Parameter zu identifizieren, die einen signifikanten Einfluss auf die Ergebnisse haben und die im Fokus stehen sollten. Neben den signifikant wirkenden Parametern können auch solche identifiziert werden, die nur einen geringen Einfluss auf die Ergebnisse haben und daher weniger Berücksichtigung erfordern.

### 3.5 Validierung mit Experten-Qualitätskontrolle und Umweltsystemkontext

Zur Validierung des Workflows untersuchen drei Experten manuell drei Zeitreihen auf Duplikate und Anomalien. Alle drei Experten arbeiten im hydrologischen Kontext mit Daten und verfügen über mehrjährige Berufserfahrung. Die gleichen drei Zeitreihen durchlaufen auch den Workflow zur Qualitätssicherung. Der Vergleich zwischen manueller und automatischer Anomalieerkennung ist eine Methode, welche häufig in der Literatur verwendet wird (Talagala et al., 2019). Auch Schäfer et al. (2023) nutzen diesen Vergleich, um die SaQC-Tests zu validieren. Die Zeitreihen stammen von Grundwassermessstellen in der Staufener Bucht und wurden aus den zuvor für die Entwicklung des Workflows verwendeten Daten zufällig ausgewählt. Die Anweisung zur Qualitätskontrolle für die Experten lautet wie folgt:

”Das Ziel ist es, mögliche Fehler, Ausreißer, doppelte Zeitstempel, eingefrorene Werte oder andere Anomalien zu identifizieren.”

Tabelle 4: Beispielhafte Darstellung einer Confusionmatrix

	wahre Klasse		
	GOOD	BAD	
berechnete Klasse	GOOD	True GOOD (TG)	False GOOD (FG)
	BAD	False BAD (FB)	True BAD (TB)

Es wird die Annahme getroffen, dass die Experten bei dem Auftreten von Duplikaten entweder alle Duplikate gleichermaßen markieren oder nur den ersten Zeitstempel weiter untersuchen. Bei dem Vergleich zwischen Expertenflags und automatischen Flags werden nur die duplikatbereinigten Daten genutzt. Bei mehrfachem Vorkommen von Zeitstempeln wird also nur der erste Datenpunkt analysiert. Die Unterschiede und Gemeinsamkeiten zwischen den Ergebnissen der manuellen und der automatisierten Qualitätssicherung werden mit Hilfe einer Confusionmatrix (Fehlermatrix, Verwirrungsmatrix) quantifiziert. Tabelle 4 zeigt den strukturellen Aufbau einer solchen Matrix. Bei einer Confusionmatrix wird zwischen einer *wahren Klasse* und einer *berechneten Klasse* differenziert. In diesem Zusammenhang ist von *TRUE GOOD* (TG) und *TRUE BAD* (TB) die Rede, wenn beide Klassen für die positive bzw. negative Kategorie stimmen. Als *FALSE GOOD* (FG) wird ein Datenpunkt dann klassifiziert, wenn die berechnete Klasse diesen als positiv einordnet, während die wahre Klasse für negativ stimmt, und umgekehrt für *FALSE BAD* (FB) (Davis und Goadrich, 2006). Im Vergleich der manuellen und automatischen Anomalieerkennung bezieht sich die

wahre Klasse auf die manuellen und die berechnete Klasse auf die automatischen Flags. Ein Datenpunkt gilt in der wahren Klasse als *BAD*, wenn einer der drei Experten diesen als Anomalie markiert. Ein Datenpunkt gilt in der berechneten Klasse als *BAD*, wenn mindestens ein Test diesen als Anomalie markiert. Für die Quantifizierung der Güte der automatischen Qualitätskontrolle im Vergleich zur manuellen Qualitätskontrolle werden die Kennzahlen Sensitivität, Spezifität, Präzision und F-Score über die Formeln [8] bis [11] berechnet und verglichen (Davis und Goadrich, 2006, Caelen, 2017).

$$\text{Recall} = \frac{\text{TB}}{\text{TB} + \text{FG}} \quad (8)$$

$$\text{Spezifität} = \frac{\text{TG}}{\text{TG} + \text{FB}} \quad (9)$$

$$\text{Präzision} = \frac{\text{TB}}{\text{TB} + \text{FB}} \quad (10)$$

$$\text{F-Score} = \frac{2 \cdot \text{Recall} \cdot \text{Präzision}}{\text{Recall} + \text{Präzision}} \quad (11)$$

Der Recall-Wert gibt an, wie viele der durch die Experten identifizierten Datenpunkte auch von dem Workflow erkannt wurden. Ein niedriger Recall bedeutet, dass viele der von den Experten markierten Anomalien vom System übersehen wurden. Die Präzision hingegen misst den Anteil der korrekt erkannten Anomalien an allen vom System als Anomalien markierten Datenpunkten. Ein niedriger Präzisionswert zeigt an, dass das System im Vergleich zur manuellen Markierung viele falsche Anomalien identifiziert hat. Die Spezifität gibt den Prozentsatz der korrekt als normal identifizierten Datenpunkte im Verhältnis zu allen tatsächlich normalen Datenpunkten an. Eine niedrige Spezifität zeigt an, dass das System viele normale Datenpunkte fälschlicherweise als Anomalien gekennzeichnet hat. Der F-Score ist das harmonische Mittel aus Precision und Recall und schafft ein ausgewogenes Verhältnis zwischen ihnen. Ein hoher F-Score zeigt ein ausgeglichenes Verhältnis zwischen Precision und Recall an, während ein niedriger F-Score auf ein Ungleichgewicht oder eine schlechte Leistung bei Precision oder Recall hinweist.

Um die Expertenflags differenziert zu betrachten und mit den Monte-Carlo Simulationen zu vergleichen, wird eine Multiklassen-Confusionmatrix genutzt. Diese hat nicht nur zwei Klassen (*GOOD* und *BAD*), sondern verschiedene Stufen von *BAD* (1/3, 2/3, 3/3) abhängig davon, wie viele Experten einen Datenpunkt als Anomalie markiert haben, bzw. in wie vielen Monte-Carlo Simulationen von 100 ein Datenpunkt als *BAD* markiert wurde.

Es wird auch geprüft, welche Tests die Ursache für die automatischen Flags bei Datenpunkten sind, die sowohl von Experten als auch durch den automatischen Workflow ausgewählt wurden, sowie bei Datenpunkten, die nur durch den automatischen Workflow erzeugt wurden.

Abschließend wird die Anomalieerkennung visuell im Zusammenhang mit dem Niederschlag und manuellen Vergleichsmessungen betrachtet, um erste Zusammenhänge zwischen der Anomalieerkennung und der Umweltsystem-Dynamik für die weitere Forschung aufzuzeigen.

## 4 Ergebnisse

### 4.1 Qualitätskontrolle

#### 4.1.1 Duplikate

Bei den 20 Messstationen treten zwischen 0 und 67,7 Prozent Duplikate auf (Mittelwert = 20,1%). Bei fünf Messstationen werden zwischen 10 und 23 Datenpunkte entfernt, da nicht nur gleiche Zeitstempel identifiziert wurden, sondern sich die Abstichwerte auch unterschieden haben. Duplikate stehen nicht im Zusammenhang zur Messfrequenz (ANOVA, p-Wert = 0,424) und der Datenpunktanzahl (linearer Trend, p-Wert = 0,2).

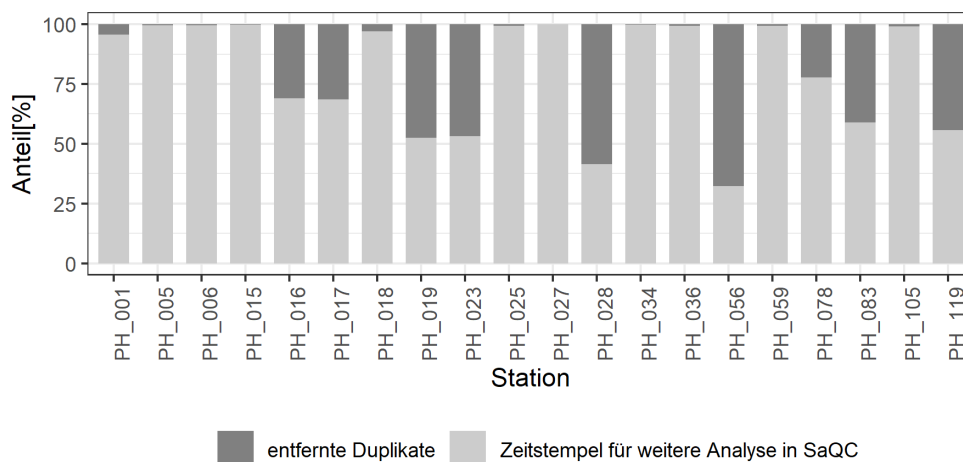


Abbildung 4: Anteil an Duplikaten pro Station

#### 4.1.2 SaQC-Testparameter

Die Annahme einer Normalverteilung der Daten muss fast für alle Verteilungen bei allen Stationen (Anderson-Darling-Test: p-Wert < 0,05) verworfen werden. Einzig die Verteilung der Differenzen der Tagesmittelwerte der Station PH 027 konnte statistisch als Normalverteilung erkannt werden (Anderson-Darling-Test: p-Wert = 0,07424). Abbildungen 19 bis 22 im Anhang zeigen die zugrundeliegenden Verteilungen zur Berechnung aller Parameter für alle Stationen. Die Verteilungen der Differenzen der gemessenen Werte bzw. der Tagesmittelwerte aller Stationen sind eher spitz mit schweren Schwänzen (Kurtosis > 3). Dennoch ähneln die Verteilungen der Abstichwerte und der Differenzen zwischen den einzelnen Datenpunkten bzw. der Tagesmittelwerte optisch einer Normalverteilung.

Die Histogramme in Abbildung 5 veranschaulichen beispielhaft für die Station PH 001 die zugrundeliegenden Verteilungen, die zur Berechnung der Parameter *min* und *max* (Range (2)), *thresh* (Offset), *window* (Constant) und *thresh* (Jump) verwendet werden. Die errechneten und genutzten Parameter sind ebenfalls in den Abbildungen dargestellt. Für die Station PH 001 liegen alle ermittelten Parameterwerte außerhalb der Hauptverteilung. Obwohl die Verteilung der Abstichwerte eher nach rechts verzogen ist, hat sie einen ausgeprägten Ausläufer nach links



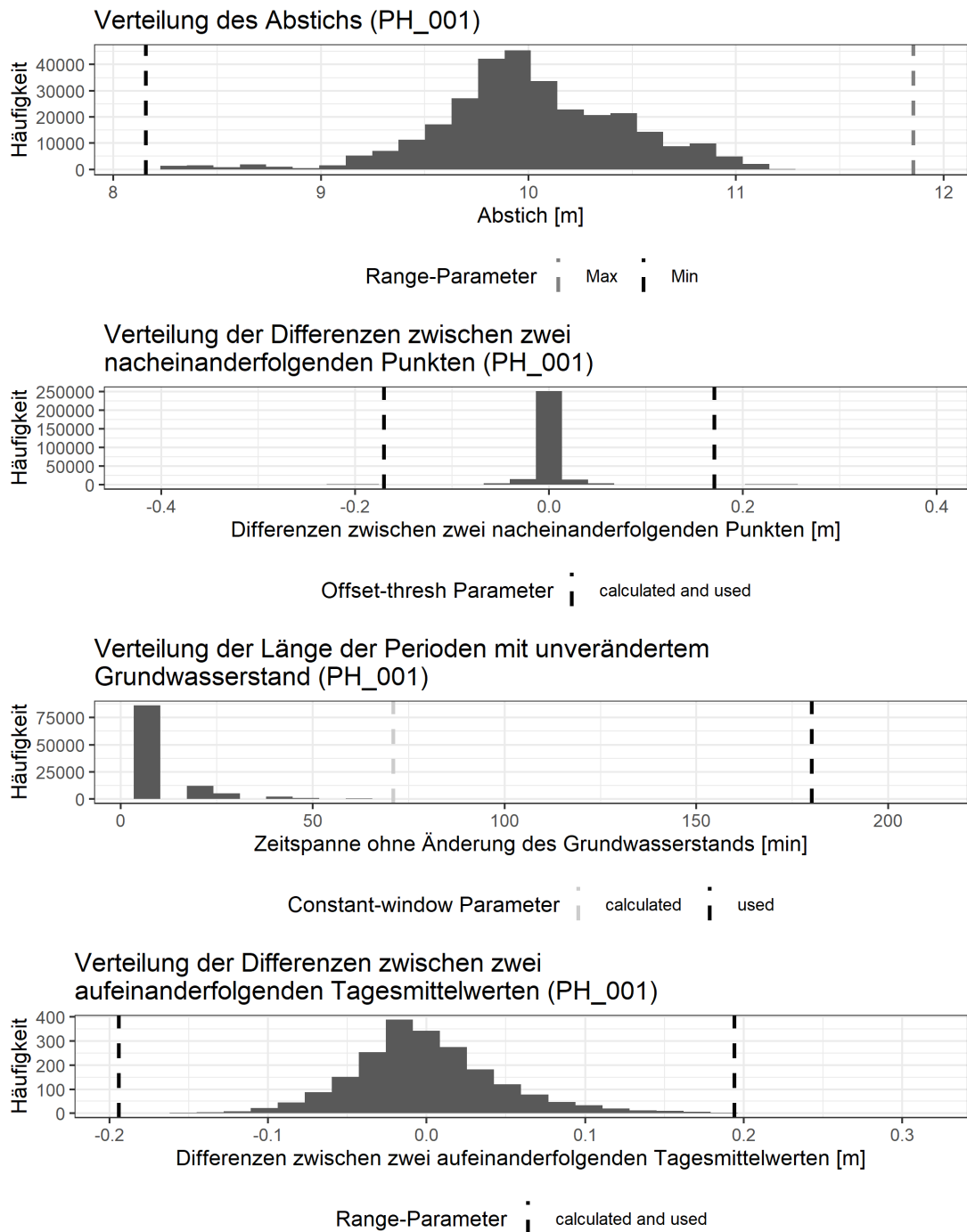


Abbildung 5: Zugrundeliegende Verteilungen für die Berechnung der Parameter *min* und *max* (Range (2)), *thresh* (Offset), *window* (Constant) und *thresh* (Jump), mit den berechneten und genutzten Parametern markiert.

(Kurtosis = 5, Schiefe = 0,66). Dieser Schweif wird jedoch durch den ermittelten *min*-Parameter von 8,16 als Schwellenwert durch den Range (2) nicht als Anomalie erkannt. Die Verteilung der Differenzen zweier nacheinanderfolgender Datenpunkte, welche zur Berechnung des *thresh*-Parameters des Offset-Tests genutzt wird, zeigt eine symmetrische Verteilung um Null mit starkem Abfall der Häufigkeiten in beide Richtungen (Kurtosis = 26, Schiefe = 0,22). Es lassen sich jedoch zwei zusätzliche Cluster um  $\pm 0,2$  erkennen. Bei Verwendung der Parameterberechnung von  $Mittelwert \pm 4 * Standardabweichung$  werden die Differenzen, welche in diesen Clustern liegen, als Anomalien betrachtet. Für den Parameter *window* (Constant) beträgt der berechnete Wert 71 Minuten. Dieser Wert wird jedoch gemäß des Expertengrenzwertes auf 180 Minuten angehoben. Die Verteilung zur Berechnung des *thresh*-Parameters (Jump) zeigt ähnlich wie die Verteilung der Differenzen zweier nacheinanderfolgender Datenpunkte eine spitze, symmetrische Verteilung um Null, die zu beiden Seiten stark abfällt (Kurtosis = 5, Schiefe = 0,66). Die Parametergrenzen liegen außerhalb der Hauptverteilung.

Abbildung 6 zeigt die Verteilung der berechneten Parameter in Boxplots für alle Stationen. Die Abbildung teilt die Stationen pro Parameter in zwei Gruppen, basierend auf ihrer Messfrequenz: 10 und 60 Minuten. Die Stationen mit einer Messfrequenz von 10 Minuten weisen eine geringere Varianz für alle dargestellten Parameter auf als die Werte der Stationen mit 60-minütiger Messfrequenz (Tabelle 5). Zwischen den zwei Messfrequenzen ist für die statistisch ermittelten Parameter *window* (Constant) (ANOVA: p-Wert = 0,006) und *thresh* (Jump) (ANOVA: p-Wert = 0,012) der Unterschied signifikant.

Tabelle 5: Der Varianzunterschied der statistisch ermittelten Parameter unterteilt in Stationen mit 10 *min* und 60 *min* zeitlicher Auflösung

Parameter	Varianz bei 10 <i>min</i> Messfrequenz	Varianz bei 60 <i>min</i> Messfrequenz
<i>min</i> (Range (2))	0.49	3.28
<i>max</i> (Range (2))	0.13	2.4
<i>thresh</i> (Offset)	0.004	0.06
<i>window</i> (Constant)	6325	733931
<i>thresh</i> (Jump)	0.002	0.005

Für die Parameter *thresh* (Offset) und *window* (Constant) sind die Expertengrenzen in Abbildung 6 farblich markiert. Bei den Stationen mit 10-minütiger Messfrequenz liegen die Parameterwerte für den *thresh*-Parameter (Offset) alle innerhalb der Expertengrenzen. Im Gegensatz dazu liegen die berechneten Parameterwerte bei 7 von 8 Stationen mit 10-minütiger Messfrequenz für den Parameter *window* (Constant) unterhalb der gegebenen Untergrenze. Bei den Stationen mit 60-minütiger Messfrequenz liegen die Werte bei 9 von 12 Stationen für beide Parameter innerhalb der Expertengrenze. Jeweils eine Station liegt darüber und zwei Stationen darunter.

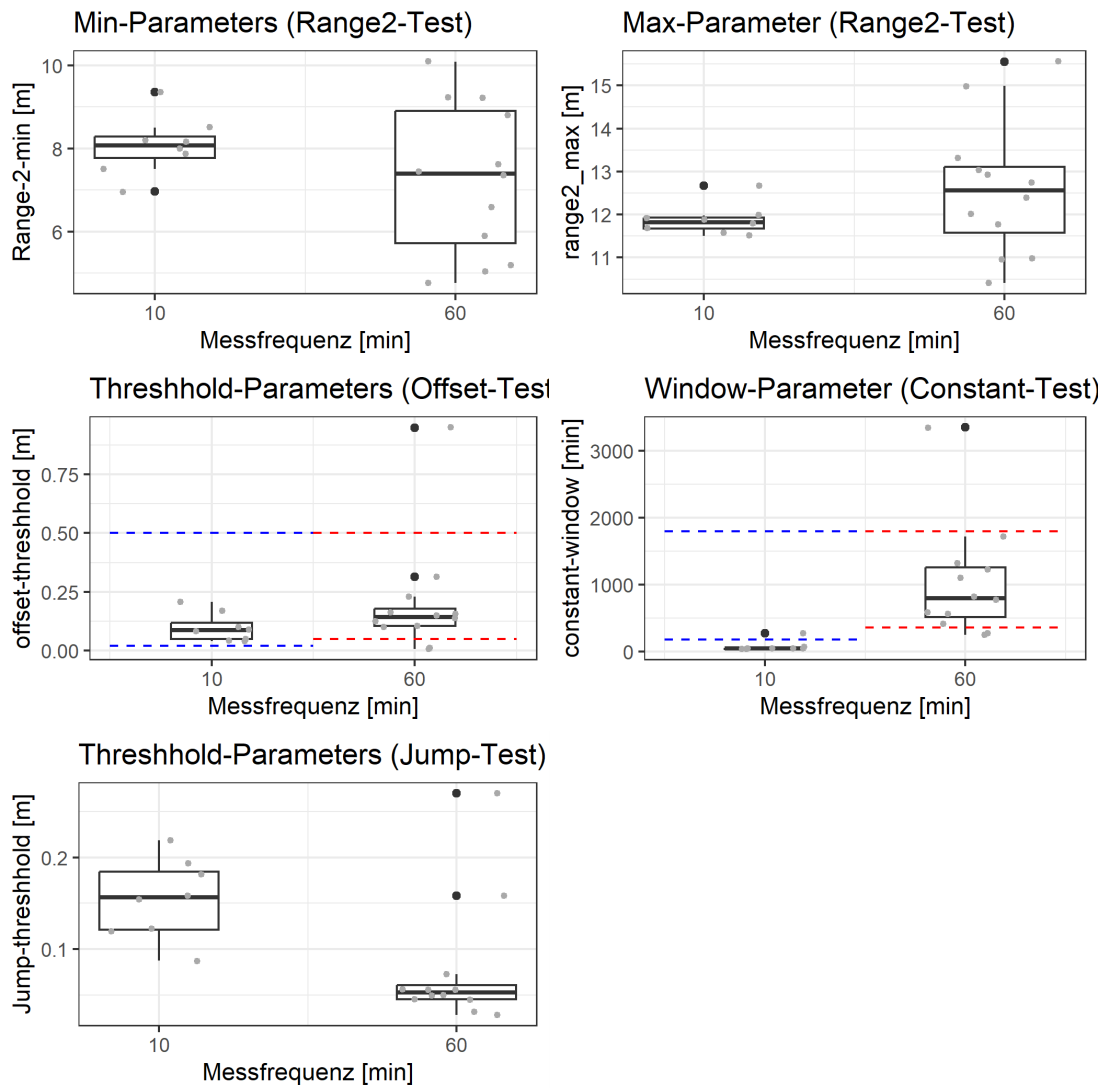


Abbildung 6: Boxplots der ermittelten Parameter aller Stationen mit farbig markierten Expertengrenzen.

### 4.1.3 Anomalieerkennung

Der Anteil der geflaggtten Datenpunkte ist für alle Stationen in Abbildung 7 dargestellt und in Abbildung 8 im Bezug zur Testererkennung untergliedert. Hierbei ist der relative Anomalieanteil an allen duplikatbereinigten Daten einer Zeitreihe zwischen den 10min (Mittelwert = 1,6 %) und 60min (Mittelwert = 4,8 %) signifikant verschieden (Anova: p-Wert = 0,00554).

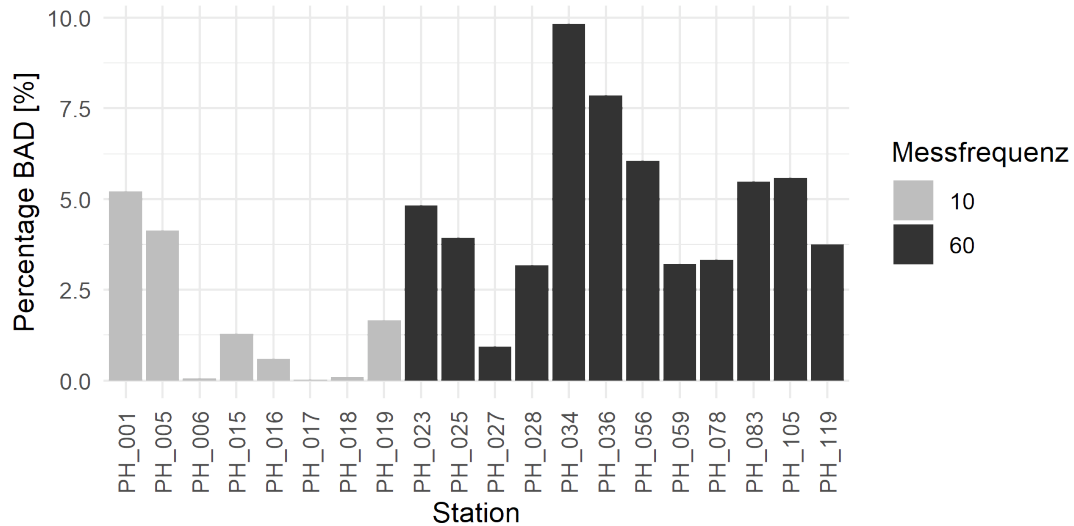


Abbildung 7: Prozentualer Anteil der als Anomalie markierten Punkte an den Duplikat-bereinigten Daten



Abbildung 8: Anteil aller Tests an der Gesamtanzahl der Flags pro Station

Bei der 60-min-Messfrequenz erkennt der Constant Test (Mittelwert = 90%) im Gegensatz zur 10-min-Messfrequenz (Mittelwert = 3,3%) die meisten Anomalien.

Bei der 10-min-Messfrequenz erkennen der LOF und Offset Test (Mittelwert = 54,2%, 48,9%) die meisten Anomalien. Die wenigsten Anomalien erkennen bei den 10-min-Messfrequenzen der Constant-Test und der Jump-Test (Mittelwert: 3,3%, 6,3%). Für die 60-min-Messfrequenzen markiert der Offset- und LOF-Test die wenigsten Datenpunkte als Anomalien (Mittelwert: 1,5%, 2,8%). Zusammenfassend zeigt Abbildung 9, dass unabhängig von der Messfrequenz im Mittel 12,4% Datenpunkte von mehreren Tests markiert werden. Der flagstärkste Test ist mit 55,4% aller Flags der Constant- und der flagschwächste mit 7,6% der Jump-Test.

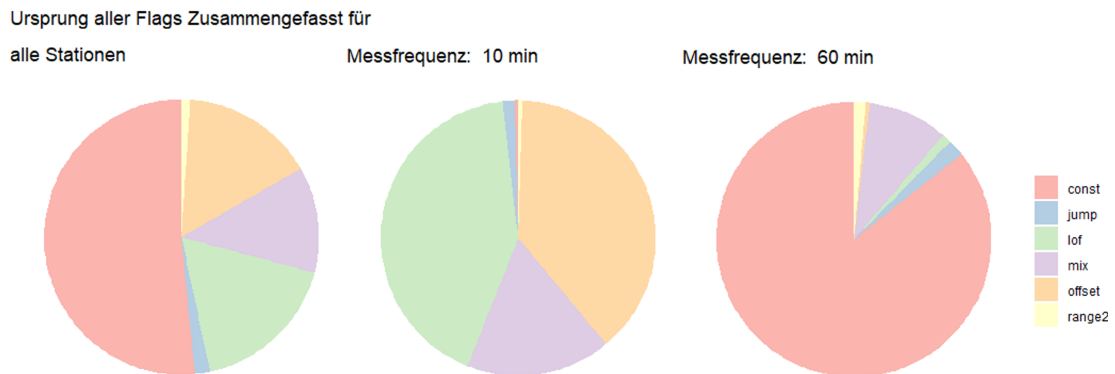


Abbildung 9: Anteil aller Tests an der Gesamtanzahl der Flags zusammengefasst auf alle Stationen und unterteilt in die Stationen mit 10-, bzw 60-minütiger Messfrequenz. Die Kategorie "Mix" bezieht sich auf Punkte, welche von mehr als einem Test geflagged wurden.

Abbildung 10 zeigt die Zeitreihen der Rohdaten mit Anomalie-Markierungen und die flagbereinigten Zeitreihen der Stationen PH 028, PH 016 und PH 017. Für die Stationen PH 016 und PH 028 werden Anomalien durch alle sechs QC-Tests erkannt. Bei PH 017 werden durch den Constant-Test keine Anomalien markiert. Bei PH 016 werden über die gesamte Zeitreihe hinweg durch den Offset- und LOF-Test viele Punkte markiert. Die QC-Test bereinigten Zeitreihen der Stationen PH 016 und PH 017 sind nahezu identisch. Für die bereinigte Zeitreihe von Station PH 028 lässt sich ebenfalls ein ähnliches Verhalten des Grundwassers erkennen, jedoch ist die Grundstruktur weniger variabel. Für alle anderen 17 Stationen ist die automatische QC, inklusive der bereinigten Zeitreihen, im Anhang auf Abbildungen 23 bis 40 dargestellt.

## 4.2 Anomalieerkennung mit Testparameterunsicherheit

Die Variabilität der relativen Anomalieerkennung im Unsicherheitsbereich ( $\pm 20\%$ ) ist in Abbildung 11 dargestellt. Im Mittel zeigen 44% der Stationen einen signifikanten linearen Trend im Unsicherheitsbereich der Testparameter. In der weiteren Untergliederung in die spezifischen Testparameter (Tabelle 6) zeigen über 80% der Stationen einen signifikanten linearen Trend bei *max* (Range (2)), *window* (Constant) und *thresh* (LOF). Für die anderen Testparameter liegt der Stations-

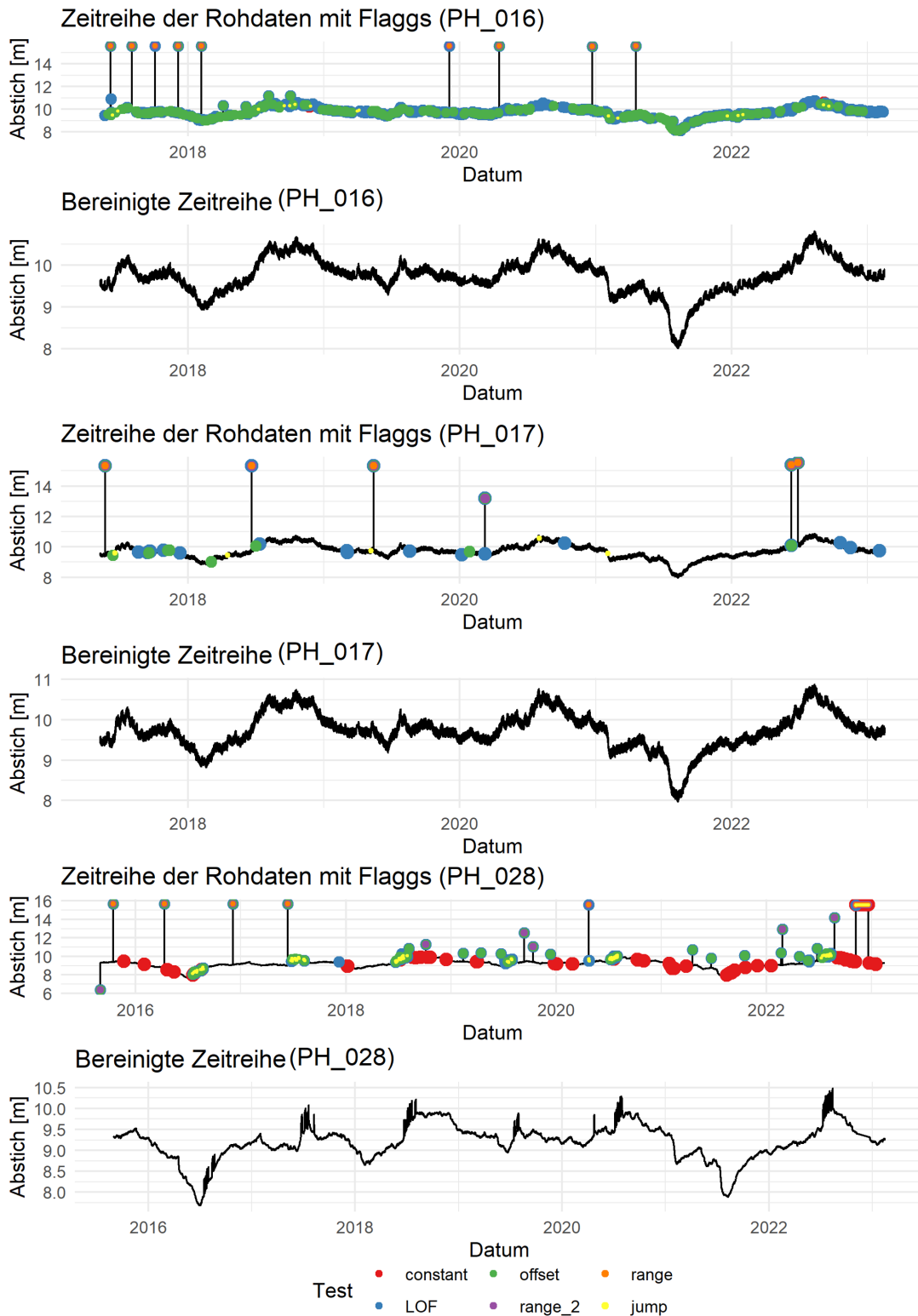


Abbildung 10: Geflaggte und flagbereinigte Zeitreihe der Stationen PH 016, PH 017 und PH 028

anteil mit signifikantem linearen Trend zwischen 15% und 45%. Für die Parameter *thresh* (Jump) und *window* (Constant) ist optisch für die meisten Stationen ein konstanter, linearer Abfall innerhalb des Unsicherheitsbereiches (zwischen -20% und +20%) erkennbar. Bei den Parametern *max* (Range (2)), *thresh* (LOF) und *thresh* (Offset) liegt die Änderung der Anomalieerkennung hauptsächlich im Unsicherheitsbereich von -20 bis -10% des statistisch ermittelten Parameterwerts. Insgesamt ist eine Sensitivität in der Anomalieerkennung durch die Parameterunsicherheit gegeben.

Tabelle 6: Anzahl der Stationen (von insgesamt 20), welche bei Änderung eines Parameters einen signifikanten Trend aufweisen.

Test	Parameter	Anzahl der Stationen mit signifikantem Trend
Range (2)	<i>min</i>	4
	<i>max</i>	18
Offset	<i>thresh</i>	9
	<i>tolerance</i>	8
	<i>window</i>	4
Constant	<i>window</i>	16
Lof	n	4
	<i>thresh</i>	17
Jump	<i>window</i>	3
	<i>thresh</i>	5

Abbildungen 12 visualisieren, wie oft ein Datenpunkt in der Unsicherheitsanalyse von einem Test maximal als Anomalie markiert wurde. Bei allen Validierungsstationen dominieren der Offset- und LOF-Test. Bei PH 028 zeigen sie ein Muster mit Konzentrationen um die Sommermonate in 2016 bis 2020 und 2022. Neben den markanten Ausreißern markiert Range (2) in PH 016 und PH 017 vier zeitgleiche Cluster um außergewöhnlich hohe (Winter 2018/19, 2020/21 und 2022/23) und niedrige (Sommer 2021) Abstichwerte. Bei PH 028 werden konstant Anomalien durch Range (2) erkannt, wobei auch hier die gleichen Cluster wie bei PH 016 und PH 017 erkannt werden.

### 4.3 Vergleich zwischen automatischer und Experten-Qualitätskontrolle

Abbildung 13 zeigt die durch die Experten geflaggt, duplikatbereinigten Zeitreihen. Es wird unterschieden, ob ein Datenpunkt durch keinen, einen, zwei oder drei Experten als Anomalie markiert wurde. Es ist erkennbar, dass die Experten bei groben Anomalien gleichermaßen flaggen, bei weniger klaren Anomalien jedoch oft keine Übereinstimmung auftritt. Es ist außerdem zu erwähnen, dass nur zwei der drei Experten Duplikate in den Zeitstempeln entdeckt haben.

Alle Datenpunkte, die bei der manuellen Kontrolle von mindestens einem Experten als Anomalie markiert wurden, werden auf Abbildung 14 mit allen Datenpunkten verglichen, die durch den automatischen Workflow geflaggt wurden. Wie auch schon beim Vergleich zwischen den drei Experten fällt auf, dass die

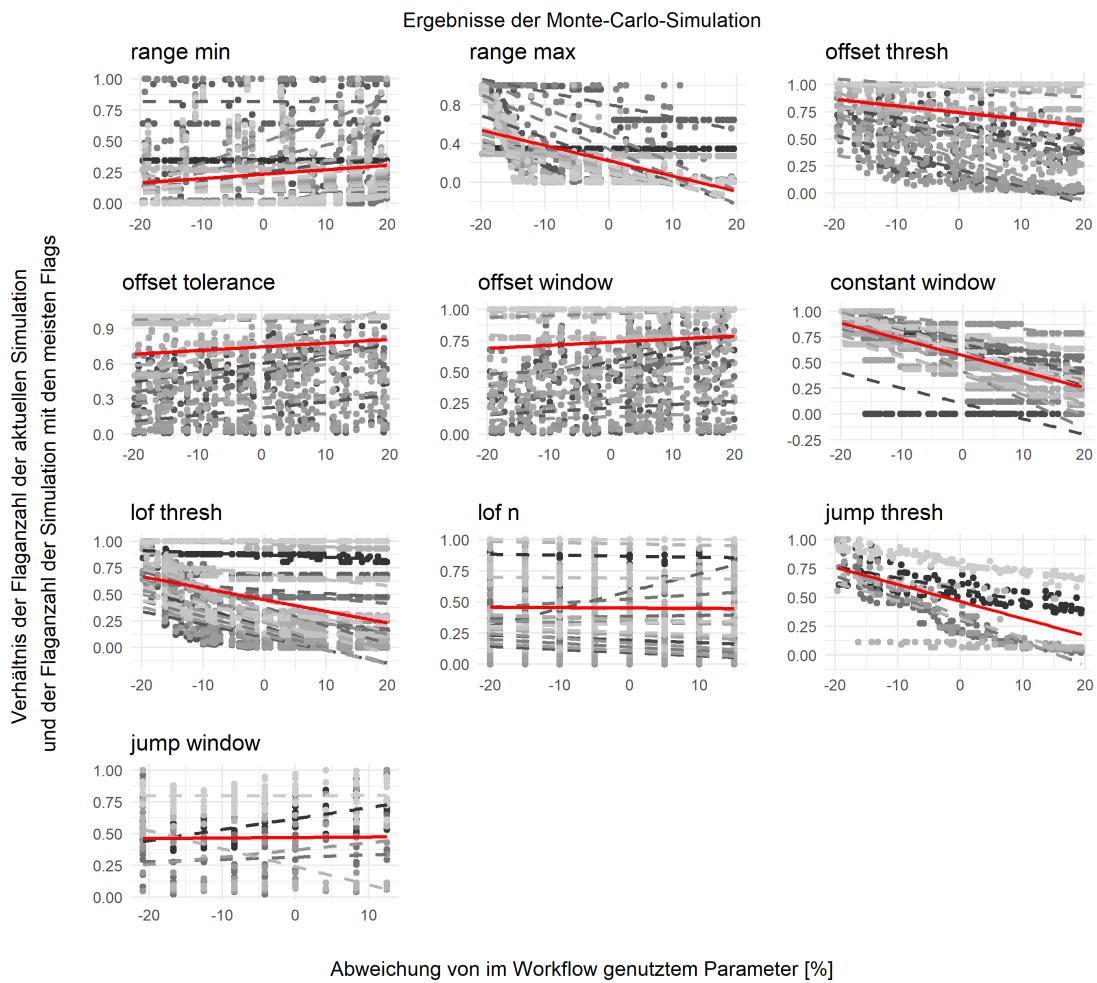


Abbildung 11: Ergebnisse der Monte-Carlo-Analyse pro Parameter, mit der Unsicherheit (Prozentuale Abweichung von dem im Workflow genutzten Parameter) auf der x-Achse und dem Verhältnis der Flaganzahl der aktuellen Simulation zu der Flaganzahl der Simulation mit den meisten Flags auf der y-Achse. Der übergeordnete lineare Trend ist in rot dargestellt.





Abbildung 12: Anomalien, welche durch mindestens einen Test bei allen MC-Simulationen als Anomalie markiert wurden (jeweils oben) und die maximale Anzahl an Flags bei 100 Monte-Carlo-Simulationen durch einen Test (jeweils unten). Bei mehreren Tests mit der gleichen Anzahl an Anomaliezeichnungen wird folgendermaßen priorisiert: Range (2), Offset, LOF, Constant, Jump.

automatisch markierten Flags und die von mindestens einem Experten markierten Datenpunkte bei groben Anomalien, die durch Range (1&2) erkannt wurden, übereinstimmen. Es können jedoch auch viele Datenpunkte ausgemacht werden, die entweder nur durch den automatischen Workflow oder nur durch die Experten erkannt werden.

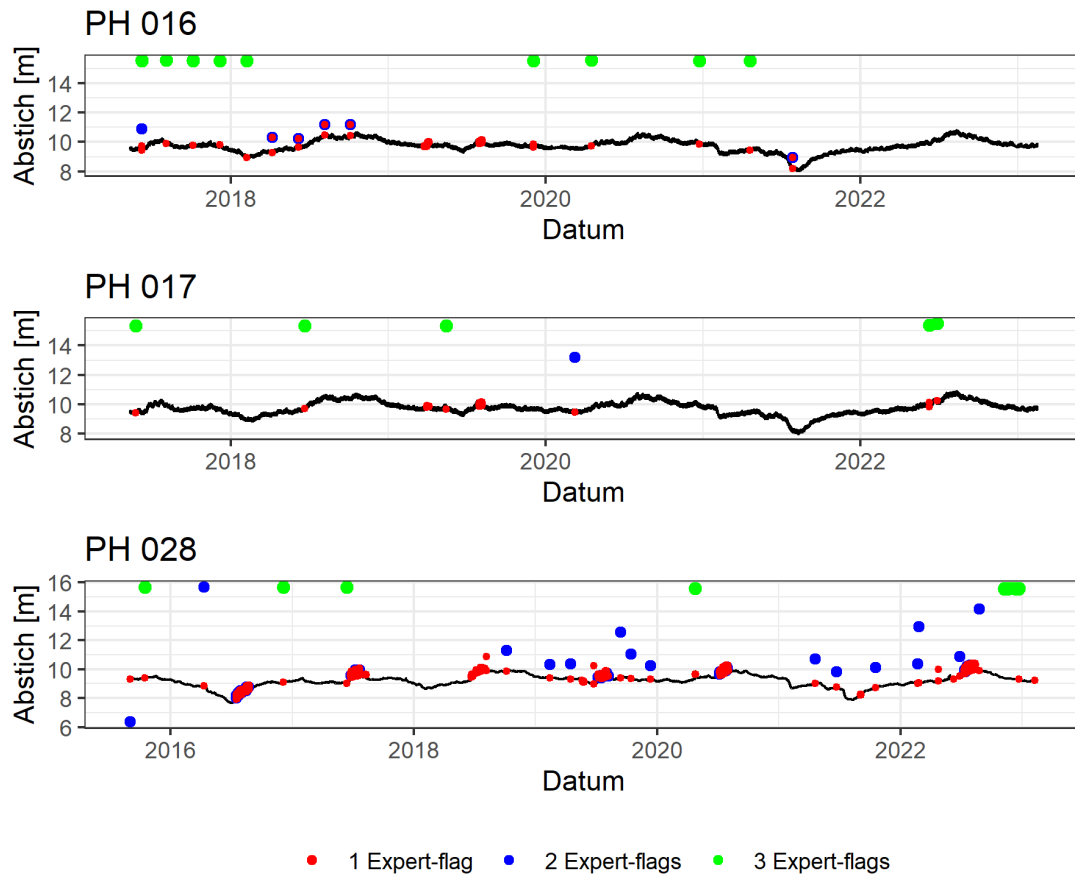


Abbildung 13: Expertenflags für alle drei duplikatbereinigten Validierungszeitreihen, farblich markiert je nach Anzahl der Experten, die den Punkt markiert haben

Die Confusionmatrix in den Tabellen 7 bis 9 quantifiziert diese Übereinstimmungen und Unterschiede. In allen drei Validierungszeitreihen werden über 93% der Datenpunkte sowohl durch die manuelle als auch durch die automatische Prüfung als *GOOD* markiert. Diese Übereinstimmung in den True *GOOD* Fällen wird auch durch die Spezifität von über 0.98 bei allen drei Stationen bestätigt. Der Recall-Wert liegt im Mittel bei 0.1272 und die Präzision bei 0.3196. Die Kennzahlen deuten darauf hin, dass einige Punkte im Vergleich zu den Experten zu viel markiert werden (*FP*), und dass noch mehr Punkte durch den Workflow nicht erkannt werden, welche durch mindestens einen Experten geflaggt sind (*FG*). Im Vergleich zwischen den automatisch markierten Datenpunkten und denen, die von mindestens zwei von drei Experten markiert wurden, ändert sich dieses Bild deutlich. Ein einzelner der drei Experten ist alleine für durchschnittlich 89% aller Markierungen verantwortlich. Somit reduziert sich die *FALSE GOOD* Klasse bei den Stationen PH 016 und PH 017 auf Null. Bei der Station PH 028 bleiben nur 71 von ursprünglich 2495 Markierungen übrig, die durch mindestens zwei Ex-

perten, nicht aber durch den automatischen Workflow als Anomalie identifiziert wurden. Diese Punkte liegen innerhalb von Zeiträumen mit hoher Variabilität, die in Kapitel 4.4 als Ereignisse der Wasserentnahme zu Berechnungszwecken identifiziert werden. Durch diese andere Betrachtungsweise der Expertenflags steigt der Recall im Mittel auf 0.98 und die Präzision fällt leicht auf 0.271. Diese Werte unterstreichen, dass der Workflow einen sehr hohen Anteil aller Anomalien erkennt, welche auch von der Mehrheit der Experten als solche gewertet werden. Gleichzeitig produziert die automatische QC aber auch Flags, welche durch die Experten nicht bestätigt werden. Der Anteil aller als *FALSE BAD* von allen als Anomalie markierten Punkten liegt zwischen 46% und 99% (Mittelwert = 73%). Der mittlere F-Score steigt von 0.1675 auf 0.375, was eine höhere Übereinstimmung des Workflow mit den mehrheitlich geflaggtten Datenpunkten durch die Experten unterstreicht. Alle Kennzahlen sind im Anhang in Tabelle x dargestellt.

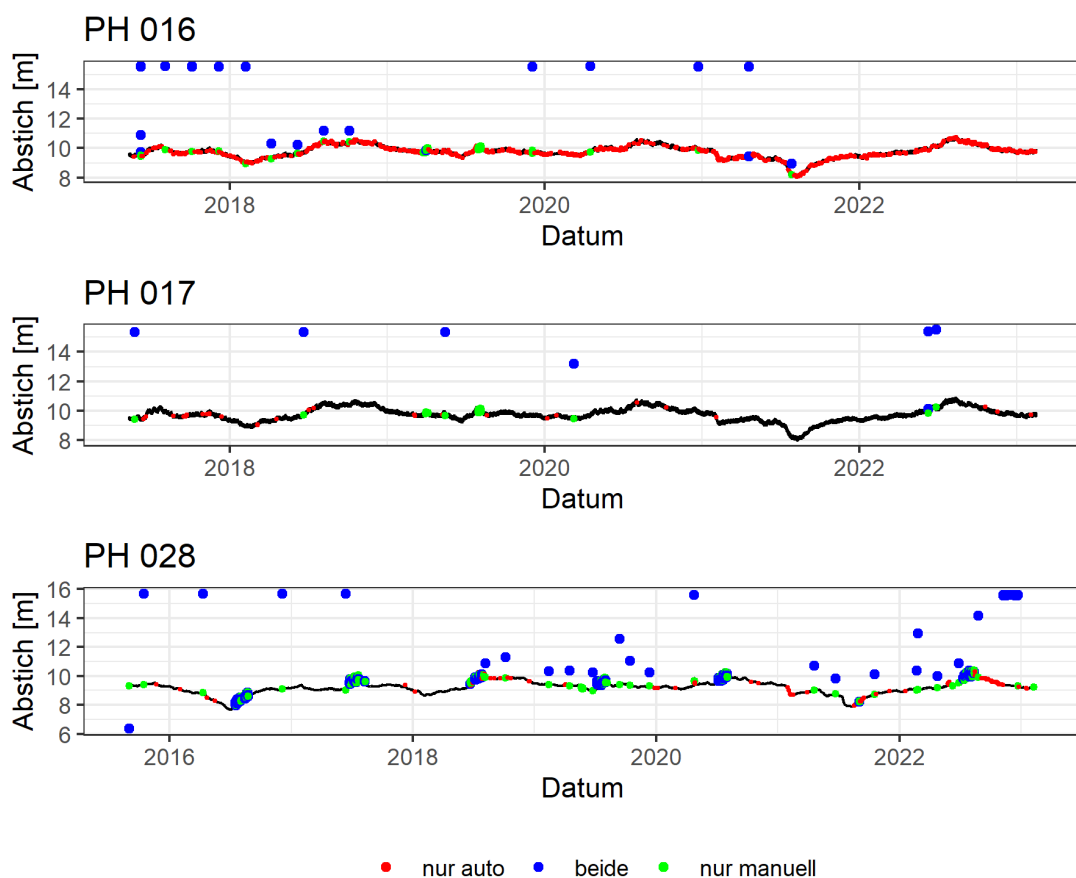


Abbildung 14: Validierungszeitreihen mit den Datenpunkten farbig markiert, welche nur durch das automatische, nur durch das manuelle oder durch beide Verfahren erkannt wurden

Tabelle 7: Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 016

PH 016	Manuelle Flags			
		GOOD	BAD	All
Automatische Flags	GOOD	298848	2692	301540
	BAD	1772	34	1806
	All	300620	2726	302347

Tabelle 8: Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 017

PH 017	Manuelle Flags			
		GOOD	BAD	All
Automatische Flags	GOOD	301494	1840	303334
	BAD	56	21	77
	All	301550	1861	303411

Tabelle 9: Confusionmatrix der Duplikat-bereinigten Validierungszeitreihe PH 028

PH 028	Manuelle Flags			
		GOOD	BAD	All
Automatische Flags	GOOD	60823	2495	63318
	BAD	693	1390	2083
	All	61516	3885	65401

Abbildung 15 zeigt, durch welche Tests die Markierungen im automatischen Workflow entstanden sind, wenn sie:

1. auch manuell durch die Experten entdeckt wurden und
2. wenn sie ausschließlich durch den Workflow markiert wurden.

Bei allen Stationen markieren die Tests LOF, Offset und Range (1&2) Datenpunkte, die auch durch die Experten markiert wurden. Die Tests Constant, LOF und Offset sind hauptverantwortlich für die Kategorie *FALSE BAD*. Es kann eingeordnet werden, dass Flags, die sowohl durch den automatischen als auch durch den manuellen Workflow entstanden sind, für die Stationen PH 016 und PH 017 durch die Tests LOF, Offset, Range (1) und Range (2) erkannt wurden. Dabei erkennt der LOF-Test alleine im Mittel dieser beiden Stationen 96% aller Überschneidungen zwischen Experten und Workflow. Außerdem wurden 5 Datenpunkte als Duplikat-2 markiert. Bei Station PH 028 sind die Tests Constant (48%), Range (1) (76%) und Range (2) (78%) hauptverantwortlich für Überschneidungen der automatischen und manuellen Flags. Aber auch der LOF- und Offset-Test markieren zwischen 9 und 11 % der Datenpunkte, die auch durch mindestens einen Experten markiert wurden.

Datenpunkte, die ausschließlich durch den automatischen Workflow als Anomalien erkannt wurden, wurden bei den zwei Stationen mit 10-minütigen Messfrequenzen (PH 016 und PH 017) zu 52% durch den LOF- und zu 44% durch den Offset-Test markiert. Bei der Station mit 60-minütiger Messfrequenz (PH 028) markierte der Constant-Test mit 80% die meisten Datenpunkte als Anomalien,

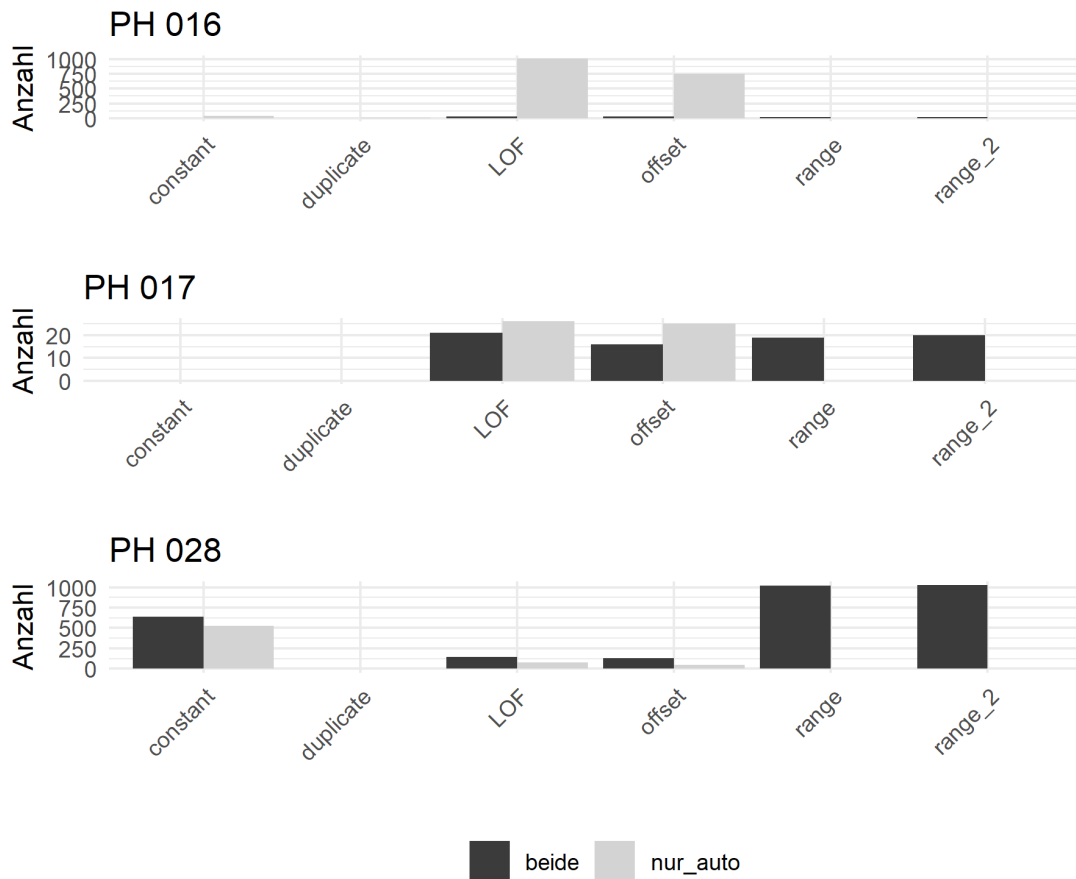


Abbildung 15: Test-Ursprung bei Punkten, welche nur durch die automatischen Tests bzw. durch diese und die Experten erkannt wurden.

die nicht durch die Experten markiert wurden.

Die Tabellen 10 bis 12 vergleichen die manuellen und automatischen Flags, differenziert zwischen der Anzahl der Markierungen durch die Experten und der Anzahl der Markierungen der Monte-Carlo-Durchläufe. Bei allen Stationen werden alle Anomalien, die von allen drei Experten markiert wurden, auch in über 66% der Durchläufe erkannt. Für PH 016 und PH 017 kann zusammengefasst werden, dass ca. 99% aller Flags durch einen Experten entstanden sind. Auch bei PH 028 überwiegen die Flags durch einen einzelnen Experten mit 69%. In der Analyse aller Validierungsstationen zeigte sich, dass Datenpunkte, die von allen drei Experten einstimmig als Anomalien identifiziert sind, konsistent auch in der Kategorie *bis 3/3* aller 100 Monte-Carlo-Simulationen als Anomalien detektiert werden. Für PH 016 und PH 017 ist das auch für alle Datenpunkte der Fall, die durch zwei Experten markiert wurden. Datenpunkte, die nur durch einen Experten als Anomalie markiert wurden, werden bei PH 016 und PH 017 in 98% von keinem Test in keiner Simulation markiert. Bei PH 028 werden alle Datenpunkte in mindestens einer Simulation durch mindestens einen Test erkannt. Durch einen Blick auf Abbildung 12 wird deutlich, dass der Ursprung dieser vielen Markierungen der Range-Test ist.

Tabelle 10: Multiclass-Confusionmatrix der Validierungszeitreihe PH 016 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden.

PH 016		Manuell				
		GOOD	BAD 1/3	BAD 2/3	BAD 3/3	All
Auto	GOOD	252750	2633	0	0	255383
	bis 1/3	41998	56	0	0	42054
	bis 2/3	5048	4	0	0	5052
	bis 3/3	825	8	6	19	858
	All	297621	2701	6	19	303347

Tabelle 11: Multiclass-Confusionmatrix der Validierungszeitreihe PH 017 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden.

PH 017		Manuell				
		GOOD	BAD 1/3	BAD 2/3	BAD 3/3	All
Auto	GOOD	263404	1813	0	0	265217
	bis 1/3	36678	27	0	0	36705
	bis 2/3	1450	0	0	0	1450
	bis 3/3	12	1	1	19	33
	All	300544	1841	1	19	302405

Tabelle 12: Multiclass-Confusionmatrix der Validierungszeitreihe PH 028 mit der Unterscheidung zwischen Datenpunkten, welche von einem, zwei oder drei Experten geflaggt wurden und dem gegenüber Datenpunkte, welche durch den flaggstärksten Test der Monte-Carlo-Simulation in bis zu 1/3, 2/3 und 3/3 der Simulationen geflagged wurden.

PH 028		Manuell				
		GOOD	BAD 1/3	BAD 2/3	BAD 3/3	All
Auto	GOOD	0	0	0	0	0
	bis 1/3	59117	1878	59	0	61054
	bis 2/3	1959	654	22	0	2635
	bis 3/3	440	159	53	1060	1712
	All	61516	2691	134	1060	65401

#### 4.4 Anomalien im Umweltsystem Kontext

Die einzelnen Markierungen der Experten und des automatischen Workflows werden im Kontext des Umweltsystems und anthropogenen Einflüssen betrachtet, um Zusammenhänge der Anomalieerkennung mit übergeordneten Events aufzuzeigen. Abbildung 16 zeigt beispielhaft drei Muster, die in allen Validierungsstationen vorhanden sind:

1. Grobe Ausreißer werden von manueller und automatischer Qualitätsprüfung gleichermaßen identifiziert. Mindestens ein Experte (grün) markiert die Punkte jeweils nach einem Ausreißer (blau) ebenfalls als Anomalie. (PH 016 Beispiel)
2. Ein einzelner Experte markiert Datenpunkte, die sich optisch nicht von den benachbarten Datenpunkten unterscheiden (PH 017 Beispiel zwischen dem 01.08.2019 und dem 07.08.2019)
3. Die manuelle und automatische Qualitätskontrolle erkennen Einflüsse durch Grundwasserentnahmen nicht eindeutig (PH 028 Beispiel).

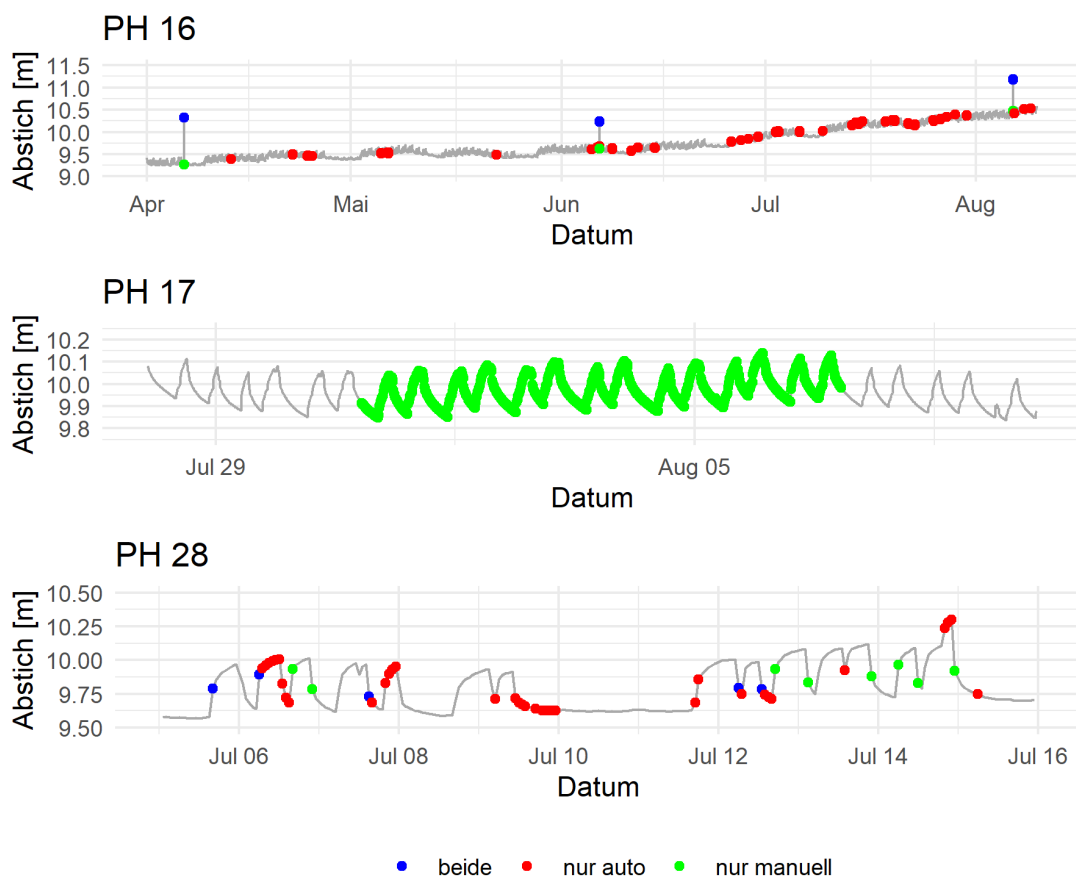


Abbildung 16: Teilzeitreihen der Validierungsstationen mit Anomaliezeichnungen der Experten und des Workflows zur Analyse einzelner Events.

In allen Stationen mit 10-min-Messfrequenz und bei PH 028 sind üblicherweise bestimmte Muster bei Wasserentnahmen zu erkennen. Diese Stationen liegen zwischen 21 und 430 Metern von einem Entnahmebrunnen entfernt (Mittelwert = 136 Meter). Die Entnahmephasen sind durch zunächst starke, plötzliche und dann abflachende Anstiege der Abstichmessungen charakterisiert. Anschließend ist ein schneller Rückgang zum Ursprungsniveau erkennbar (Retike et al., 2022). Eine solche Phase des Absinkens und wieder Anschwellen des Grundwasserspiegels erstreckt sich in der Regel über einen Zeitraum von einigen Stunden. Die Stationen PH 016 und PH 017 weisen diese Muster das gesamte Jahr über auf, während sie bei PH 028 nur phasenweise erkennbar sind. In Rücksprache mit der Badenova AG ist bekannt, dass der 21 Meter entfernte Brunnen zur Station PH 028 ein Beregnungsbrunnen ist und in den Sommermonaten Grundwasser zur Bewässerung der umliegenden Agrarflächen entnimmt. Weder die automatische noch die manuellen Anomalieerkennungen scheinen die Muster der Wasserentnahmen gut einordnen zu können und markieren scheinbar zufällig Werte innerhalb dieser Phasen. Die Entnahmemuster treten bei den 10-min-gemessenen Stationen täglich auf und prägen die Dynamik der Zeitreihen. Bei den Stationen mit 60-min-Messfrequenz können diese Muster grundsätzlich nur vereinzelt in den Sommermonaten erkannt werden.

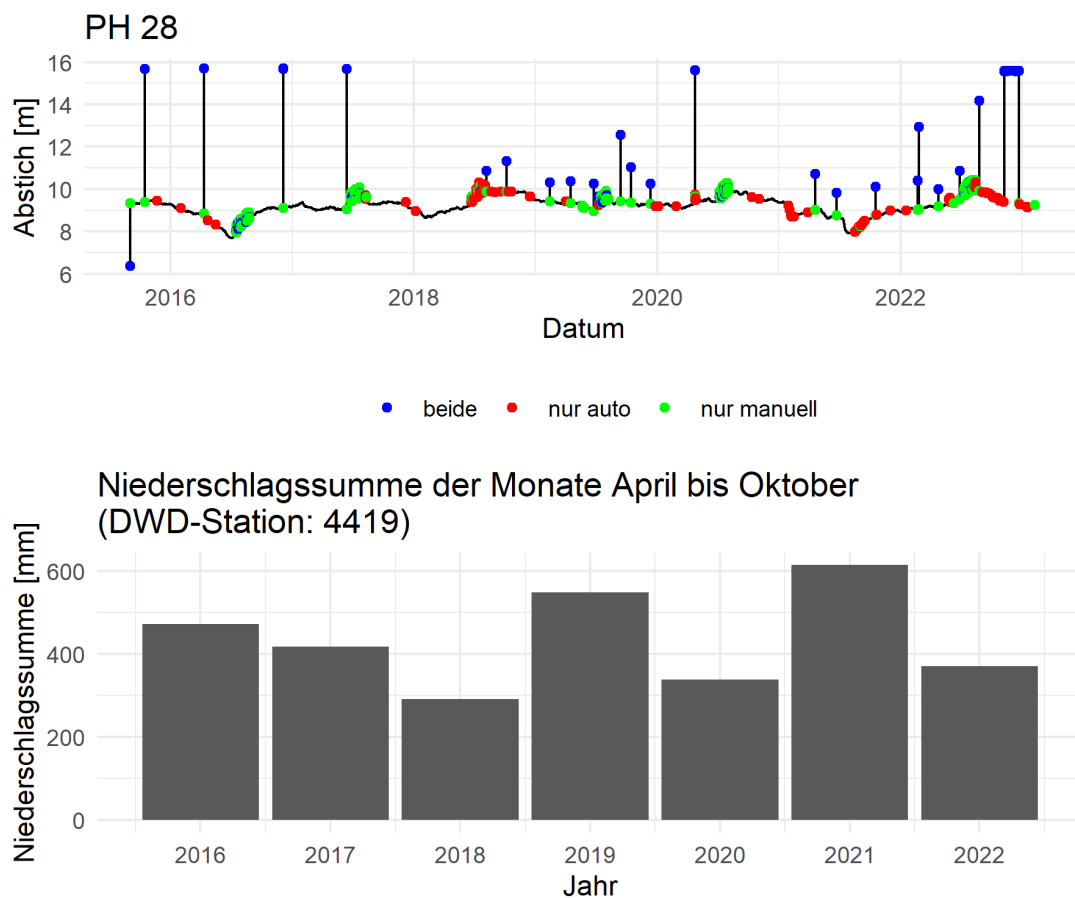


Abbildung 17: Zeitreihe der Rohdaten der Station PH 028 mit Markierungen durch die Experten und den manuellen Workflow und Niederschlagssummen der Monate April bis Oktober für die Jahre 2016 bis 2022



Die Zeitreihe der Station PH 028, inklusive der Markierungen durch die Experten und den automatischen Workflow, wird in Abbildung 17 mit den Niederschlagssummen der Monate April bis Oktober für die Jahre 2016 bis 2022 verglichen. In den Sommermonaten der Jahre 2016 bis 2020 und im Jahr 2022 können die oben genannten Entnahmestrukturen mit hoher Variabilität erkannt werden. Das Jahr 2022 weist diese Strukturen nicht auf. Gleichzeitig ist dieses Jahr das mit der höchsten Niederschlagssumme in den Monaten April bis Oktober, was den Rückschluss zulässt, dass in diesem Jahr weniger Grundwasser zur Bewässerung benötigt wurde.

Bei den Stationen PH 006 und PH 083 wurden durch den automatischen Workflow Sprünge in den Zeitreihen erkannt. Da manuelle Handmessungen dieser Stationen nur bis vor den Sprüngen verfügbar sind, ist ein Vergleich hier nicht möglich. Um grundsätzlich den Nutzen eines Vergleichs zwischen Logger- und Handmessungen zu untersuchen, werden die Zeitreihen beider Messverfahren in Abbildung 18 dargestellt. Der Vergleich zeigt, dass die Schwankungen im Jahresverlauf zwischen manuellen und Logger-Messungen gut übereinstimmen, aber kurzfristige Schwankungen nicht validiert werden können. Für einzelne Ausreißer ist diese Übereinstimmung nicht gegeben. Bei PH 028 ist außerdem zu erkennen, dass der Abstichwert der Handmessungen ca. zwei Meter unter den Werten der automatischen Messung liegt. Der Grund hierfür liegt möglicherweise an einer falschen Messoberkante. Auch Retike et al. (2022) nutzen den Vergleich zwischen manuellen und automatischen Messungen, um Sprünge in den Zeitreihen zu erkennen.

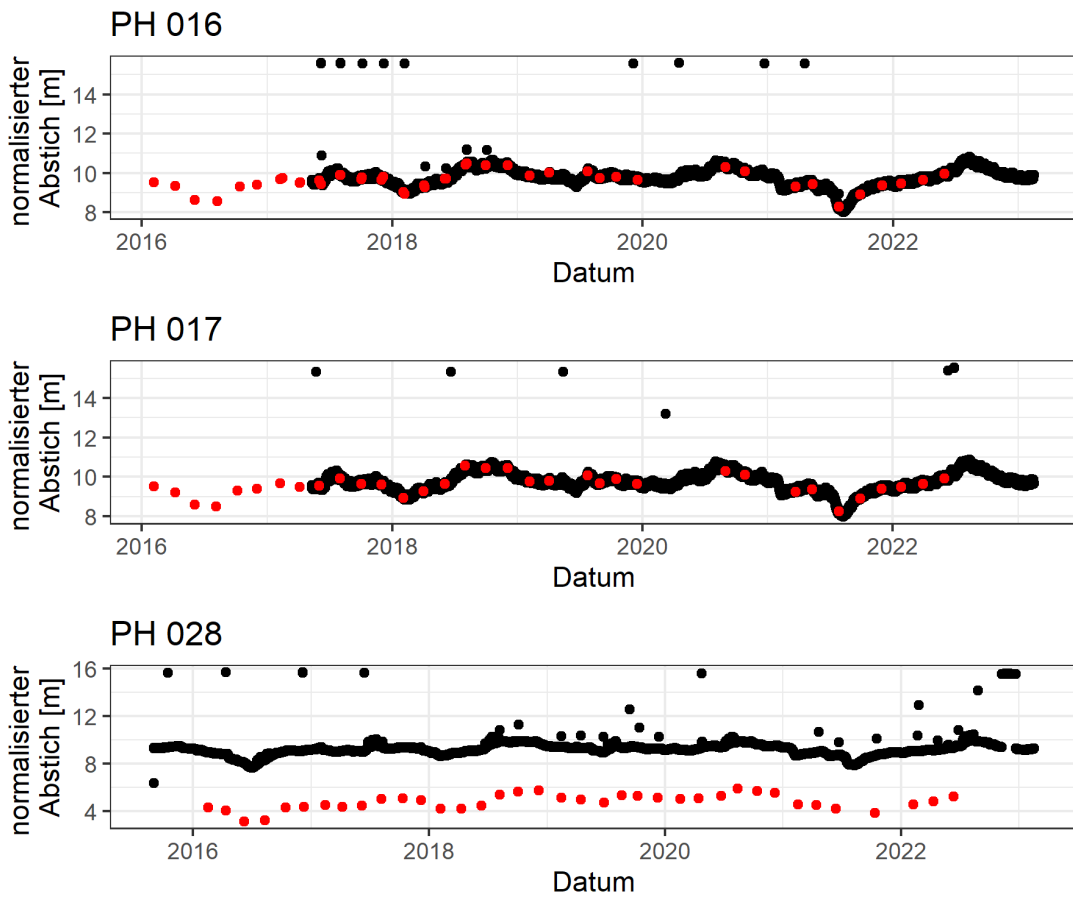


Abbildung 18: Manuelle Abstichmessungen im Vergleich zu den automatischen Loggermessungen für die Stationen PH 006 und PH 083

## 5 Diskussion

### 5.1 Anomalieerkennung mit Testparameterunsicherheit

Die durchgeführte Monte-Carlo-Simulation beschreibt die Sensitivität der Anomalieerkennung gegenüber Unsicherheiten in den Testparametern. Es wird deutlich, dass nicht alle Tests gleich sensibel gegenüber Unsicherheiten reagieren. Während einige Parameter einen minimalen Einfluss auf die Anomalieerkennung haben, sind andere für den Output des Workflows von entscheidender Bedeutung. *max* (Range (2)), *window* (Constant) und *thresh* (LOF) zeigen im Unsicherheitsbereich bei über 80% der Stationen eine signifikante Änderung der Anomalieerkennung. Verschiebungen der Parameter *thresh* und *tolerance* (Offset) innerhalb des Unsicherheitsbereichs führen ebenfalls teilweise zu signifikant anderen Anomalieerkennungen. Bei der 10-min-Messfrequenz ist der statistisch ermittelte *window* (constant) durch den Expertengrenzwert korrigiert, sodass constant weniger Anomalien als bei der 60-min-Messfrequenz ohne Grenzwertkorrektur erkennt. Somit ist für die automatische Anomalieerkennung die Parameterberechnung und die Expertengrenzwahl relevant, deren korrekte Ermittlung auch im wissenschaftlichen Diskurs uneindeutig ist (Antonetti und Zappa, 2018, Taylor und Loescher, 2013).

Allgemein reduziert die statistische Parametrisierung der Tests subjektive Entscheidungen, trifft aber die Annahme einer Normalverteilung der Daten und benötigt Expertengrenzen. Bei der in dieser Arbeit gewählten Berechnungsvorschrift der Testparameter wird eine gewählte Anzahl an Standardabweichungen zu dem Mittelwert einer Verteilung addiert, die im Diskurs zwischen 2 und 6 beträgt (Shulski et al., 2014; Hubbard et al., 2005). Taylor und Loescher (2013) und Shulski et al. (2014) argumentieren, dass diese willkürliche Wahl an Standardabweichungen besser nachvollziehbar, replizierbar und anpassbar ist als die manuelle Expertenkontrolle, welche einen deutlich größeren Anteil an Willkür mit sich bringt. Dennoch kann die in dieser Arbeit getroffene Annahme einer Normalverteilung der Daten statistisch nicht bestätigt werden. Um diese Annahme zu umgehen, beziehen sich Taylor und Loescher (2013) auf den Zentralen Grenzwertsatz und ziehen eine konstruierte Stichprobenverteilung heran.

Vrugt et al. (2008) und Kavetski et al. (2006) nutzen bekannte Optimierungstechniken wie die Bayesianische Statistik und die Monte-Carlo-Simulation mittels Markov-Ketten, um die Testparameter im Bezug auf die gewünschten Ergebnisse zu optimieren. Die Nutzung dieser Techniken ist auch möglich, um die Anzahl der Standardabweichungen in den Formeln zur Parametrisierung der Tests zu optimieren. Dabei können die manuellen Expertenmarkierungen als Ziel der Optimierung gesetzt werden. Beim Optimierungsziel steht die Frage im Vordergrund, ob es wichtiger ist, alle tatsächlichen Ausreißer zu erkennen und dafür auch Markierungen zu akzeptieren, welche durch einen Experten als *FALSE BAD* identifiziert werden können, oder ob die Parameter lieber so gewählt werden, dass die *FALSE BAD*-Kategorie reduziert wird und die Tests möglicherweise in Folge daraus einige echte Anomalien nicht erkennen. Vor dem Hintergrund der Unsicherheiten der automatischen Qualitätskontrolle wird sie gegenwärtig nicht eigenständig, sondern zur Unterstützung der Expertenentscheidung verwendet (Arbesser et al., 2016; Gschwandtner und Erhart, 2018; Retike et al., 2022; Hollenberg et al., 2011).

## 5.2 Vergleich zwischen manueller und automatischer Qualitätskontrolle

Zur Beantwortung der Frage, ob ein automatischer Workflow die manuelle Kontrolle durch Experten nachbilden und verbessern kann, wurde ein Vergleich zwischen automatisch und manuell qualitätsgesicherten Daten durchgeführt. Die Ergebnisse geben Aufschluss über verschiedene Aspekte:

Obwohl alle drei Experten denselben Auftrag zur Qualitätskontrolle erhielten, kam es im Mittel bei lediglich 10 % der Fälle zu einer Übereinstimmung zwischen allen drei Experten. Dabei wurden 89 % der markierten Datenpunkte nur von einem Experten geflaggt. Auch Duplikate wurden nur von zwei von drei Experten erkannt. Diese Beobachtung unterstreicht, dass subjektive Expertenbeurteilungen bezüglich der Markierung von Anomalien nicht immer konsistent sind. Jeder Schritt der manuellen Qualitätskontrolle ist potenziell durch persönliche Vorurteile oder Bias beeinflusst (Polz et al., 2023). Um diesen Bias zu minimieren und eine bessere Nachvollziehbarkeit und Reproduzierbarkeit zu gewährleisten, braucht es präzisere Anweisungen und Richtlinien für die Experten sowie klarere Dokumentationsanforderungen während des Qualitätskontrollprozesses (Jones et al., 2018). Ebenso ist es wichtig, einen Verwendungszweck der Daten zu definieren, da die Datenqualität immer im Hinblick auf ihren Nutzen beurteilt wird (International Organization for Standardization, 2023). Auch bei der *Badenova AG* findet keine Dokumentation der Qualitätskontrolle durch die Experten statt. In Hinsicht auf das Kriterium *Nachvollziehbarkeit* und *Reproduzierbarkeit* besteht hier somit noch Verbesserungspotenzial.

Durch die inkonsistente Qualitätskontrolle durch Experten ist zu hinterfragen, ob die Wahl der *wahren Klasse* als alle Datenpunkte, welche von mindestens einem Experten markiert wurden, eine gute Wahl ist. Retike et al. (2022) setzt bei ihrem Ansatz zur manuellen Qualitätskontrolle daher auch auf die Kombination aus den Einschätzungen verschiedener Experten.

Datenpunkte, die von mindestens zwei von drei Experten als Anomalie identifiziert werden, werden auch durch die automatische Kontrolle in durchschnittlich 98 % der Fälle erkannt. Gleichzeitig sind durchschnittlich 73% aller durch den Workflow als Anomalie markierten Datenpunkte als *FALSE BAD* zu werten. Was sich durch einen hohen Recall (Mittelwert = 0.98) und einem weniger hohen Präzision Das Aufkommen von vielen falsch markierten Anomalien ist ein bekanntes Problem der automatischen Kontrolle (Kunkel et al., 2005; Schmidlin et al., 1995). Die nachträgliche Überprüfung durch einen geschulten Experten ist eine Möglichkeit, um Markierungen zu beurteilen, die möglicherweise zu viel geflaggt wurden (Taylor und Loescher, 2013; Arbesser et al., 2016). Obwohl diese Nachprüfung durch Experten ein Element der subjektiven Entscheidungsfindung einführt und die potenzielle Zeiteinsparung verringert, ist der Prozess dennoch effizienter als eine vollständig manuelle Prüfung (Taylor und Loescher, 2013). Die automatisierte Vorauswahl an Anomalien liefert nicht nur eine Entscheidungsgrundlage, sondern reduziert auch die Anzahl der zu prüfenden Datenpunkte.

Bei der Untersuchung der *TRUE BAD* und *FALSE BAD* konnte kein einzelner Test ausgemacht werden, welcher für besonders viele Unterschiede zu den manuellen Flags verantwortlich ist, während er keine *TRUE BAD* flaggt. Das zeigt, dass der Workflow keine redundanten Tests integriert.

Auffallend ist, dass Datenpunkte während spezifischer Ereignisse, wie zum Beispiel Wasserentnahmen, sowohl automatisch als auch manuell als Anomalien erkannt werden. Kritisch zu betrachten ist, dass die Wasserentnahmen einem regelmäßigen Verlauf folgen, beide Kontrollen diese Ereignisse jedoch nur unregelmäßig als Anomalie identifizieren. Zusätzlich unterscheiden sich die markierten Datenpunkte innerhalb dieser Events zwischen manueller und automatischer Kontrolle erheblich. Die in Zeitreihen auftretenden Muster, die durch Wasserentnahmen entstehen, werden in hydrologischen Studien häufig identifiziert (Ha et al., 2021; Rau et al., 2019). Je nach weiterer Nutzung der Daten können diese Muster entweder gelöscht oder ignoriert werden (Retike et al., 2022). Soll beispielsweise der Trend des Grundwasserspiegels über einen längeren Zeitraum betrachtet werden, sind Schwankungen innerhalb eines Tages vernachlässigbar. Bei der Analyse von genauen Wasserentnahmemengen sind diese Muster von großer Wichtigkeit und müssen in den Daten erhalten bleiben. Hier wäre eine Zweitkontrolle durch einen geschulten Experten eine valide Option (Durre et al., 2008).

### 5.3 Grenzen und Stärken des Workflows

Die Analyse des hier vorgestellten Workflows beschränkt sich auf 20 Messstationen der Badenova AG. Es wird keine Vergleichsanalyse mit Daten anderen Ursprungs vorgenommen. Dadurch lassen sich die gewonnenen Erkenntnisse nicht direkt auf andere Standorte skalieren. Hier liegt gleichzeitig auch die Möglichkeit einer anschließenden Forschung um die Wirksamkeit des Workflows auch in anderen hydrogeologischen Bereichen zu validieren.

Beim Vergleich des Workflows mit drei unabhängigen Experten kann den mehrheitlich bzw. einstimmig markierten Datenpunkten ein hohes Maß an Vertrauen entgegengebracht werden (Polz et al., 2023). Um jedoch die individuellen Vorgehensweisen der Experten nachvollziehen zu können, ist es ratsam, ihre Arbeitsschritte zu dokumentieren.

Bei der Wahl der Tests wurden keine Tests integriert, welche Zeitreihen auf die Anomalien Drift, Rauschen und Lücken untersuchen. Sowohl in den Empfehlungen der DWA als auch der WMO werden diese Tests nahegelegt, da Sensordaten oft von diesen Arten der Anomalien betroffen sind (World Meteorological Organization, 2021; Hollenberg et al., 2011). Darüber hinaus ist anzumerken, dass die Prüfungen im vorgestellten Arbeitsablauf durch eine strategische Abfolge bei der finalen Implementierung effizienter gestaltet werden können. Die Aneinanderreihung verschiedener Tests zur Erkennung unterschiedlicher Anomalien ist auch in wissenschaftlichen Studien ein übliches Vorgehen (Panagopoulos et al., 2021; Campbell et al., 2013). Bei der Entwicklung des SaQC wurde die Möglichkeit integriert, Datenpunkte, die bereits durch einen Test markiert wurden, von weiteren Tests auszuschließen, was in diesem Workflow jedoch nicht verwendet wurde (Schäfer et al., 2023).

Die automatische Markierung der Datenpunkte wird in dem vorgestellten Workflow nur in *BAD* und *GOOD* unterteilt. Hier wären weitere Abstufungen sinnvoll. Beispielsweise die Einführung eines Ampelsystems mit der zusätzlichen Kategorie *Suspicious* (Campbell et al., 2013). Auch dafür bietet das SaQC bereits eine integrierte Möglichkeit (Schäfer et al., 2023).

Eine zufällige Wahl der Validierungszeitreihen ist sinnvoll, um eine vollkommen

objektive Auswahl zu treffen. Jedoch kann es auch von Vorteil sein gezielt Zeitreihen zu wählen, welche Interessante Strukturen und Anomalien aufweisen. Bei den zufällig gewählten Zeitreihen können keine Sprünge erkannt werden, was folglich keine Validierung des Jump-Tests zulässt.

Taylor und Loescher (2013) schlagen eine Verbesserung der Parametrisierung vor, indem Informationen aus räumlich und zeitlich benachbarten Beobachtungen einbezogen werden sollen. Da Grundwasserstände von verschiedenen Stationen innerhalb eines Aquifers von Natur aus eine hohe Korrelation aufweisen, kann diese Art von Daten zusätzliche Informationen über die natürlichen Schwankungen der gemessenen Daten liefern und so dazu beitragen, genauere Schwellenwerte festzulegen. Es können Gewichtungsfaktoren eingeführt werden, die die räumlichen und zeitlichen Informationen berücksichtigen. Diese Faktoren legen fest, wie stark die benachbarten Daten in die Berechnungen einfließen.

Die Testparameter werden immer anhand der vorhandenen Daten ermittelt. Die statistische Parametrisierung birgt daher eine große Anpassungsfähigkeit an neue Begebenheiten. Da sich die Verteilung von Grundwasserständen aufgrund von Grundwasserabsenkungen im Laufe der Zeit erheblich verschieben kann, ist diese Flexibilität von großer Bedeutung. Eine Herausforderung bei diesem dynamischen Ansatz ist jedoch die Notwendigkeit, zuverlässige Flags zu liefern, die sich nicht nachträglich ändern, da diese Daten in vielen Fällen als Entscheidungsgrundlage dienen (Sartirana et al., 2022). Es muss daher ein Weg gefunden werden, der ein Gleichgewicht zwischen Flexibilität und Kontinuität der Qualitätskontrollen gewährleistet. Eine mögliche Lösung könnte darin bestehen, die Parameter auf der Grundlage der Daten der letzten 10 Jahre zu berechnen und nur die noch nicht gekennzeichneten Daten einer Qualitätskontrolle zu unterziehen (Taylor und Loescher, 2013). Da die klimatologischen Mittelwerte ebenfalls alle 10 Jahre neu berechnet werden, kann auf diese Weise sichergestellt werden, dass die durchgeführten Analysen sowohl einen wasserrechtlichen Bezugsrahmen als auch eine konsistente Qualitätskontrolle aufweisen (World Meteorological Organization, 2017).

## 6 Schlussfolgerungen

Der neu entwickelte Workflow zur automatischen QC von Grundwasserzeitreihen markiert bei 20 Sensor-Stationen im Mittel 3.2% der Datenpunkte als Anomalien. Die Anomalieerkennung unterscheidet sich zwischen Stationen unterschiedlicher Messintervalle (10min: 1.6%; 60min: 4,8%). Bei 60min erkennt der Constant (90%) die meisten Anomalien, wohingegen bei 10min Offset (48.9%) und LOF (54.2%) dominieren. Bei 10min ist der statistisch ermittelte window (constant) durch einen im Workflow gewählten Expertengrenzwert korrigiert, sodass constant bei 10min (3.3%) signifikant weniger Anomalien als bei den 60min (90%) ohne Grenzwertkorrektur erkennt. Durchschnittlich werden 12.4% der markierten Anomalien durch mehrere Tests erkannt, sodass alle 5 Tests (Range, Offset, Constant, LOF, Jump) für eine differenzierte Anomalie Erkennung notwendig sind. Der Vergleich mit 3 Expertenkontrollen an 3 Sensor-Stationen zeigt die automatische QC eine gute Übereinstimmung der als *GOOD* markierten Datenpunkte (Spezifität = 0.98 bis 0.99), aber geringe Werte bei Recall (0.01 bis 0.36) und Präzision (0.02 bis 0.67). Werden nur manuelle Anomalieerkennung genommen, in denen die Experten mehrheitlich übereinstimmen (10% aller manuell markierten Datenpunkte), steigt der Recall im Mittel auf 0.98. 73% der automatischen Anomalieerkennung sind durch keinen Experten bestätigt (False BAD), liegen aber in der Größenordnung von der manuellen Anomalieerkennung, die nur auf einen Experten zurückzuführen sind (89%). Manuelle und automatische QC erkennen nicht eindeutig und nicht einheitlich die Zeiträume, in denen ein Einfluss durch nahegelegene Grundwasserentnahme vorliegt. Insgesamt zeigen beide QC-Verfahren Unsicherheiten, wovon die Unsicherheit der automatische QC aufgrund eines standardisierten Workflows und eindeutiger Dokumentation weiter untersucht werden konnte. Im Mittel zeigen 44% der Stationen einen signifikanten linearen Trend im Unsicherheitsbereich der Testparameter, wozu über 80% der Stationen einen signifikanten linearen Trend bei max (Range (2)), window (Constant) und thresh (LOF) haben. Durch zu hohen max (Range (2)) werden niederschlagsreiche Winter als Anomalie markiert und durch zu niedrigen window (Constant) zu viele Stagnationen im Grundwasserstand, die auf stabilen Grundwasserständen bei Stationen mit 60min Messintervall zurückzuführen sind. Zeiträume von nahen Grundwasserentnahmen werden zufällig durch LOF und Offset markiert. Die statistische Parametrisierung in der automatischen QC ist somit sensitiv gegenüber den örtlichen Begebenheiten, markiert aber ohne Expertengrenzen auch plausible Messungen als Anomalie. Zusammenfassend demonstriert die Masterarbeit die Machbarkeit einer automatischen QC für Grundwasserzeitreihen mit dem Vorteil, Unsicherheiten durch die Standardisierung und Dokumentation im Workflow untersuchen zu können und damit die QC nachvollziehbar und kontinuierlich zu optimieren. Weiterhin ist die Interaktion durch den Experten notwendig, der Parametergrenzwerte setzt und komplexe Muster wie Entnahmemuster in der automatischen Anomalieerkennung identifiziert. Für die weitere Optimierung der automatischen QC sind Manuelle Expertenkontrolle unabdingbar, weshalb die manuelle QC im ersten Schritt standardisiert und besser dokumentiert werden muss. Auch im Sinne der nationalen Wasserstrategie, ist das Ziel ein qualitätsgesichertes Echtzeit-Monitoring für das Grundwasserressourcenmanagement bundesweit zu etablieren.

## Literatur

- Adamowski, K., & Hamory, T. (1983). A Stochastic Systems Model of Groundwater Level Fluctuations. *Journal of Hydrology*, *62*, 129–141.
- Antonetti, M., & Zappa, M. (2018). How can expert knowledge increase the realism of conceptual hydrological models? A case study based on the concept of dominant runoff process in the Swiss Pre-Alps. *Hydrol. Earth Syst. Sci.*, *22*, 4425–4447. <https://doi.org/10.5194/hess-22-4425-2018>
- Arbesser, C., Spechtenhauser, F., Mühlbacher, T., & Piringer, H. (2016). Visplause: Visual Data Quality Assessment of Many Time Series Using Plausibility Checks. *IEEE Transactions on Visualization and Computer Graphics*. <https://doi.org/10.1109/TVCG.2016.2598592>
- BadenovaNetze. (2023). *Wasserwerke und Anlagen* [Zugriff am [17.08.2023]]. <https://wasser.badenovanetze.de/ueber-uns/wasserwerke-und-anlagen/>
- Bakker, M., & Schaars, F. (2019). Solving Groundwater Flow Problems with Time Series Analysis: You May Not Even Need Another Model. *Groundwater*, *57*(6), 826–833. <https://doi.org/10.1111/gwat.12927>
- Bannick, C., Engelmann, B., Fendler, R., Frauenstein, J., Ginzky, H., Hornemann, C., Ilvonen, O., Kirschbaum, B., Penn-Bressel, G., Rechenberg, J., Richter, S., Roy, L., & Wolter, R. (2008). *Grundwasser in Deutschland* (L. Keppner & B. Kirschbaum, Hrsg.). Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU). [www.bmu.de](http://www.bmu.de)
- Bekesi, G., McGuire, M., & Moiler, D. (2008). Groundwater Allocation Using a Groundwater Level Response Management Method—Gnangara Groundwater System, Western Australia. *Springer*.
- Betting, D., Selz, M., Morhard, A., Kern, F.-J., & Schrempp, S. (2006). *Nitrathaushalt und Eintragungspotentiale der Trinkwassergewinnungsgebiete – Auskunftsplattform Gewässerschutz: Zusammenfassender Abschlußbericht* (Techn. Ber.) [Gefördert aus dem Innovationsfonds Klima- und Wasserschutz der badenova AG & Co. KG]. badenova AG & Co. KG. Freiburg im Breisgau.
- Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2020). A review on outlier/anomaly detection in time series data. *ACM*.
- Branisavljevic, N., Kapelan, Z., & Prodanovic, D. (2011). Improved real-time data anomaly detection using context classification. *Journal of Hydroinformatics*, *13*(39), 307–323. <https://doi.org/10.2166/hydro.2011.042>
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *MOD 2000*, 1-58113-218–2/00/05.
- Bundesministerium für Umwelt, Nukleare Sicherheit und Verbraucherschutz. (2023). *Nationale Wasserstrategie* [Kabinettsbeschluss vom 15. März 2023]. [https://www.bmu.de/fileadmin/Daten\\_BMU/Download\\_PDF/Binnengewasser/nationale\\_wasserstrategie\\_2023\\_bf.pdf](https://www.bmu.de/fileadmin/Daten_BMU/Download_PDF/Binnengewasser/nationale_wasserstrategie_2023_bf.pdf)
- Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*, 429–450. <https://doi.org/10.1007/s10472-017-9564-8>
- Campbell, J. L., Rustad, L. E., Porter, J. H., Taylor, J. R., Dereszynski, E. W., Shanley, J. B., Gries, C., Henshaw, D. L., Martin, M. E., Sheldon, W. M.,



- & Boose, E. R. (2013). Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. *BioScience*, *63*(7), 574–585. <https://doi.org/10.1525/bio.2013.63.7.10>
- Clemens-Meyer, F. H. L. R., Lepot, M., Blumensaat, F., Leutnant, D., & Gruber, G. (2021). Data validation and data quality assessment. In J.-L. Bertrand-Krajewski, F. Clemens-Meyer & M. Lepot (Hrsg.), *Metrology in Urban Drainage and Stormwater Management: Plug and Pray* (S. 327–390). IWA Publishing. [https://doi.org/10.2166/9781789060119\\_0327](https://doi.org/10.2166/9781789060119_0327)
- Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd International Conference on Machine Learning*.
- Deutscher Wetterdienst. (2023). Freier Zugang zu vielen Klimadaten des Climate Data Centers (CDC) [Zugriff am [17.08.2023]].
- Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., & Vose, R. S. (2010). Comprehensive Automated Quality Assurance of Daily Surface Observations. *Journal of Applied Meteorology and Climatology*, *49*, 1616–1633. <https://doi.org/10.1175/2010JAMC2375.1>
- Durre, I., Menne, M. J., & Vose, R. S. (2008). Strategies for Evaluating Quality Assurance Procedures. *J. Appl. Meteor. Climatol.*, *47*, 1785–1791. <https://doi.org/10.1175/2007JAMC1706.1>
- Erdbrügger, J., van Meerveld, I., Seibert, J., & Bishop, K. (2022). Shallow groundwater level time series and a groundwater chemistry survey from a boreal headwater catchment. *Earth System Science Data Discussions*. <https://doi.org/10.5194/essd-2022-114>
- Erhan, L., Ndubuaku, M., Di Mauro, M., Song, W., Chen, M., Fortino, G., Bagdasar, O., & Liotta, A. (2020). Smart anomaly detection in sensor systems: A multi-perspective review [Journal Pre-proof, Received: 15 June 2020, Revised: 12 September 2020, Accepted: 4 October 2020]. *Information Fusion*. <https://doi.org/https://doi.org/10.1016/j.inffus.2020.10.001>
- Faybishenko, B., Versteeg, R., Pastorello, G., Dwivedi, D., Varadharajan, C., & Agarwal, D. (2022). Challenging problems of quality assurance and quality control (QA/QC) of meteorological time series data. *Stochastic Environmental Research and Risk Assessment*, *36*, 1049–1062. <https://doi.org/https://doi.org/10.1007/s00477-021-02106-w>
- Fiebrich, C. A., Morgan, C. R., McCombs, A. G., Jr., P. K. H., & McPherson, R. A. (2010). Quality Assurance Procedures for Mesoscale Meteorological Data. *Journal of Atmospheric and Oceanic Technology*, *27*, 1565–1582. <https://doi.org/10.1175/2010JTECHA1433.1>
- Gantz, J., & Reinsel, D. (2012). The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. <https://api.semanticscholar.org/CorpusID:112313325>
- Gschwandtner, T., Aigner, W., Miksch, S., Gärtner, J., Kriglstein, S., Pohl, M., & Suchy, N. (2014). TimeCleanser: A Visual Analytics Approach for Data Cleansing of Time-Oriented Data. *i-KNOW '14*. <https://doi.org/http://dx.doi.org/10.1145/2637748.2638423>
- Gschwandtner, T., & Erhart, O. (2018). Know Your Enemy: Identifying Quality Problems of Time Series Data. *2018 IEEE Pacific Visualization Symposium (PacificVis)*. <https://doi.org/10.1109/PacificVis.2018.00034>

- Ha, K., Lee, E., An, H., Kim, S., Park, C., Kim, G.-B., & Ko, K.-S. (2021). Evaluation of Seasonal Groundwater Quality Changes Associated with Groundwater Pumping and Level Fluctuations in an Agricultural Area, Korea. *Water*, *13*(1), 51. <https://doi.org/https://doi.org/10.3390/w13010051>
- Haaf, E., Giese, M., Heudorfer, B., Stahl, K., & Barthel, R. (2020). Physiographic and Climatic Controls on Regional Groundwater Dynamics. *Water Resources Research*, *56*, e2019WR026545. <https://doi.org/10.1029/2019WR026545>
- Halder, S., Roy, M. B., & Roy, P. K. (2020). Analysis of groundwater level trend and groundwater drought using Standard Groundwater Level Index: a case study of an eastern river basin of West Bengal, India. *Springer Nature Switzerland AG*.
- Harrison, R. L. (2010). Introduction to Monte Carlo Simulation. *AIP Conference Proceedings*, *1204*, 17. <https://doi.org/10.1063/1.3295638>
- Hollenberg, A., Kochh, J., Libuda, J., Milke, H., Ristenpart, E., Ruß, H.-J., Sitzmann, D., Uhl, M., & Weiß, G. (2011). *DWA-M 181*. DWA Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall e. V.
- Horsburgh, J. S., Reeder, S. L., Jones, A. S., & Meline, J. (2015). Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software*, *70*, 32–44. <https://doi.org/10.1016/j.envsoft.2015.04.002>
- Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., & Osugi, T. T. (2005). Performance of Quality Assurance Procedures for an Applied Climate Information System. *Journal of Atmospheric and Oceanic Technology*, *22*, 105–112.
- International Organization for Standardization. (2023). Smart water management — Part 2: Data management guidelines [CD stage. Available from <https://www.iso.org>].
- Jackson, R. E. (1974). Time-Series Analysis of Groundwater Hydrographs from Surficial Deposits on the Canadian Shield. *Canadian Journal of Earth Sciences*, *11*, 177–188.
- Jeong, G., Yoo, D.-G., Kim, T.-W., Lee, J.-Y., Noh, J.-W., & Kang, D. (2021). Integrated Quality Control Process for Hydrological Database: A Case Study of Daecheong Dam Basin in South Korea (A. Ye, Hrsg.). *Water*, *13*, 2820. <https://doi.org/10.3390/w13202820>
- Jones, A. S., Horsburgh, J. S., & Eiriksson, D. P. (2018). Assessing subjectivity in environmental sensor data post processing via a controlled experiment. *Ecological Informatics*, *46*, 86–96. <https://doi.org/https://doi.org/10.1016/j.ecoinf.2018.05.001>
- Jousma, G., Attanayake, P., Chilton, J., Margane, A., Martínez Navarrete, C., Melo, M. T., López Guerrero, P. N., Polemio, M., Roelofsen, F., Sharma, S., Streetly, M., Subah, A., & Yaqoubi, A. A. (2006). *Guideline on: Groundwater monitoring for general reference purposes* (Report) [Revised March 2008, Report nr. GP 2008-1]. International Groundwater Resources Assessment Centre (IGRAC). Utrecht, The Netherlands.
- Junker, B., Wendt, O., Essler, H., & Lamprecht, K. (1977). *Hydrogeologische Karte Baden-Württemberg - Kaiserstuhl-Markgräflerland: Erläuterungen*

- zur Hydrogeologischen Karte von Baden-Württemberg 1:50000 [5 Karten, 12 Anlagen, 8 Schnitte]. Geologisches Landratsamt Baden-Württemberg.
- Kaffashzadeh, N., Kleinert, F., & Schultz, M. G. (2019). A New Tool for Automated Quality Control of Environmental Time Series (AutoQC4Env) in Open Web Services.
- Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, 42, W03408. <https://doi.org/10.1029/2005WR004376>
- Kunkel, K. E., Easterling, D. R., Hubbard, K., Redmond, K., Andsager, K., Kruk, M. C., & Spinar, M. L. (2005). Quality Control of Pre-1948 Cooperative Observer Network Data. *Journal of Atmospheric and Oceanic Technology*, 22, 1691–1705.
- Law, A. G. (1974). *Stochastic Analysis of Groundwater Level Time Series in the Western United States* (HYDROLOGY PAPERS Nr. 68). Colorado State University. Fort Collins, Colorado.
- Lin, H., Gharehbaghi, A., Zhang, Q., Band, S. S., Paie, H. T., Chau, K.-W., & Mosavig, A. (2022). Time series-based groundwater level forecasting using gated recurrent unit deep neural networks.
- Mirzavand, M., & Ghazavi, R. (2014). A Stochastic Modelling Technique for Groundwater Level Forecasting in an Arid Environment Using Time Series Methods. *Springer Science+Business Media Dordrecht*.
- Nevulis, R. H., Davis, D. R., & Sorooshian, S. (1989). Analysis of Natural Groundwater Level Variations for Hydrogeologic Conceptualization, Hanford Site, Washington. *Water Resources Research*, 25(7), 1519–1529.
- Nippes, K.-R., & Hettich, R. (1988). *Untersuchung zur Abschätzung der Infiltration der Möhlin im Bereich des Wasserwerks Hausen an der Möhlin im Sommer 1987 und Winter 1988* (Untersuchung). Freiburger Energie- und Wasserversorgungs-AG. Freiburg.
- Panagopoulos, Y., Konstantinidou, A., Lazogiannis, K., Papadopoulos, A., & Dimitriou, E. (2021). A New Automatic Monitoring Network of Surface Waters in Greece: Preliminary Data Quality Checks and Visualization (B. Karthikeyan & V. Kanakoudis, Hrsg.). *Hydrology*, 8, 33. <https://doi.org/10.3390/hydrology8010033>
- Patle, G. T., Singh, D. K., Sarangi, A., Rai, A., Khanna, M., & Sahoo, R. N. (2015). Time Series Analysis of Groundwater Levels and Projection of Future Trend.
- Peter, A. (1998). *Dreidimensionale, instationäre Grundwasserströmungsmodellierung für das Einzugsgebiet des Wasserwerkes Hausen an der Möhlin* [Unveröffentlichte Diplomarbeit, Institut für Hydrologie, Albert-Ludwigs-Universität].
- Polz, J., et al. (2023). Expert Flagging of Commercial Microwave Link Signal Anomalies: Effect on Rainfall Estimation and Ambiguity of Flagging. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 1–5. <https://doi.org/10.1109/ICASSPW59220.2023.10193654>
- Porter, J. H., Hanson, P. C., & Lin, C.-C. (2012). Staying afloat in the sensor data deluge. *Trends in Ecology and Evolution*, 27(2), 121–129. <https://doi.org/10.1016/j.tree.2011.11.009>

- Rau, G., Post, V., Shanafield, M., Krekeler, T., Banks, E., & Blum, P. (2019). Error in hydraulic head and gradient time-series measurements: A quantitative appraisal. *Hydrol. Earth Syst. Sci.*, *23*(9), 3603–3629. <https://doi.org/https://doi.org/10.5194/hess-23-3603-2019>
- Retike, I., Bikse, J., Kalvans, A., Delina, A., Avotniece, Z., Zaadnoordijk, W. J., Jemeljanova, M., Popovs, K., Babre, A., Zelenkevičs, A., & Baikovs, A. (2022). Rescue of groundwater level time series: How to visually identify and treat errors. *Journal of Hydrology*, *605*, 127294. <https://doi.org/https://doi.org/10.1016/j.jhydrol.2021.127294>
- Rinderer, M., van Meerveld, H. J., & McGlynn, B. L. (2019). From Points to Patterns: Using Groundwater Time Series Clustering to Investigate Subsurface Hydrological Connectivity and Runoff Source Area Dynamics. *Water Resources Research*, *55*, 5784–5806.
- RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC. Boston, MA. <http://www.rstudio.com/>
- SADC-GMI. (2019). *SADC Framework for Groundwater Data Collection and Data Management* (Technical Report). SADC Groundwater Management Institute (SADC-GMI). Bloemfontein, South Africa. <http://www.sadc-gmi.org>
- Sartirana, D., Rotiroti, M., Bonomi, T., Amicis, M. D., Nava, V., Fumagalli, L., & Zanotti, C. (2022). Data-driven decision management of urban underground infrastructure through groundwater-level time-series cluster analysis: the case of Milan (Italy). *Hydrogeology Journal*, *30*, 1157–1177. <https://doi.org/10.1007/s10040-022-02494-5>
- Schäfer, D., Palm, B., Lünenschloß, P., Schmidt, L., Bumberger, J., Ziegner, N., Gey, R., & Gransee, F. (2023). *System for automated Quality Control - SaQC* (Version 2.4.1) [Available at: <https://rdm-software.pages.ufz.de/saqc/index.html>]. <https://doi.org/10.5281/zenodo.8092184>
- Schmidlin, T. W., Wilks, D. S., McKay, M., & Cember, R. P. (1995). Automated quality control procedure for the "water equivalent of snow on the ground" measurement. *Journal of Applied Meteorology*, *34*, 143–151.
- Schmidt, L., Schäfer, D., Geller, J., Lünenschloss, P., Palm, B., Rinke, K., Rebmann, C., Rode, M., & Bumberger, J. (2023). System for automated Quality Control (SaQC) to enable traceable and reproducible data streams in environmental science. *Preprint submitted to Elsevier*, 1–30. <https://git.ufz.de/rdm-software/SaQC>
- Shulski, M. D., You, J., Krieger, J. R., Baule, W., Zhang, J., Zhang, X., & Horowitz, W. (2014). Quality Assessment of Meteorological Data for the Beaufort and Chukchi Sea Coastal Region using Automated Routines. *Arctic*, *67*(1), 104–112.
- Sturtevant, C., Metzger, S., Nehr, S., & Foken, T. (2021). Quality Assurance and Control. In T. Foken (Hrsg.), *Springer Handbook of Atmospheric Measurements* (S. 49–106). Springer Nature Switzerland AG. [https://doi.org/10.1007/978-3-030-52171-4\\_3](https://doi.org/10.1007/978-3-030-52171-4_3)
- Talagala, P. D., Hyndman, R. J., Leigh, C., Mengersen, K., & Smith-Miles, K. (2019). A Feature-Based Procedure for Detecting Technical Outliers in Water-Quality Data From In Situ Sensors. *Water Resources Research*, *55*, 8547–8568. <https://doi.org/10.1029/2019WR024906>

- Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., . . . Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2022.03.014>
- Taylor, J. R., & Loescher, H. L. (2013). Automated quality control methods for sensor data: a novel observatory approach [CC Attribution 3.0 License]. *Biogeosciences*, *10*, 4957–4971. <https://doi.org/10.5194/bg-10-4957-2013>
- Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*(12), n/a–n/a. <https://doi.org/10.1029/2007WR006720>
- Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, *3*. <https://doi.org/https://doi.org/10.1038/sdata.2016.18>
- Woolf, K., McGibbon, D., Misrole, M., Mkali, A., & Flügel, T. (2023). *Guidance Document on Groundwater Data Collection* (WRC Report No. TT 905/22). Umvoto Africa (Pty) Ltd. South Africa.
- World Meteorological Organization. (2017). *WMO Guidelines on the Calculation of Climate Normals* [WMO-No. 1203. Available from: [publications@wmo.int](mailto:publications@wmo.int)].
- World Meteorological Organization. (2021). *Guidelines on Surface Station Data Quality Control and Quality Assurance for Climate Applications* (2021 edition) [WMO-No. 1269. Available from <https://public.wmo.int/en/meteoterm>].
- WRRL. (2000). Richtlinie 2000/60/EG des Europäischen Parlaments und des Rates [zur Schaffung eines Ordnungsrahmens für Maßnahmen der Gemeinschaft im Bereich der Wasserpolitik]. *Europäisches Parlament und Rat der Europäischen Union*. <https://eur-lex.europa.eu/>
- Yang, Y., Guan, H., Batelaan, O., McVicar, T. R., Long, D., Piao, S., Liang, W., Liu, B., Jin, Z., & Simmons, C. T. (2016). Contrasting responses of water use efficiency to drought across global terrestrial ecosystems. *Scientific Reports*, *6*, 8. <https://doi.org/10.1038/srep23284>
- Zamani, M. G., Moridi, A., & Yazdi, J. (2022). Groundwater management in arid and semi-arid regions. *Arabian Journal of Geosciences*, *15*, 14. <https://doi.org/10.1007/s12517-022-09546-w>

## A Appendix

Tabelle 13: Die berechneten Parameter für die Tests Range (2), Offset, Constant, Jump

Station	Offset			Range (2)		Constant	Jump
	<i>thresh</i> [m]	<i>tolerance</i> [m]	<i>window</i> [m]	<i>max</i> [m]	<i>min</i> [m]	<i>window</i> [min]	<i>thresh</i> [m]
PH_001	0.170	0.085	100	11.854	8.156	71	0.194
PH_005	0.208	0.104	100	11.783	6.971	48	0.219
PH_006	0.089	0.045	100	11.981	9.352	41	0.087
PH_015	0.100	0.050	100	12.668	7.508	273	0.182
PH_016	0.041	0.021	100	11.498	7.980	51	0.122
PH_017	0.051	0.027	100	11.588	7.855	43	0.120
PH_018	0.043	0.021	100	11.700	8.501	47	0.158
PH_019	0.084	0.042	100	11.911	8.221	46	0.154
PH_023	0.138	0.069	600	12.382	5.039	777	0.045
PH_025	0.106	0.053	600	12.932	8.800	407	0.056
PH_027	0.008	0.004	600	12.013	10.092	251	0.032
PH_028	0.163	0.082	600	10.959	7.438	282	0.158
PH_034	0.012	0.006	600	13.043	7.618	1095	0.056
PH_036	0.948	0.474	600	12.740	6.588	3351	0.050
PH_056	0.315	0.158	600	15.554	5.188	1231	0.270
PH_059	0.125	0.062	600	10.984	4.762	556	0.073
PH_078	0.229	0.114	600	14.988	9.221	818	0.028
PH_083	0.149	0.074	600	11.776	7.358	1719	0.045
PH_105	0.156	0.078	600	13.308	9.231	1334	0.050
PH_119	0.103	0.051	600	10.407	5.894	598	0.057

Tabelle 14: Die Kennzahlen Recall, Präzision, Spezifität und F-Score abhängig von der Definition der wahren Klasse. Ein Datenpunkt wird hier als Anomalie gewertet, wenn 1. mindestens ein Experte ihn markiert oder 2. die Experten mehrheitlich für eine Anomalie stimmen.

	PH 016		PH 017		PH 028	
	ab 1	ab 2	ab 1	ab 2	ab 1	ab 2
Recall	0.01	1	0.01	1	0.36	0.94
Präzision	0.02	0.01	0.27	0.26	0.67	0.54
Spezifität	0.99	0.99	0.99	0.99	0.98	0.98
F-Score	0.02	0.03	0.02	0.41	0.47	0.69

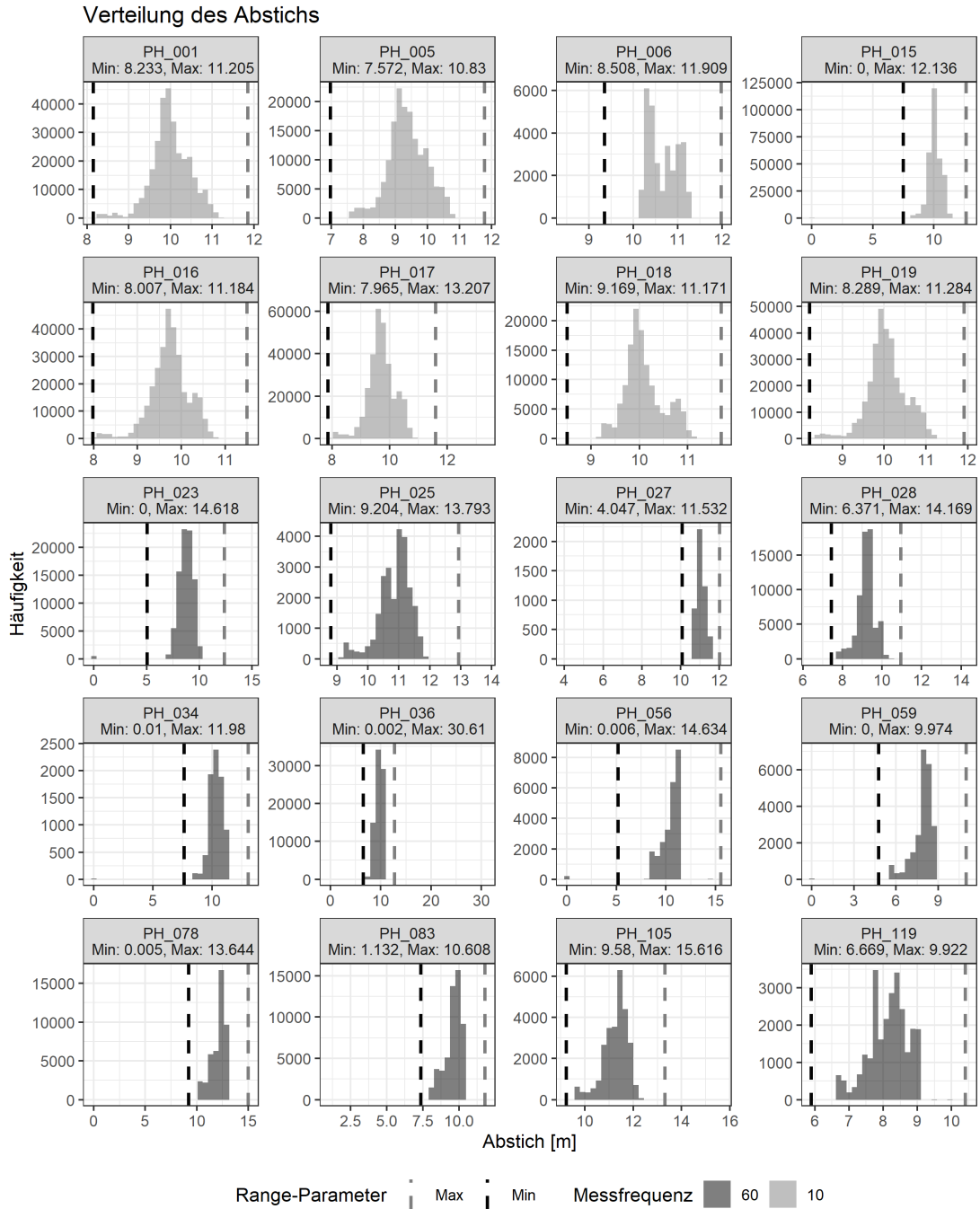


Abbildung 19: Verteilung der Abstichwerte

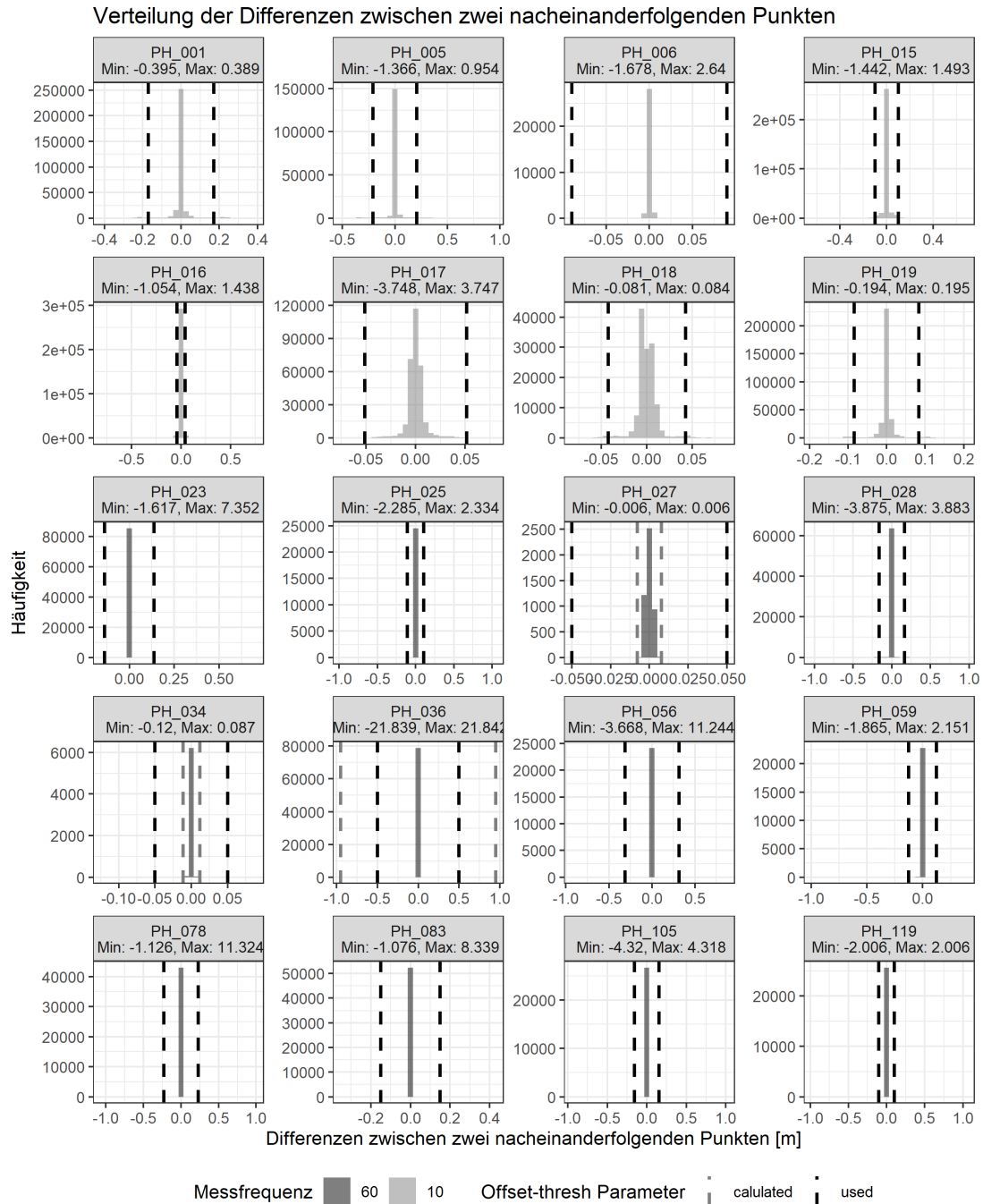


Abbildung 20: Verteilung der Differenzen zwischen zwei nacheinanderfolgenden Werten



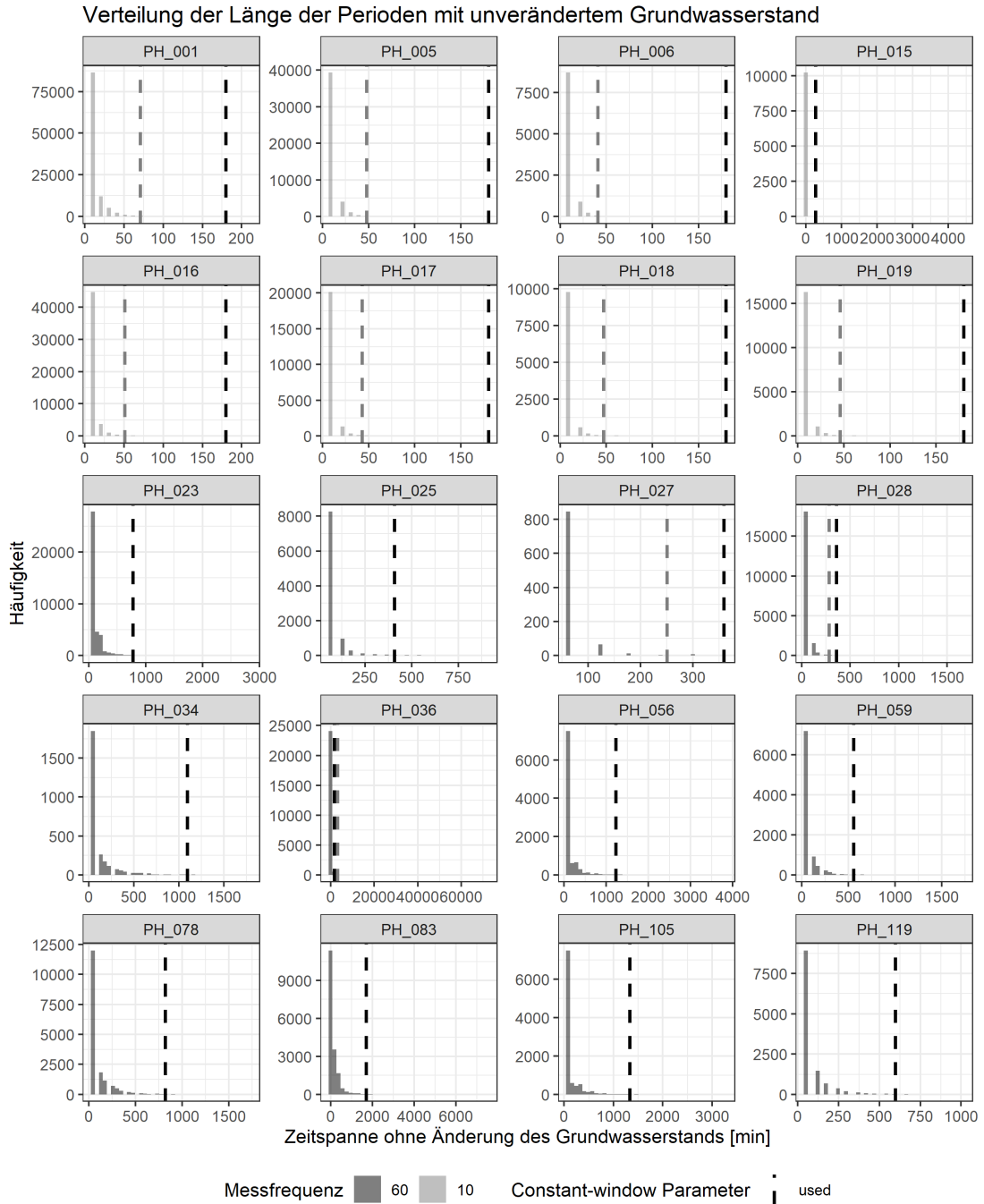


Abbildung 21: Verteilung der Perioden mit unverändertem Grundwasserstand

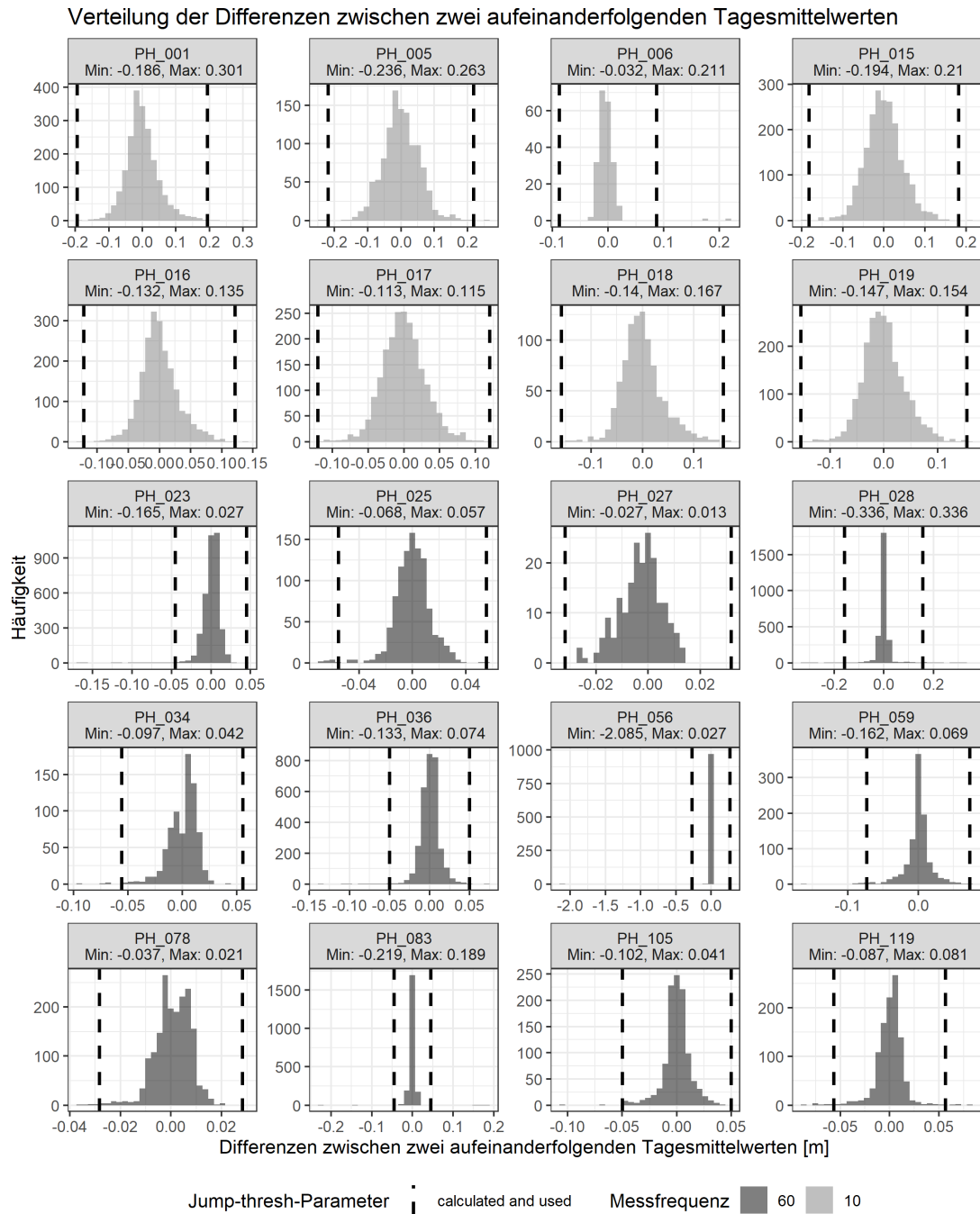


Abbildung 22: Verteilung der Differenzen zwischen zwei nacheinanderfolgenden Tagesmittelwerten

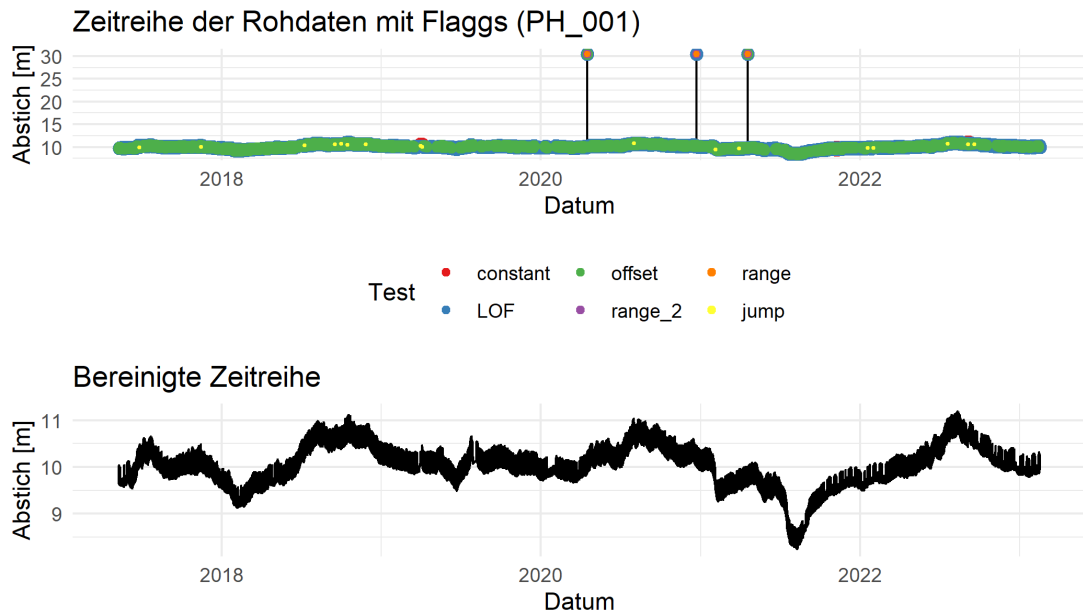


Abbildung 23: Geflaggte und flagbereinigte Zeitreihe der Station PH 001

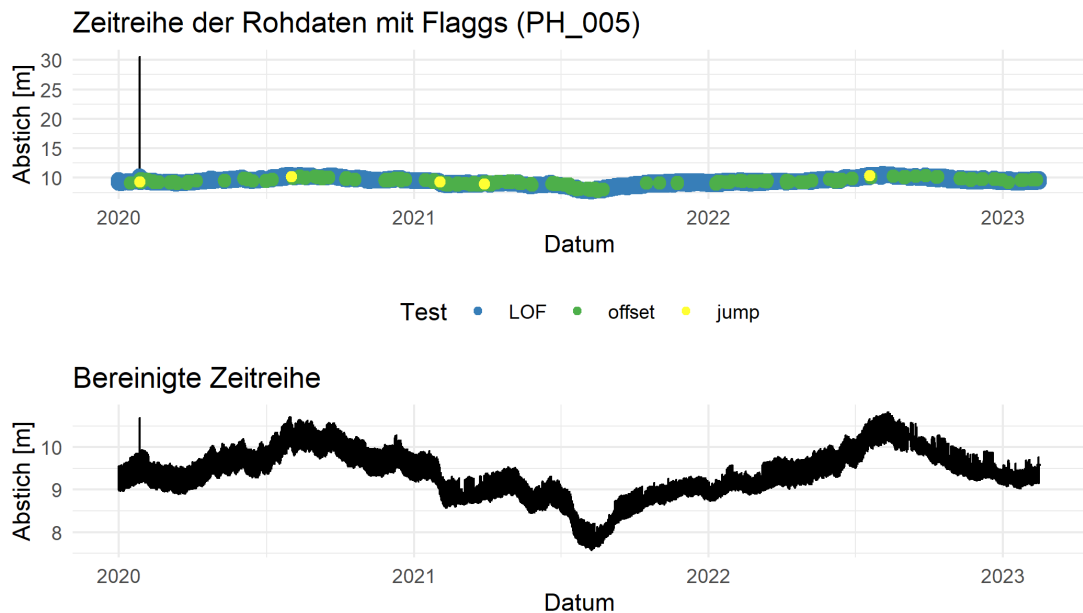


Abbildung 24: Geflaggte und flagbereinigte Zeitreihe der Station PH 005

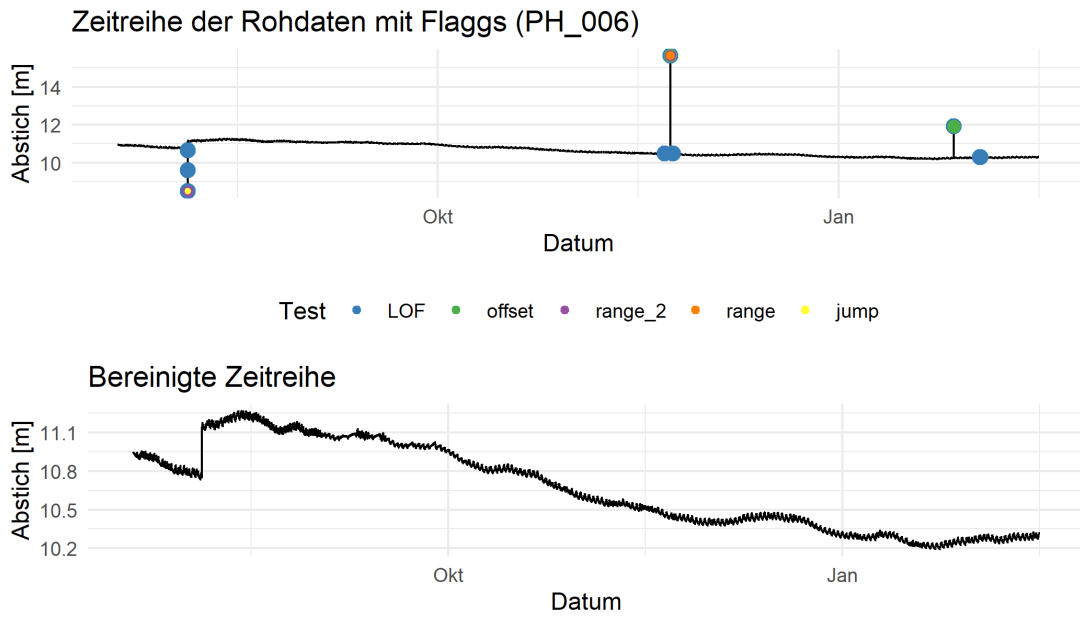


Abbildung 25: Geflaggte und flagbereinigte Zeitreihe der Station PH 006

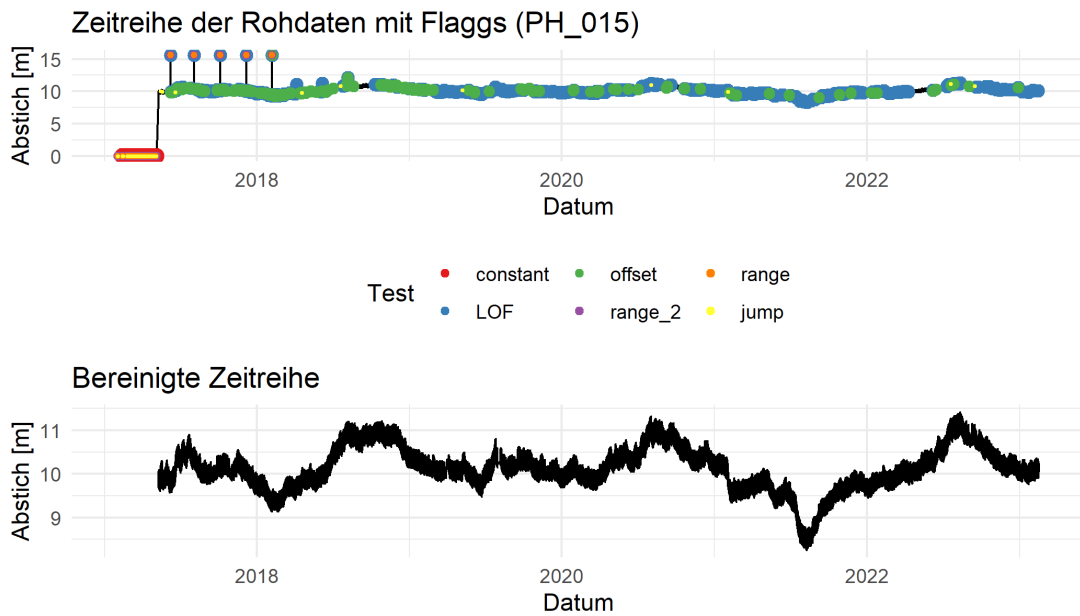


Abbildung 26: Geflaggte und flagbereinigte Zeitreihe der Station PH 015

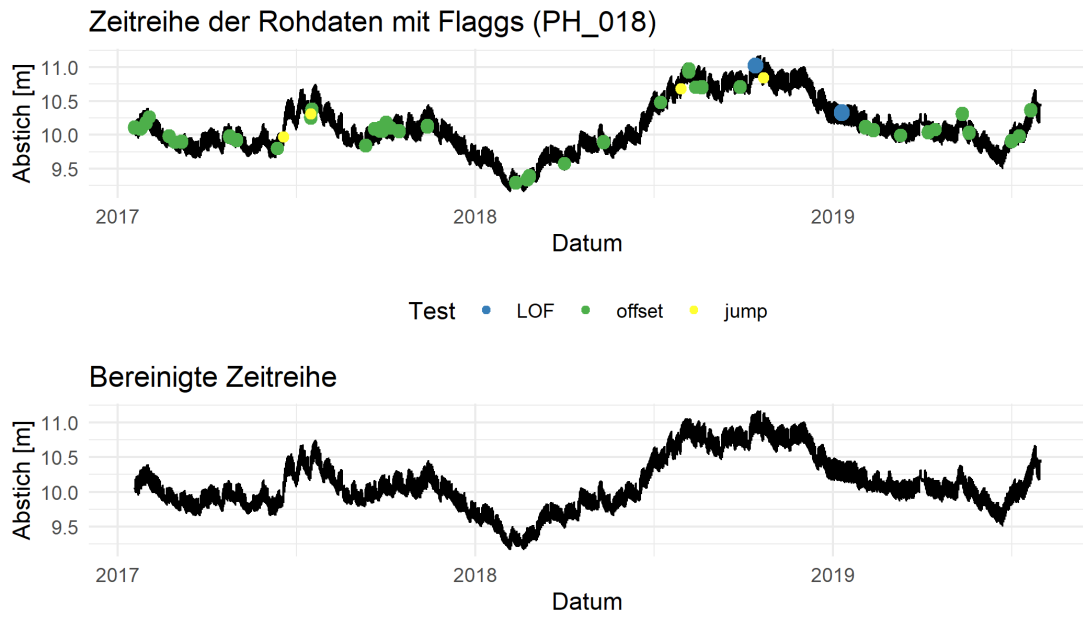


Abbildung 27: Geflaggte und flagbereinigte Zeitreihe der Station PH 018

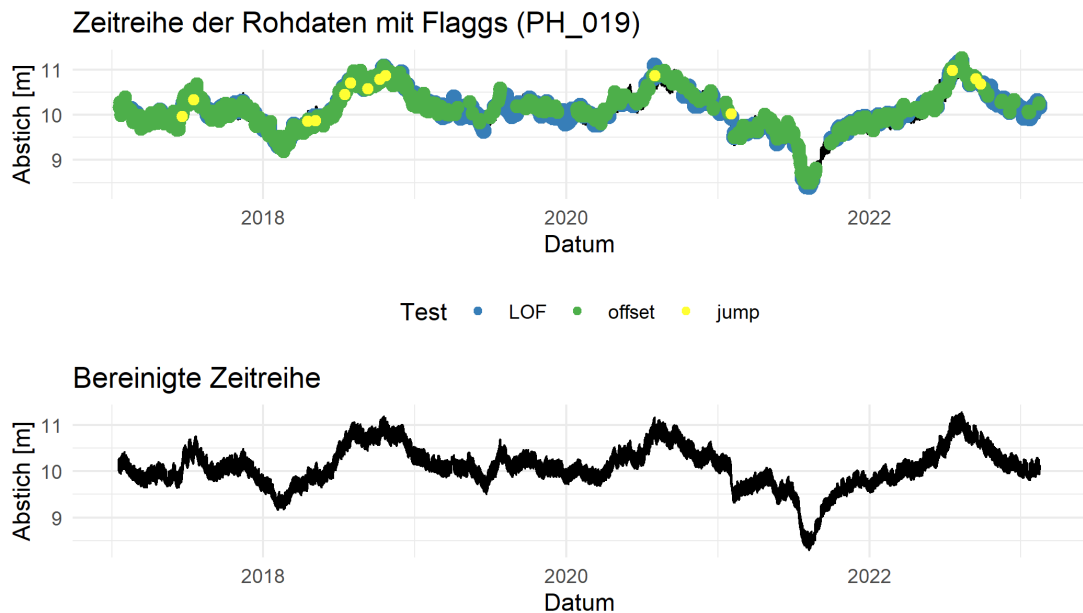


Abbildung 28: Geflaggte und flagbereinigte Zeitreihe der Station PH 019

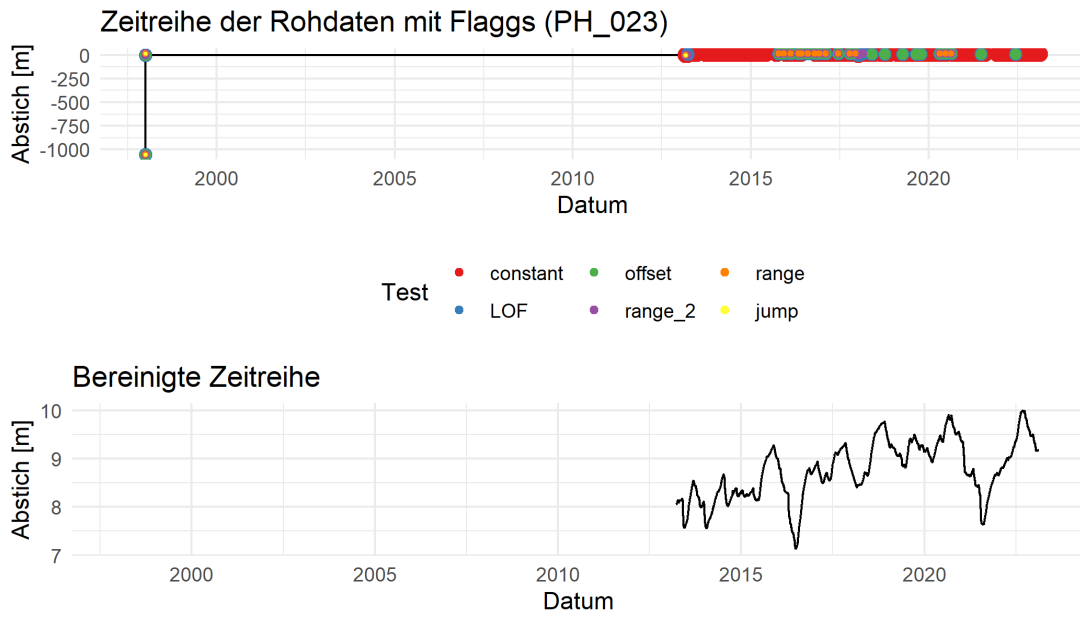


Abbildung 29: Geflaggte und flagbereinigte Zeitreihe der Station PH 023

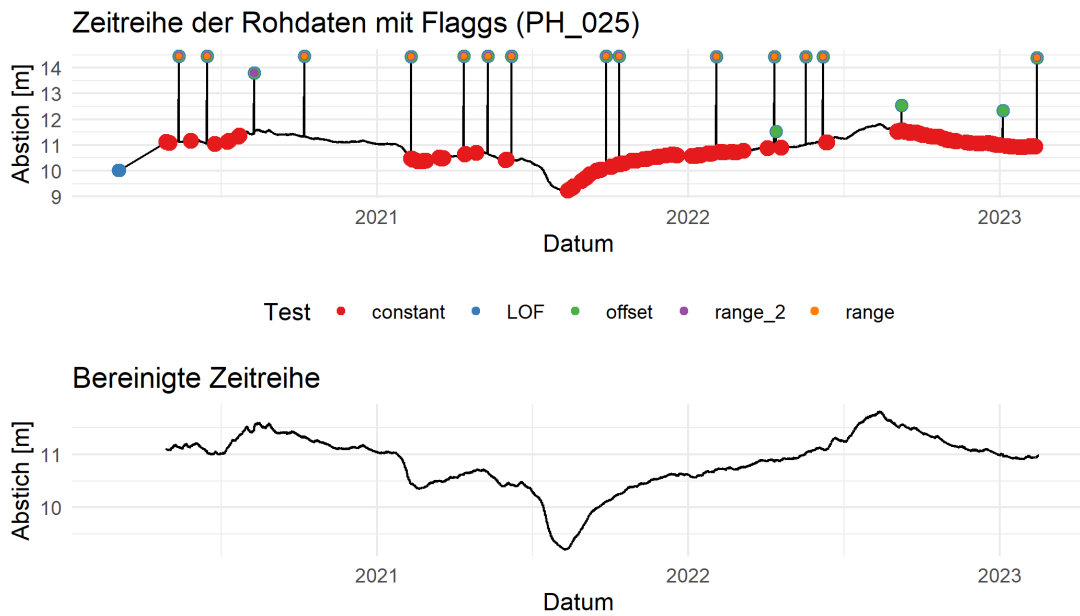


Abbildung 30: Geflaggte und flagbereinigte Zeitreihe der Station PH 025

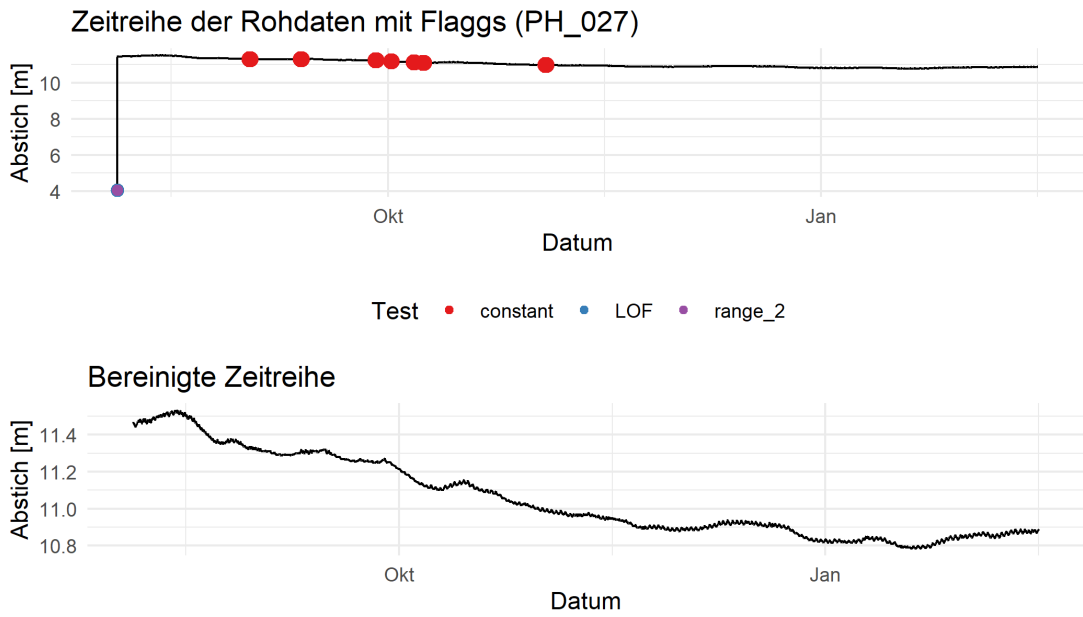


Abbildung 31: Geflaggte und flagbereinigte Zeitreihe der Station PH 027

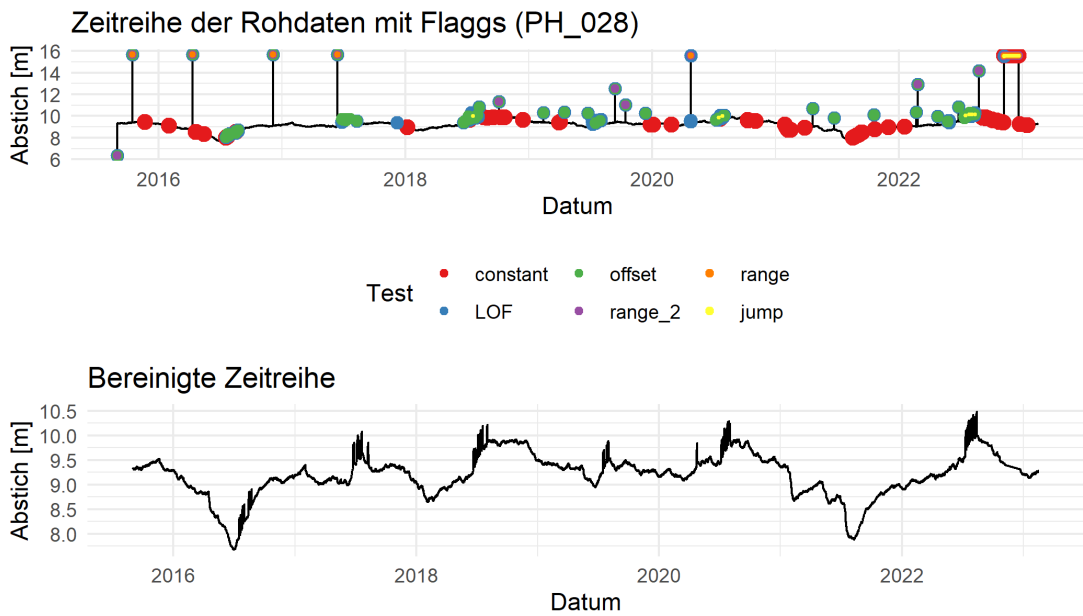


Abbildung 32: Geflaggte und flagbereinigte Zeitreihe der Station PH 028

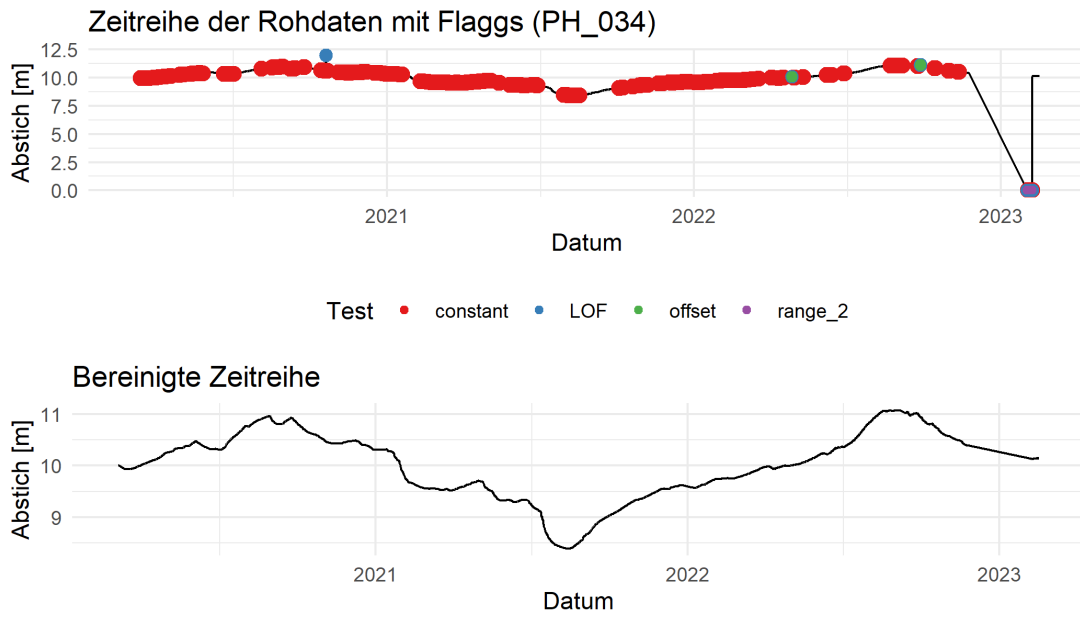


Abbildung 33: Geflaggte und flagbereinigte Zeitreihe der Station PH 034

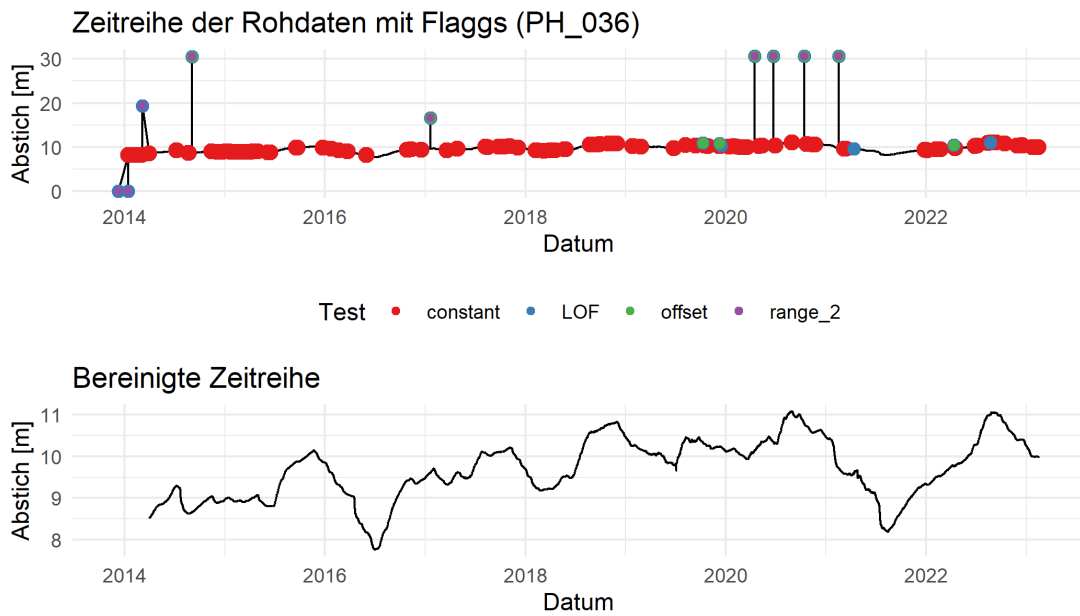


Abbildung 34: Geflaggte und flagbereinigte Zeitreihe der Station PH 036



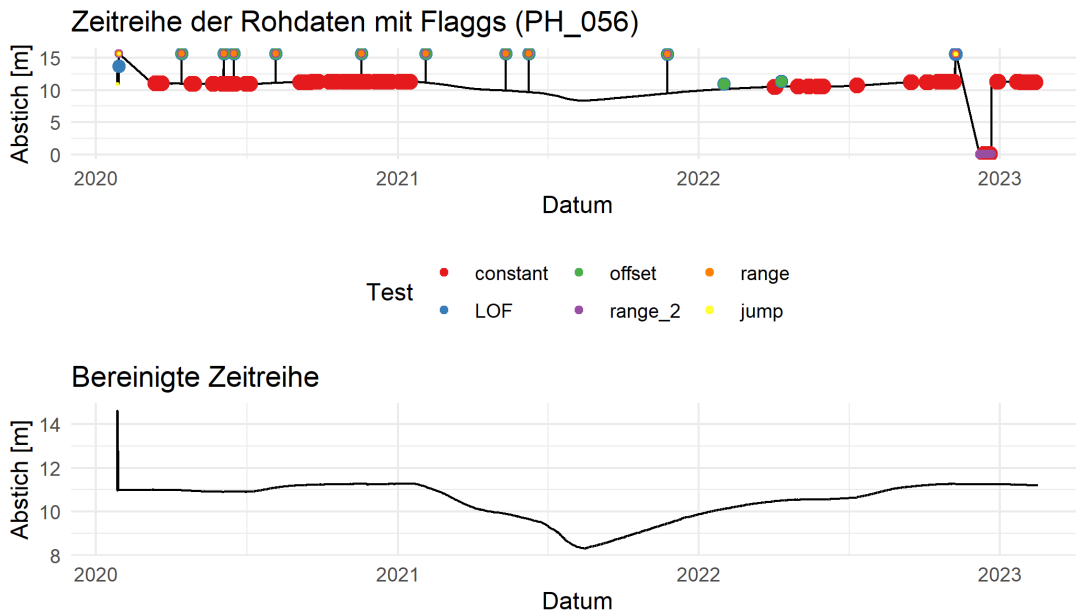


Abbildung 35: Geflaggte und flagbereinigte Zeitreihe der Station PH 056

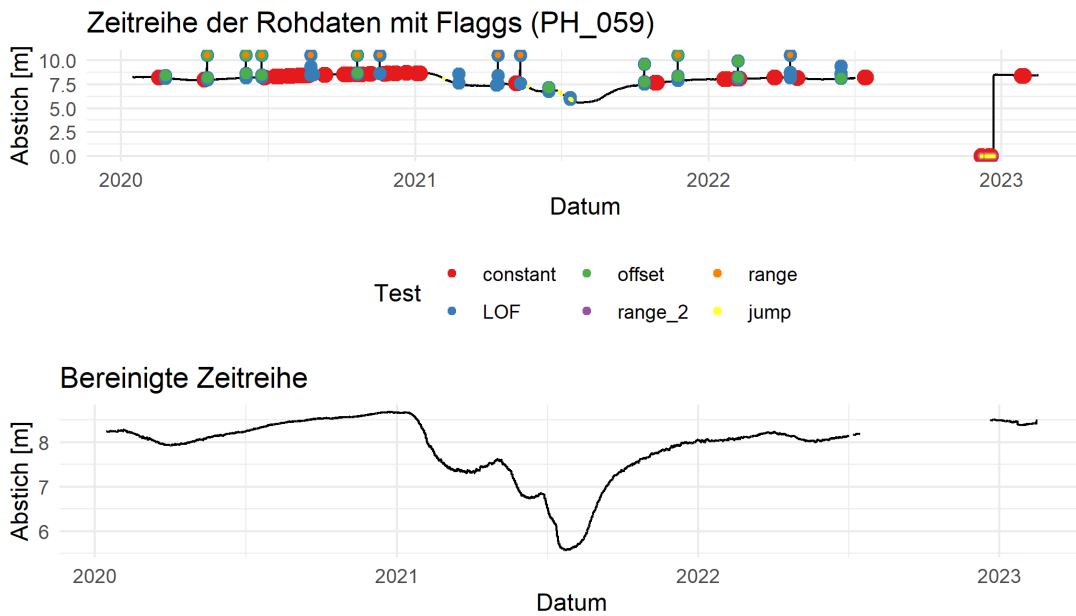


Abbildung 36: Geflaggte und flagbereinigte Zeitreihe der Station PH 059

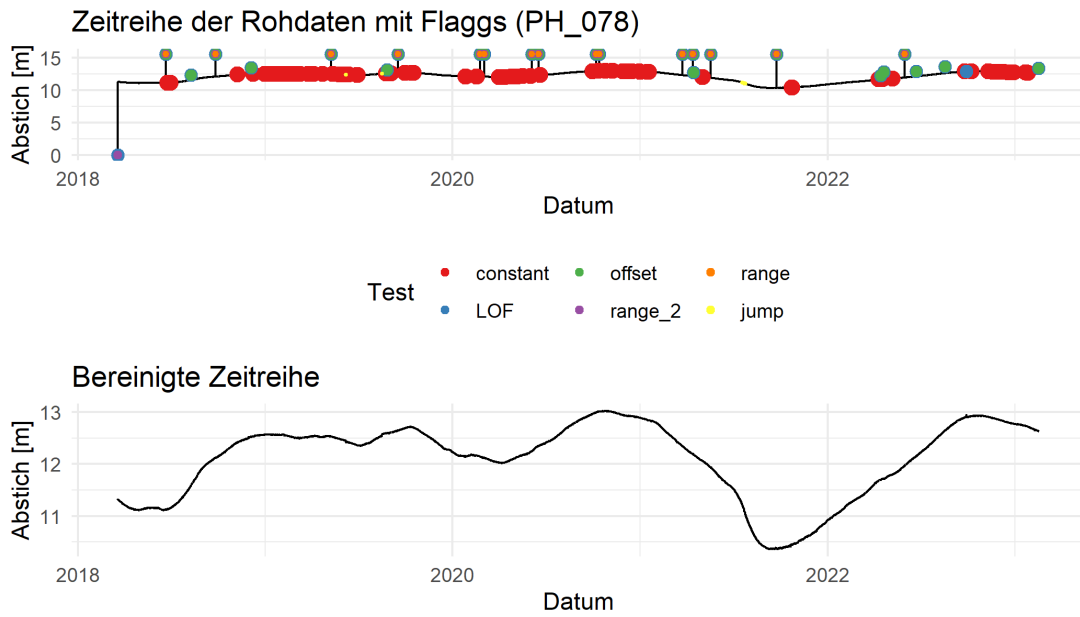


Abbildung 37: Geflaggte und flagbereinigte Zeitreihe der Station PH 078

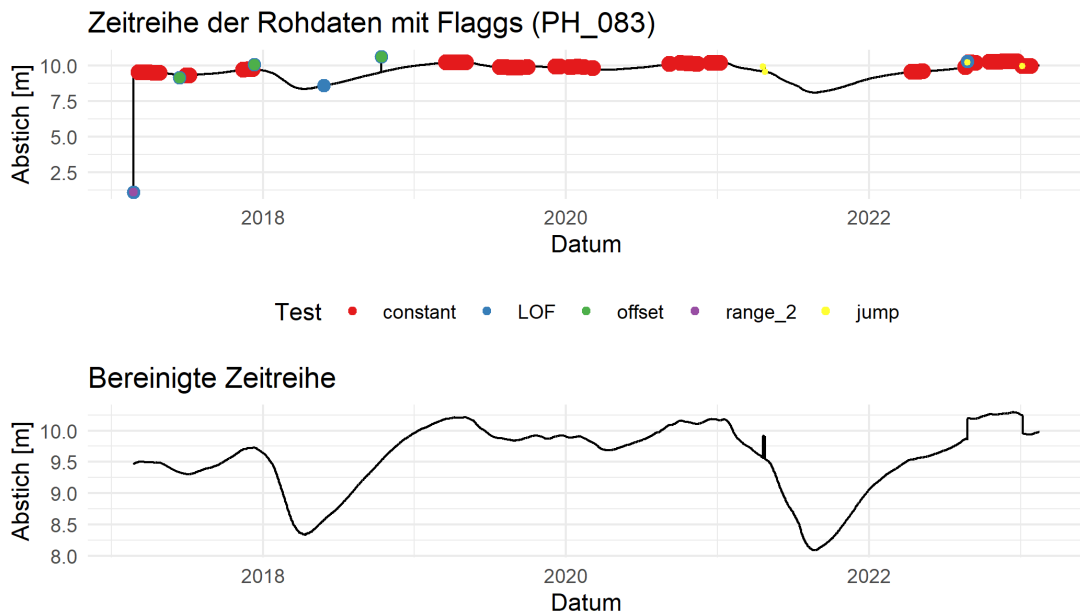


Abbildung 38: Geflaggte und flagbereinigte Zeitreihe der Station PH 083

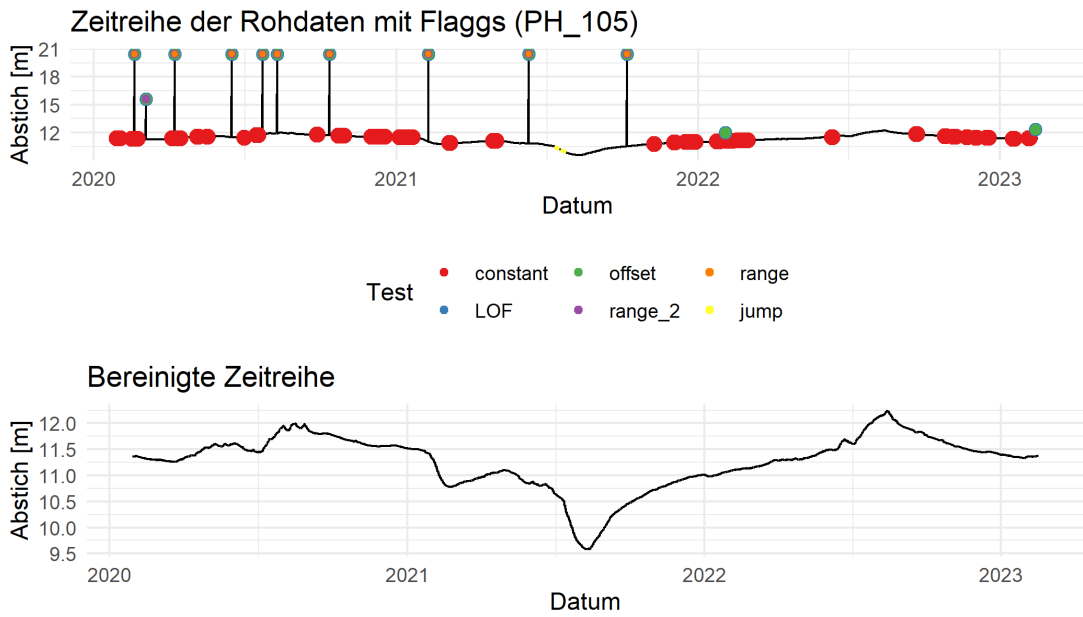


Abbildung 39: Geflaggte und flagbereinigte Zeitreihe der Station PH 105

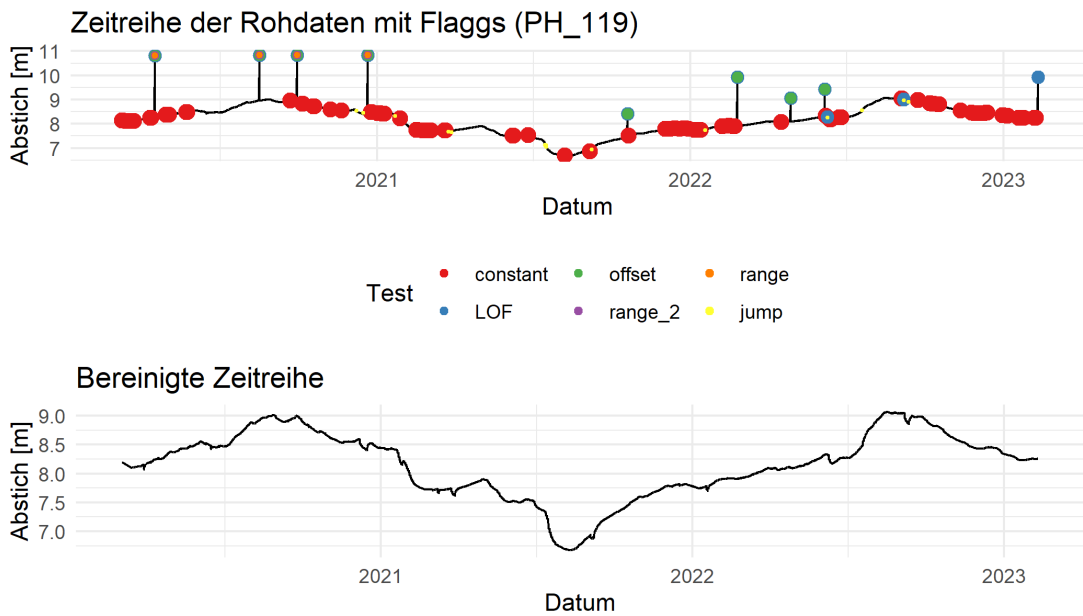


Abbildung 40: Geflaggte und flagbereinigte Zeitreihe der Station PH 119

# Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass die Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt wurde.

Ort, Datum

Unterschrift