Deltares

Stappenplan voor optimalisatie van hoogfrequente waterkwaliteitsdata

Waterschap Aa en Maas

microLAN On-line Biomonitoring Systems

AQUON Wateronderzoek en advies

Deze activiteit is mede gefinancierd door TKI Watertechnologie uit de PPS-innovatie programmasubsidie van het Ministerie van Economische Zaken

Opdrachtgever	Tki Watertechnologie

Documentgegevens	
Versie	0.1
Datum	11-09-2024
Projectnummer	11208383-000
Document ID	11208383-000-BGS-0003
Pagina's	83
Classificatie	
Status	definitief

Auteur(s)		
	Joep Appels (microLAN) Niels van Aarle (AQUON) Sander van Eijke (AQUON) Frank van Herpen (Waterschap Aa en Maas) Eppe Nieuwenhuis (AQUON) Joachim Rozemeijer (Deltares) Kevin Ouwerkerk (Deltares)	

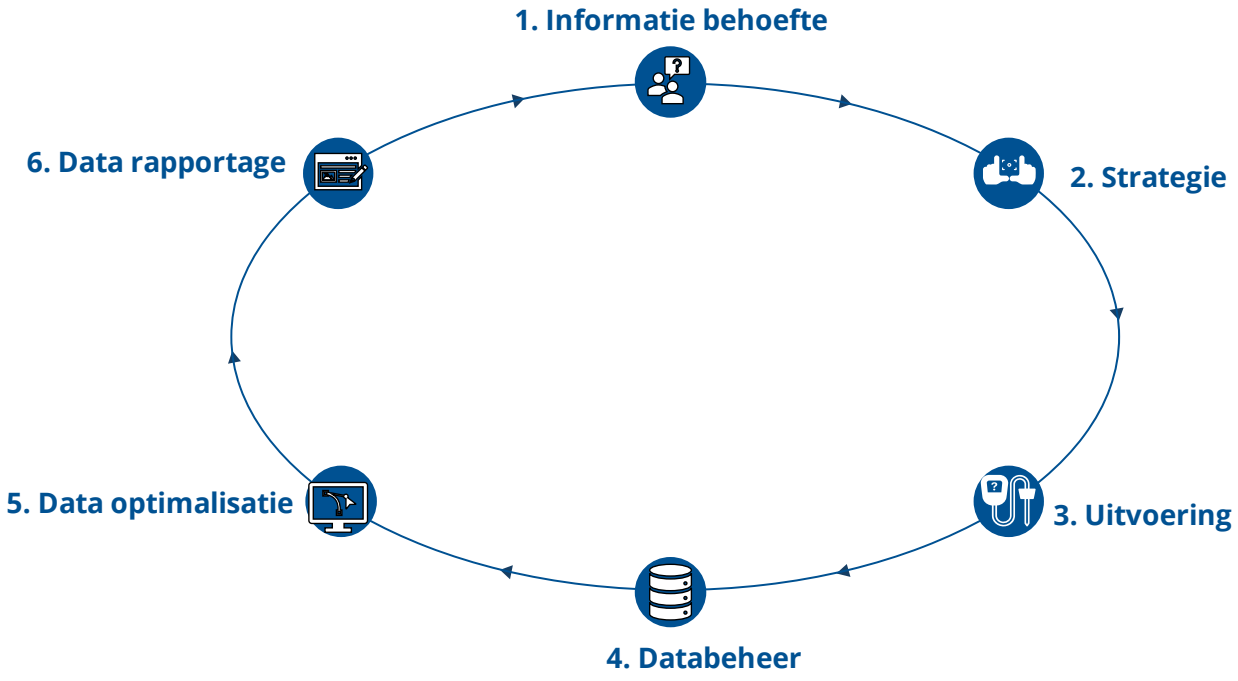
Inhoudsopgave

Inhoudsopgave		**3**
0	**Inleiding**	**4**
1	**Informatie behoefte**	**7**
2	**Strategie**	**10**
3	**Uitvoering**	**13**
4	**Data beheer**	**16**
5	**Data optimalisatie**	**20**
6	**Data rapportage**	**25**
A	**Literatuur studie**	**26**
B	**Voorbeeld data optimalisatie**	**67**

0 Inleiding

Binnen het in 2022 opgestarte Topconsortium voor Kennis en Innovatie (TKI) project Optima-HWQ werken Deltares, AQUON, Waterschap Aa en Maas en microLAN samen aan de optimalisatie van routines voor hoogfrequente waterkwaliteitsdata (bijv. sensoren en auto-analyzers). Optima-HWQ staat voor Optimal High-frequency Water Quality. We richten ons enerzijds op het near real-time detecteren van afwijkingen voor betere online visualisaties en adequaat sensoronderhoud. Anderzijds op het achteraf optimaliseren van de meetreeksen. Bijvoorbeeld door het opvullen van gaten in de meetreeks door Machine Learning algoritmes die worden gebaseerd op continue gegevens van andere parameters of door sensormetingen achteraf te corrigeren voor afwijkingen met conventionele waterkwaliteitsmetingen door gecertificeerde laboratoria. In dit document zijn de ervaringen die tijdens dit project zijn opgedaan beschreven in de vorm van een stappenplan dat kan worden doorlopen om te bepalen of hoogfrequente waterkwaliteitsmetingen noodzakelijk zijn en zo ja hoe de implementatie van de meetopstelling(en) zo goed mogelijk vormgegeven kan worden. Verdere informatie over dit project is te vinden op de public wiki pagina zoals voorbeelden voor het optimaliseren sensor data en een literatuurstudie.

Dit document beschrijft een cyclisch proces voor het meten en beheren van hoogfrequente sensoren en auto-analyzers. Het uitgangspunt hiervoor is de monitoringcyclus die veel gehanteerd wordt bij het ontwerp van waterkwaliteitsmeetnetten (zie Figuur 1 en bijv. Rozemeijer et al., 2004). In dit rapport specificeren we de monitoringcyclus voor de specifieke aspecten van de inzet van hoogfrequente waterkwaliteitsmonitoring. Het doel is nieuwe gebruikers een compleet overzicht te geven van aspecten waarmee bij de implementatie van waterkwaliteitsensoren en auto-analyzers rekening gehouden moet worden.



Figuur 1: Het cyclisch proces voor het meten en beheren van hoogfrequente sensoren en auto-analyzers voor waterkwaliteitsmetingen

Elk onderdeel van de cyclus hangt nauw met andere onderdelen samen en vormt een continue cyclus die begint bij het identificeren van de informatiebehoefte en eindigt bij de data-rapportage. Vanuit de data-rapportage volgt een evaluatie over in hoeverre de informatiebehoefte ingevuld is en of er nieuwe informatiebehoeften zijn ontstaan. Deze meetnetevaluatie kan leiden tot aanpassingen in de meetstrategie en zo wordt de cyclus opnieuw doorlopen in een geoptimaliseerd meetnet. Door de onderlinge afhankelijkheid en terugkoppeling tussen de stappen in de monitoringcyclus kan het gehele proces lastig te overzien zijn, zeker als het voor het eerst wordt uitgevoerd. Door dit samen te voegen in dit document hopen wij dit proces te vergemakkelijken en overzichtelijker te maken.

In Optima-HWQ en in dit document ligt de nadruk vooral op de stappen 4 en 5 uit de cyclus, de andere stappen zijn daarom in minder detail uitgewerkt maar voor de volledigheid wel beschreven. Hieronder volgt een kort overzicht van elke stap in dit proces. Het volledig uitvoeren van de monitoringcyclus is nodig voor het valideren en voor en optimale inzet van hoogfrequente waterkwaliteitsmetingen. Worden delen van dit proces niet of beperkt uitgevoerd dan kan men eigenlijk niet voldoende waarde hechten aan de resultaten en de betrouwbaarheid van de meetopstelling(en).

1. Informatiebehoefte
Het begint met het vaststellen van de informatiebehoefte. Dit houdt in dat hier de onderzoeksvra(a)g(en) worden geformuleerd en wordt bepaald welke gegevens nodig zijn om deze vragen te beantwoorden. Deze fase is cruciaal omdat het de basis vormt voor alle verdere stappen in het proces. Hier worden globaal de frequentie van metingen (hoogfrequente waterkwaliteitsmetingen nodig of niet) en de benodigde specificaties in beeld gebracht op basis van de onderzoeksvraag en de context van de gewenste waterkwaliteitsmonitoring. Belangrijk om hier te realiseren is dat het gebruik van sensoren een middel is en geen doel op zich. Het al dan niet inzetten van sensoren is dus afhankelijk van de informatiebehoefte, en hoeft niet altijd de beste oplossing te zijn.

2. Strategie
Op basis van de gedefinieerde informatiebehoefte en bestaande informatie over het gebied wordt een meetstrategie ontwikkeld. Dit omvat het bepalen van specifieke parameters, de nauwkeurigheid van de metingen en de precieze frequentie van de gegevensverzameling. Daarnaast worden meetlocaties en omgevingsfactoren vastgesteld en wordt een data-strategie geformuleerd om later de gegevens op een efficiënte en veilige manier te beheren en te analyseren.

3. Uitvoering
De uitvoering van de meetstrategie omvat de praktische stappen van installatie, stroomvoorziening, onderhoud, kalibratie en kostenbeheer. Hierbij is het essentieel om regelmatig de meetopstelling en de verzamelde data te evalueren en indien nodig bij te stellen, zodat op technisch niveau de gegevens continu voldoen aan de informatiebehoefte. Het kan namelijk ook zo zijn dat naast hoog frequente waterkwaliteitsmetingen ook andere gegevens nodig zijn zoals meteorologische data, afvoergegevens of grondwaterstanden. Echter gaan we niet in op het verwerken of verkrijgen van andere gegevens dan waterkwaliteitsgegevens anders dan het benoemen dat deze afhankelijk van de informatie behoefte nodig kunnen zijn.

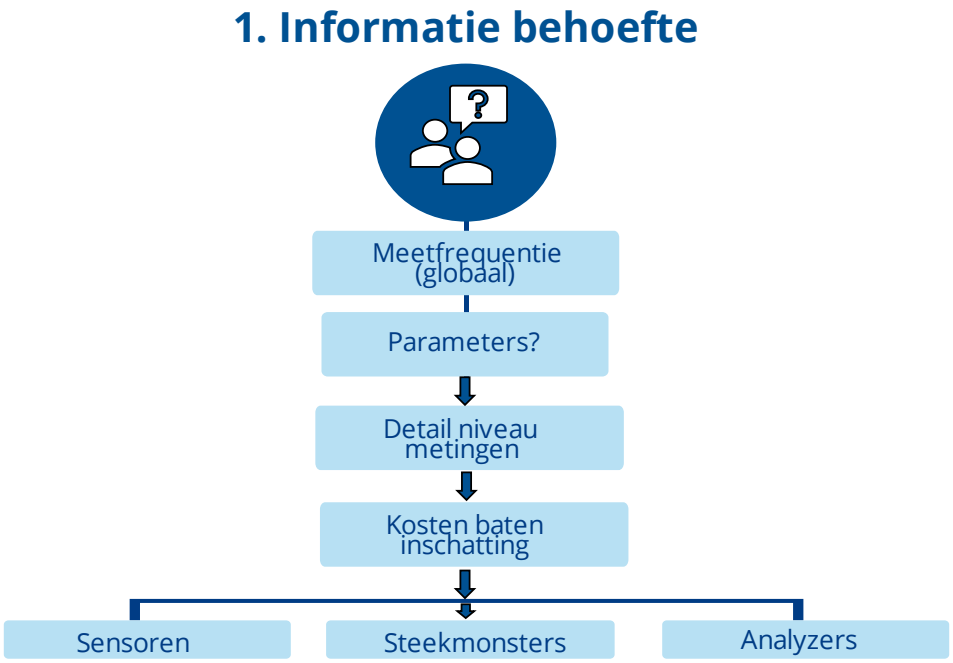
4. Data beheer
Het beheer van de verzamelde data begint met de opslag van ruwe gegevens in loggers en gaat door tot de integratie in ICT-systemen. Hier wordt aandacht besteed aan de veilige overdracht, opslag en toegang tot de data. Een gestructureerd datamodel en de naleving van standaarden, zoals de AQUO-standaard, zijn hierbij van groot belang.

5. Data optimalisatie
De optimalisatie van data omvat het valideren en corrigeren van de verzamelde gegevens. Ruwe sensordata zijn zonder optimalisatie vaak niet of minder goed bruikbaar. De optimalisatie gebeurt door het toepassen van methoden zoals het detecteren van uitschieters, driftcorrectie en het invullen van datagaten. Deze stappen zorgen ervoor dat de data betrouwbaar en bruikbaar is voor verdere analyses en rapportages.

6. Data rapportage
Ten slotte worden de geoptimaliseerde data gerapporteerd in een vorm die aansluit bij de oorspronkelijke informatiebehoefte. Dit kan afhankelijk van de doelgroep en de informatiebehoefte variëren van gedetailleerde technische rapporten tot dashboards waar de data live te zien is.

Leeswijzer
Na deze inleiding doorlopen we in dit rapport de hiervoor genoemde stappen uit de monitoringcyclus. Deze stappen worden beschreven en vooral voor de data optimalisatie (stap 5) zijn voorbeelden gegeven. Deze voorbeelden zijn voor een deel gebaseerd op methoden die verzameld zijn in een literatuuronderzoek (Bijlage A). We gaan in het hoofdrapport niet in op de scripts voor de data optimalisatie, maar die zijn wel opgenomen in Bijlage B. Deze rapportage is bedoelt voor waterkwaliteitsonderzoekers/beheerders die willen beginnen met hoogfrequente waterkwaliteitsmetingen.

1 Informatie behoefte



Informatiebehoefte
Het is als eerste van belang om na te gaan wat de informatiebehoefte is. Wat zijn de onderzoeksvragen en hoe kunnen deze worden beantwoord? Over welke parameters is informatie nodig en op welke ruimtelijke en temporele resolutie? Het is van belang om hier mee te beginnen omdat het de keuzes bepaald voor de rest van de stappen die betrekking hebben op het meten van de waterkwaliteit op een hoge frequentie en of het meten op een hoge frequentie voor de betreffende onderzoeksvraag überhaupt nodig is. Het is, anders dan in de monitoringscyclus, daarom in deze fase ook raadzaam om alvast globaal na te denken over de meetstrategie. Dit is meteen een goede realiteits-check: is de verwachting dat de onderzoeksvragen wel met voldoende betrouwbaarheid te beantwoorden zijn? Bij meerdere onderzoeksvragen is het raadzaam ze de prioriteren aangezien verschillende vragen en informatiebehoeften kunnen leiden tot verschillende meetstrategieën en snel oplopende kosten.

Meetfrequentie
Ten eerste zal aan de hand van de onderzoeksvraag en de informatiebehoefte een inschatting moeten worden gemaakt van de benodigde meetfrequentie. Het kan zijn dat voor het bepalen van de trend of toestand op één of meerdere locaties een frequentie van twee of vier weken voldoende is. Is het nodig om de relatie met neerslagbuien en waterkwaliteit vast te stellen of om betrouwbare gemiddelde concentraties of vrachten te berekenen, dan is waarschijnlijk een hogere meetfrequentie nodig. Zie bijvoorbeeld Figuur 2 waaruit de onzekerheid van gemiddelde concentraties op basis van laagfrequente metingen voor P-totaal blijkt. De lagere meetfrequentie mist duidelijk pieken die met de hogere frequentie wel worden gedetecteerd.

Verder kunnen de specifieke parameters ook bepalend zijn, niet alle parameters kunnen namelijk hoogfrequent gemeten worden. Tabel 1 kan helpen bij het maken van de keuze tussen meten met een hoge temporele resolutie, meten met een hoge ruimtelijke resolutie of een combinatie van beiden.

Tabel 1: Afwegingen voor bepalen van de ruimtelijke en temporele resolutie van het meetnet (waarbij ook combinaties van hoge ruimtelijke en temporele resoluties mogelijk zijn)

Kenmerk van de vraag	Hoogfrequent, weinig locaties (= sensoren gebruiken)	Laagfrequent, veel locaties (= geen sensoren gebruiken)
Onderzoeksgebied	Klein (een stroomgebied, een of enkele locaties)	Meerdere stroomgebieden, veel locaties inschatting over een heterogeen gebied
Variaties	Variatie in de tijd wordt duidelijk	Variatie ruimtelijk wordt zichtbaar (bronnen, hot spots)
Trends	Specifiek op één locatie, ook trends in processen detecteerbaar	Aselecte metingen, trend over steekproef van gebied of locatie.
Budget / beschikbaar fte	Budget en FTE's voor installatie, onderhoud, kalibratie en dataverwerking, bij meerdere locaties loopt dit op; conventionele monstername blijft nodig	Alleen conventionele monstername
Nauwkeurigheid van de meetreeks	Op één locatie hoog, over een groot gebied laag door enkele meetlocatie	Op één locatie laag, over een groot gebied hoog
Parameter	Beperkter aantal parameters mogelijk	Breed analysepakket mogelijk met steekmonsters

Detailniveau metingen

Een extra aandachtpunt is nog het belang van het absolute niveau van de meetwaarden. Sensoren zijn voornamelijk heel goed in het vangen van de dynamiek en tijdelijke veranderingen. Als het belangrijk is om de absolute waarden heel nauwkeurig te bepalen (zoals bij het bepalen van vrachten) dan zijn aanvullende reguliere steekmonster of debietproportionele monsters ook nodig voor het corrigeren van de meetreeks. Daarnaast kan het ook belangrijk zijn om gaten in de meetreeks op te vullen (bepalen vrachten), bijvoorbeeld aan de hand van de relatie met andere hoogfrequent gemeten parameters. Hiervoor zijn dan wel aanvullende parameters nodig die goed correleren met de parameters van interesse. Hierbij kan gedacht worden aan neerslag, grondwaterstanden of de afvoer maar ook de pH of geleidbaarheid of andere waterkwaliteitsparameters.

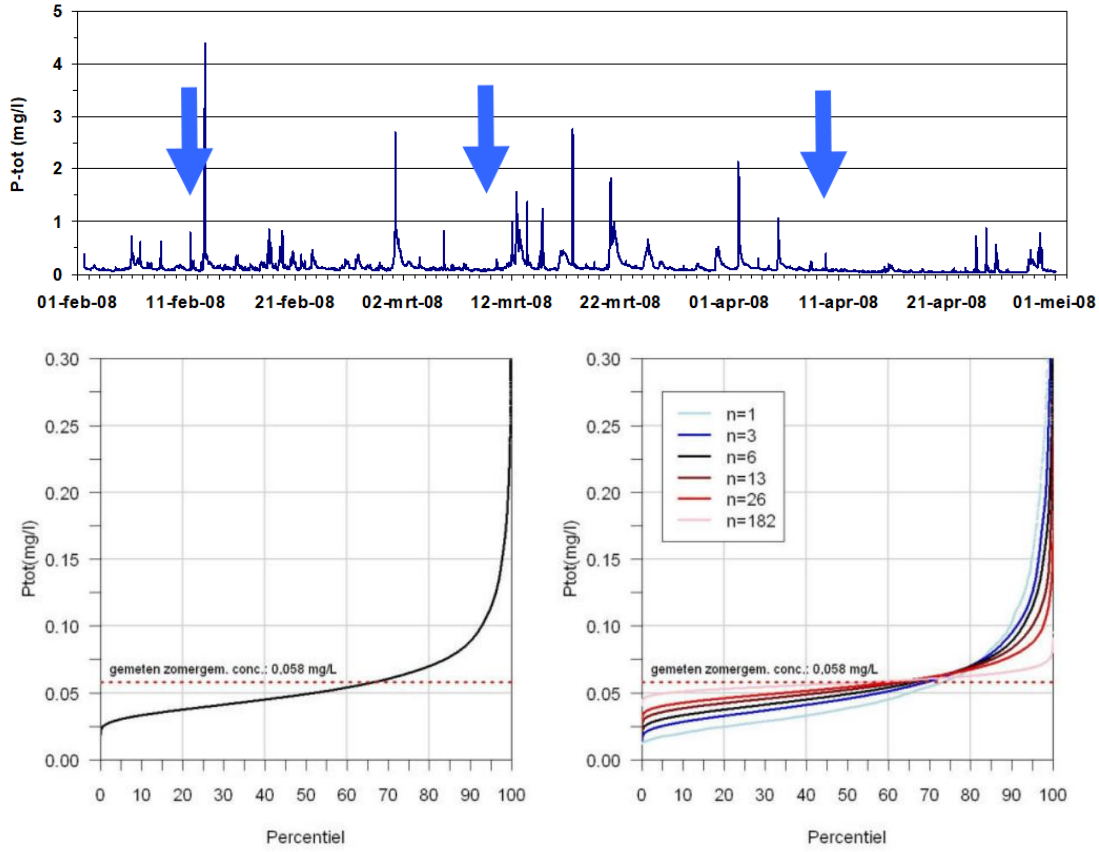
Kosten baten inschatting

Als bepaald is dat er een hoge meetfrequentie nodig is moet er vervolgens een kosten inschatting worden gemaakt. Bij kosten moet in deze context breed gedacht worden, het gaat dan niet alleen om de kosten per meting en de aanschaf van eventuele sensoren of auto-analyzers, maar ook om de beschikbare FTE's voor het uitvoeren van de meetcampagne en alles wat daar bij komt kijken inclusief het opzetten van de benodigde ICT infrastructuur. Deze kosteninschatting helpt bij het bepalen van de meerwaarde van een meetopstelling met sensoren of auto-analyzers. Het kan namelijk ook zo zijn dat het goedkoper is om heel vaak reguliere steekmonsters (of metingen met handhelds) uit te voeren dan het plaatsen van sensoren of auto-analyzers (zie Tabel 2 voor voorbeelden). Ook is een keuze mogelijk om bijvoorbeeld getriggerde autosamplers te plaatsen die pas monsters nemen indien er een afwijking wordt geconstateerd met sensoren (voor bijvoorbeeld peil, afvoer, EC). Een aantal belangrijke punten die kunnen helpen bij het maken van deze afweging zijn:

- De verwachte duur: over een langere meetperiode is de plaatsing van een meetopstelling met sensoren of auto-analyzers waarschijnlijk rendabeler.
- De ruimtelijke schaal: Hoe meer locaties hoe meer watermonsters genomen of hoe meer sensoren/auto-analyzers geplaatst moeten worden.
- Temporele variatie: zijn dag-nachtritmes of andere variaties binnen een dag belangrijk dan vallen reguliere steekmonster om praktische redenen af.
- Opschaling: is het waarschijnlijk dat er opgeschaald gaat worden dan zijn sensoren waarschijnlijk eerder rendabel dan bij een kortstondige toepassing op 1 locatie, verder heeft dit ook effect op de ICT-infrastructuur en het onderhoud. Bij auto-analyzers is dat weer minder het geval omdat ze ook een bepaalde infrastructuur nodig hebben (stroom, meetbehuizing, monstervoorbehandeling, etc).
- Snelheid data beschikbaarheid: voor bepaalde toepassingen is het prima om de data achteraf te analyseren, voor bijvoorbeeld operationele toepassingen of bij incidentele afwijkingen (bijvoorbeeld lozingen), is de data (near) realtime nodig en kan niet worden gewacht op de resultaten uit het lab. Dit heeft ook effect op de ICT-infrastructuur en data optimalisatie.

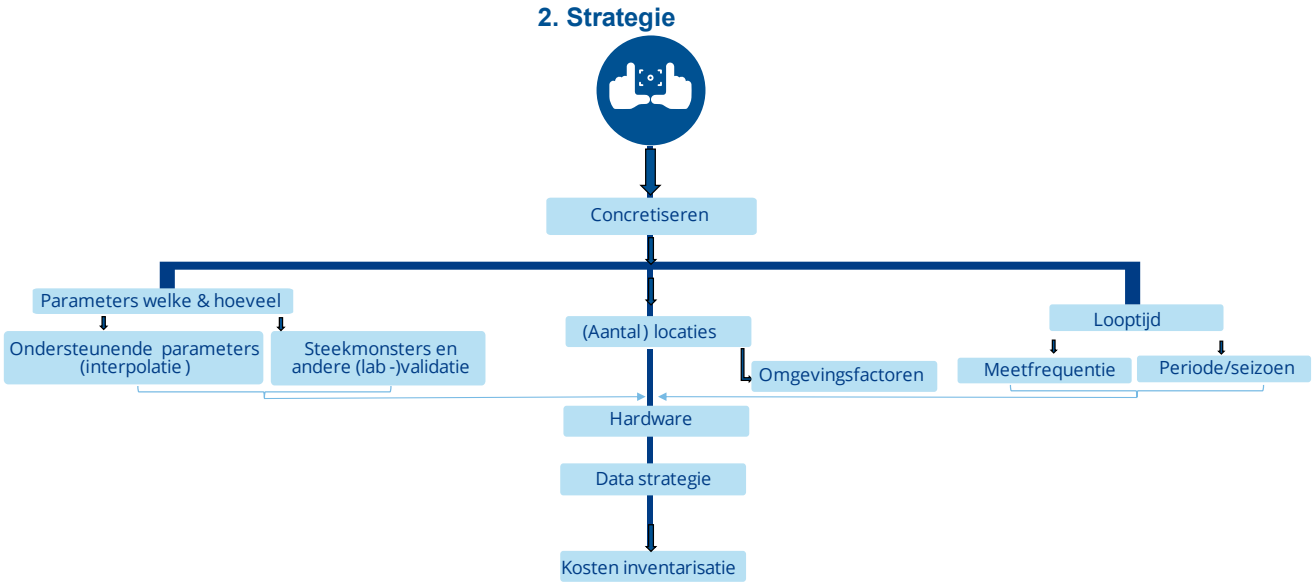
Voorbeeld

Hieronder een voorbeeld van waarom hoogfrequente metingen extra inzicht kunnen geven in de verandering van waterkwaliteit op een locatie, in dit geval de Hupselse beek.



Figuur 2: Boven: voorbeeld van een hoogfrequente meetreeks voor P-totaal (Hupselse beek, Achterhoek). De blauwe pijlen geven een reguliere maandelijkse meetfrequentie aan. Onder links: verdeling van mogelijke uitkomsten van zomergemiddelde concentraties op basis van 6 trekkingen uit de continue meetreeks. Onder rechts: verdeling van zomergemiddelde concentraties op basis van 1 tot 182 metingen (1 meting per zomerhalfjaar tot dagelijkse metingen). Zie ook Klein et al. (2012).

2 Strategie



Nadat de informatiebehoefte bekend is en er een eerste verkenning is gedaan aan de haalbaarheid (kosten, capaciteit, type en frequentie metingen) is het tijd voor de meetstrategie. Hierbij wordt vastgesteld welke parameters op welke locaties met welke frequentie en met welke meetmethodes gemeten worden.

Parameters
Bij de strategie moeten de parameters van globaal naar concreet gemaakt worden, dat wil zeggen: hoeveel parameters gaan er worden gemeten en welke precies.

- Parameters: welke en hoeveel.
- Bepalen of en zo ja hoe frequent er voor bepaalde parameters reguliere steekmonsters nodig zijn. Dit kan noodzakelijk zijn als absolute niveaus belangrijk zijn (bijv. bij bepalen van vrachten).
- Ondersteunende parameters: Voor sommige onderzoeksvragen zijn naast waterkwaliteitsmetingen ook ondersteunende parameters noodzakelijk (bv meteo, debieten, grondwaterstanden). Het kan zijn dat de benodigde gegevens al worden verzameld, anders moet er voor die data ook een meetstrategie komen.

Locaties
Bepalen en concreet maken van locaties, welke locaties wordt er bemeten en hoeveel. Daarbij komen verschillende omgevingsfactoren kijken zoals:

- Stroming
- Diepte
- Stratificatie
- Maai strategie
- Toegankelijkheid/veiligheid/bevoegdheid (mens & product)
- Opstellen veld logboek -> dit is niet afhankelijk van een meetstrategie maar moet standaard worden meegenomen (want alle sensoren hebben onderhoud nodig).

Looptijd
De looptijd concreet maken door het bepalen van**:**

- Minimale looptijd van de monitoring
- Vaststellen van relevante periodes of seizoenen
- Benodigde meetfrequentie

Hardware
Uit het concreet maken van de onderdelen kunnen verschillende keuzes gemaakt worden zoals:

- Meetbereik per parameter
- Nauwkeurigheid van de metingen per parameter
- Sensor type en fabrikant

Datastrategie
Bij de meetstrategie hoort ook al een data strategie hier in moet in ieder geval worden na gedacht over:

- Data opslag keuze (houdt ook rekening met de hoeveelheid aan data)
- ICT systemen en personeel
- Keuze data standaarden (bijv. AQUO)
- Notificaties en mogelijke opvolging
- Snelheid data beschikbaarheid (achteraf analyseren of (near)realtime)

Kosten Inventarisatie:
Maak een inventarisatie van de kosten die gepaard gaan met de aanschaf, installatie en het onderhoud van de meetapparatuur en IT-infrastructuur. Dit helpt bij het budgetteren en planning van de benodigde middelen.

Aanvullend hier op moet er bepaald worden of de materialen gekocht of gehuurd worden. Het is mogelijk om de aanschaf en het onderhoud van de materialen uit te besteden aan een externe partij (ontzorgen) die ook het hier beschreven proces gaat volgen. Ditzelfde geldt voor de data, hier kan ook voor ontzorging worden gekozen (data as a service). Belangrijk is dan wel om na te denken over vraagstukken zoals: privacy, eigenaarschap van en toegang tot de data door externe partijen.

Voorbeeld
Voor het project Sensorgestuurd boeren (waterschap Aa en Maas) was er in eerste instantie een duidelijke strategie bedacht voor het meten met sensoren. Na het uitwerken van de strategie is men begonnen met de uitvoering van de metingen. Echter was er bij de strategie voor het meten nog geen strategie voor het databeer uitgewerkt. Dit heeft er toen toe geleid dat er een aantal jaar na het beginnen met meten nog steeds veel werk in het opzetten en laten aansluiten van de ICT-infrastructuur moest worden gestopt. Dit had kunnen worden voorkomen als hier direct bij het maken van een strategie al rekening mee was gehouden. Verder is ter illustratie is hieronder een tabel weergeven met een voorbeeld van een kosten inschatting.

Tabel 2: Voorbeeld van kosten inschatting

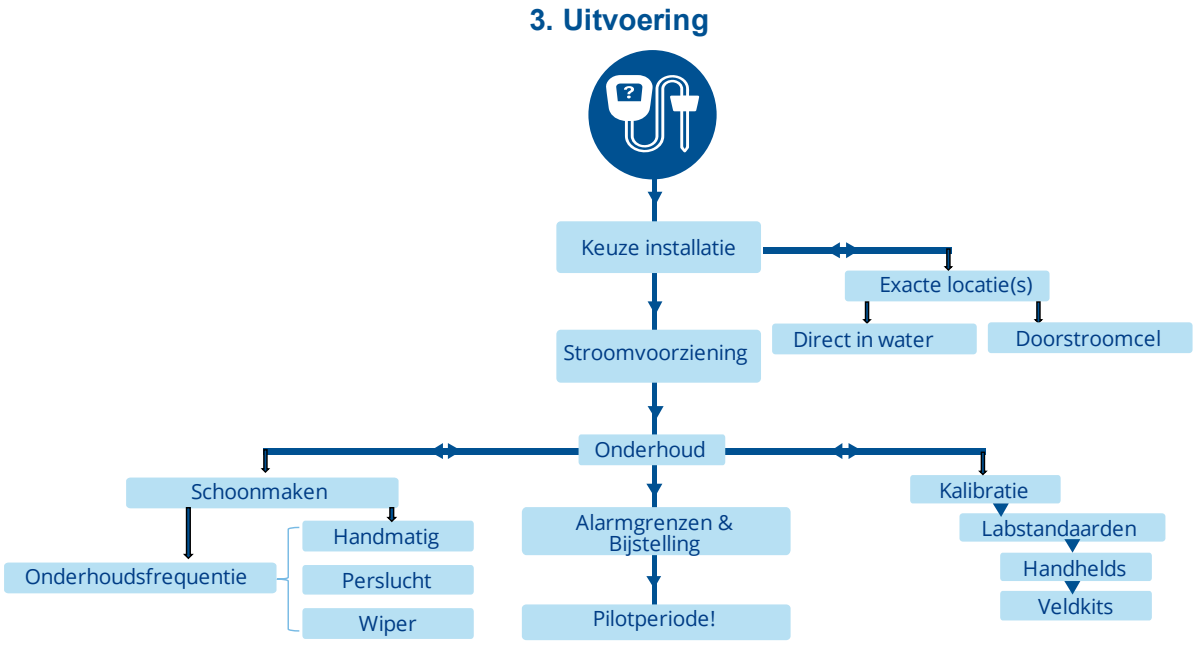
Kostenelement	Details en suggesties voor kosten en toepassingen	globale kosten (in 2024)
Kosten voor aanschaf van een sensor	Eenmalige kosten afhankelijk van het sensortype	UV-sensor[1] (Nitraat): 10-20 k€ ISE[2] (Nitraat): 5-10 k€ Auto-analyzer[3] P): 50-100 k€
Afschrijvingsperiode	Aantal jaren tot vervanging	UV-sensor[1] (Nitraat): 5-10 jaar ISE[2] (Nitraat): 0.5-5 jaar Auto-analyzer[3]: 10-15 jaar
Kosten voor het opzetten van een hoogfrequente meetopstelling	Typische installatiekosten zijn afhankelijk van de implementatiemethode, maar omvatten meestal: stroomvoorziening; materiaal voor installatie, zoals bijvoorbeeld de bouwkosten voor het plaatsen van sensoren direct in het water of in een doorstroomcel). Auto-analyzers zullen altijd in een cabine geïnstalleerd moeten worden.	Stroomvoorziening: 3-6 k€ UV[1] direct in water: 2-3 k€ ISE[2] direct in water: 1-2 k€ Cabine: 10-20 k€
Jaarlijkse bedrijfskosten	Jaarlijkse kosten voor stroom, wissers, ISE's[2], membranen, chemicaliën, enz.	UV[1]: < 0.5 k€ ISE[2]: 0.5-2 k€ Auto-analyzer[3]: 2-5 k€
Jaarlijkse kosten voor onderhoud en kalibratie	Onderhoud is sterk afhankelijk van het sensortype, maar kan bestaan uit het regelmatig handmatig reinigen van sensoren (bijvoorbeeld maandelijkse intervallen), kalibratie van de sensor, vervanging van membranen, levering van nieuwe chemicaliën en watermonsters die in het laboratorium worden geanalyseerd, bijvoorbeeld maandelijks) dit laatste is een normaal onderdeel van programma's voor steekmonsters	UV[1]: 2-5 k€ ISE[2]: 10-20 k€ Auto-analyzer[3]: 3-6 k€
Kosten voor het opzetten van een database voor sensordata en het beschikbaar stellen van data	Zou gebruik kunnen maken van de bestaande database die verder kan worden ontwikkeld om hoogfrequente gegevens op te nemen (bijvoorbeeld gegevensverzameling met een frequentie tussen 1 en 30 minuten)	10-20 k€

1 UV-sensoren zijn gebaseerd op Ultra Violet (UV) lichtabsorptie

2 ISE's zijn ionselectieve elektroden

3 Een auto-analyzer is een laboratoriumapparaat dat automatisch chemische analyses uitvoert op vloeistofmonsters

3 Uitvoering



Na dat de meetstrategie is uitgewerkt kan er een plan worden gemaakt voor de meetuitvoering. Hieronder worden de stappen en overwegingen beschreven die essentieel zijn voor een effectieve meetuitvoering.

Keuze van Installatie:
Bij het kiezen van de installatie moet worden bepaald of deze direct in het water, in een doorstoombak of in een doorstroomcel wordt geplaatst. Daarnaast is het belangrijk om te beslissen of de installatie op een vlotter (drijvend) komt te staan of niet. Hierbij moet ook aandacht besteed worden aan de veiligheid tijdens de werkzaamheden (ook bij onderhoud) en of daar nog aanvullende maatregelen voor nodig zijn (bijv. een trap, valbeveiliging etc.).

Stroomvoorziening:
De stroomvoorziening van de meetapparatuur kan op verschillende manieren worden geregeld: via een zonnepaneel, stroomkabel of accu. De keuze hangt af van de locatie, de keuze van een sensor of auto-analyzer en de beschikbaarheid van stroombronnen.

Schoonmaken:
Sommige sensoren en veel auto-analyzers hebben automatische schoonmaakroutines. Sensoren kunnen bijvoorbeeld uitgerust zijn met wissers of automatisch gereinigd worden met perslucht. Auto-analyzers hebben vaak automatische reinigingsroutines die bijvoorbeeld dagelijks doorlopen worden. Het handmatig schoonmaken van de meetapparatuur blijft ook altijd noodzakelijk. Bij een auto-analyzer wordt ook vaak een voorfiltratie gebruikt die ook weer onderhoud vergt. Regelmatig schoonmaken is cruciaal om nauwkeurige metingen te garanderen.

Onderhoudsfrequentie:
Bepaal hoe vaak onderhoud aan de meetapparatuur nodig is. Dit kan variëren afhankelijk van de omgevingsomstandigheden en de specificaties van de apparatuur. Aangezien de omgevingsomstandigheden variabel zijn, kan ook de benodigde onderhoudsfrequentie variëren. Zo kan bio-fouling bij hogere temperaturen in de zomer sneller gaan dan in de winter. Door meer sedimenttransport in natte perioden kan ook in de winter soms vaker onderhoud nodig zijn. Het is ook noodzakelijk om de onderhoudsmomenten goed te registreren en een logboek bij te houden bij voorkeur op een digitaal platform of in een asset managementsysteem om automatisering te vereenvoudigen en fouten te voorkomen. Het vastleggen van onderhoud middels foto's (bij afwijkingen of veranderingen in de omgeving) is ook een essentieel onderdeel om het onderhoud in een assetmanagement- of servicesysteem te beheren.

Kalibratie
De kalibratie van de apparatuur is afhankelijk van de te meten parameters. Dit kan worden uitgevoerd aan de hand van (gevalideerde)handsensoren, veldkits of (gecertificeerde)laboratoriumstandaarden.

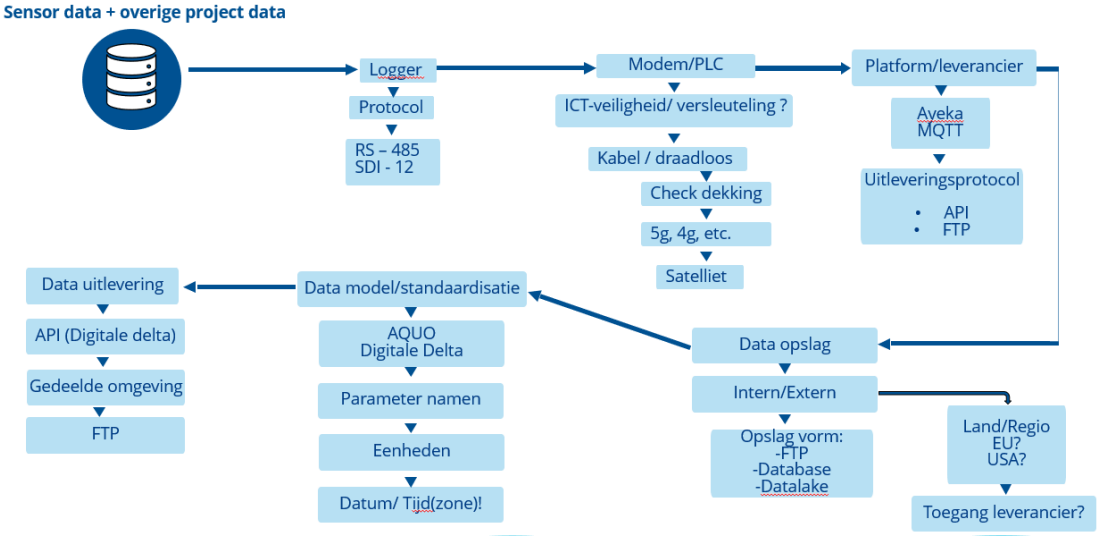
Alarmgrenzen en bijstelling
Het is essentieel om op technisch vlak de binnenkomende meetgegevens regelmatig op vooraf vastgelegde tijden te controleren en bij te stellen indien nodig, dit om er zeker van te blijven dat de meetopstelling gegevens levert van voldoende kwaliteit om te voorzien in de informatie behoefte. Dit kan gedaan worden door het bijhouden van alarmgrenzen op relevante grenswaarden. Verder kunnen triggers worden gebruikt om de onderhoud cyclus te optimaliseren, bijvoorbeeld een trigger op drift op ruis in de meetwaarden. Dit kan ook helpen met de dataoptimalisatie en een betere inschatting geven van de service behoefte.

Opstartperiode/Pilotperiode:
Er moet rekening worden gehouden met een opstartperiode of pilotperiode waarin de meetapparatuur wordt ingeregeld en geoptimaliseerd. De duur van de opstartperiode of de noodzaak van een specifieke pilotperiode is afhankelijk van de informatiebehoefte, de meetstrategie en de specifieke omgevingsfactoren. Binnen deze pilotperiode zal er naast ervaring met de technische kant van het meten ook de eerste ervaring worden opgedaan met de data die binnen komt en de mogelijke uitschieters in de data. Dit helpt daarmee ook om de volgende stappen in het de cyclus in meer detail uit te werken.

Voorbeeld
Binnen het project Sensorgestuurd Boeren bleek vervuiling van de lenzen voor optische sensoren als de UV-absorptienitraatsensor een probleem (Figuur 3). Regelmatig reinigen is noodzakelijk om goede metingen te krijgen. Sommige typen sensoren zijn daarom uitgerust met een kleine mechanische wisser. Deze wisser kan echter zelf ook verontreinigd raken of zelfs klem komen te zitten als de sensor niet goed beschermd is tegen drijfvuil. Het kan ook mis gaan als er wat hardere deeltjes in de vervuiling zitten: in combinatie met de wisser kan dit leiden tot krassen op de lenzen en daarmee tot onbetrouwbare metingen door verstrooiing van de lichtbundel. Bij verschillende sensoren is daarom ook automatische reiniging met perslucht mogelijk. Afgezien van deze geautomatiseerde reiniging met wissers or perslucht, blijft ook het handmatig schoonmaken nodig. Meestal is maandelijks handmatig onderhoud voldoende, maar in een periode met veel sedimenttransport kan onderhoud vaker nodig zijn. Dit voorbeeld is afkomstig uit het H2O artikel *Praktijkervaringen met nitraatsensoren in oppervlaktewater* hierin zijn nog meer voorbeelden te vinden.

Figuur 3: Een UV-nitraatsensor met vervuiling in een ijzerrijke kwelsloot in een zandgebied. De wisser is niet krachtig genoeg om de aangroei weg te halen en veroorzaakte daarbij krassen op de lenzen (links). Rechts na schoonmaak van de sensor (foto's: Aquon).

4 Data beheer



Als de meetuitvoering goed is uitgedacht is de volgende stap om na te denken over de opslag en het beheer van de gegevens.

Logger
Hierbij is het van belang om te beginnen op het laagste niveau wat in bijna alle gevallen een logger zal zijn (tegenwoordig vaak al gecombineerd met een modem). Zoals de naam al zegt logt deze de gegevens van de sensor en kan deze doorsturen naar een extern apparaat, meestal een modem of een computer/programmable logic controller (PLC) of een combinatie van de twee. De logger zal dit altijd doen via een bepaald communicatieprotocol (bijvoorbeeld Modbus, SDI-12 of RS-485). Het is belangrijk om te weten welk protocol van toepassing is om goed te kunnen communiceren met externe apparatuur zoals een PLC of (cloudbased) dataplatform. Omdat de gegevensoverdracht in de praktijk soms verstoord wordt, is het raadzaam de meetgegevens ook lokaal op de meetlocatie voor een langere periode op te slaan.

Modem
Het volgende niveau in de data stroom is de modem of PLC (eventueel met geïntegreerde modem). Dit kan worden gezien als de computer die met de meetopstelling communiceert en deze eventueel bestuurd.
Net als bij alle ICT systemen is het ook hier van belang om na te denken over de ICT-veiligheid en eventuele versleuteling van de data. De mate van beveiliging is afhankelijk van de context waarbinnen de data verzameld en gebruikt zal worden en van eisen van eventuele projectpartners.
Om de data van de modem/PLC te krijgen naar de locatie waar er mee gewerkt zal worden is het nodig om de data te versturen. Daarom is het nodig om na te gaan welke mogelijkheden beschikbaar zijn. In specifieke gevallen kan dit via een bekabelde verbinding maar tegenwoordig zal dit meestal draadloos zijn. Wanneer de data draadloos wordt verstuurd is het belangrijk om na te gaan of er op de meetlocatie dekking is voor het juiste netwerk (4g, 5g, Lora, NB-IoTect.). In het uiterste geval kan het ook via een satelliet verbinding.

Leverancier
Tussen de meetopstelling en de gebruiker van de gegevens zit meestal nog de leverancier van de apparatuur (modem/PLC). De leveranciers hebben vaak hun eigen platform voor het uitleveren van de gegevens. Elk platform heeft zijn voor- en nadelen en het is belangrijk om deze voor- en nadelen af te wegen. Voor het kunnen combineren van meetgegevens vanuit verschillende apparaten of voor specifieke databewerking is het veelal nodig de data door te sturen naar een eigen database en/of platform.

Uitleveringsprotocol
Een essentieel onderdeel van het uitwisselen van meetgegevens is of het uitleveringsprotocol (bijv. MQTT, S-FTP of API) goed aansluit bij de bestaande ICT-infrastructuur. Het kan veel extra werk en tijd schelen als het uitleveringsprotocol goed aansluit bij ICT-infrastructuur die al aanwezig is. Tegenwoordig zijn hiervoor gelukkig veel mogelijkheden.

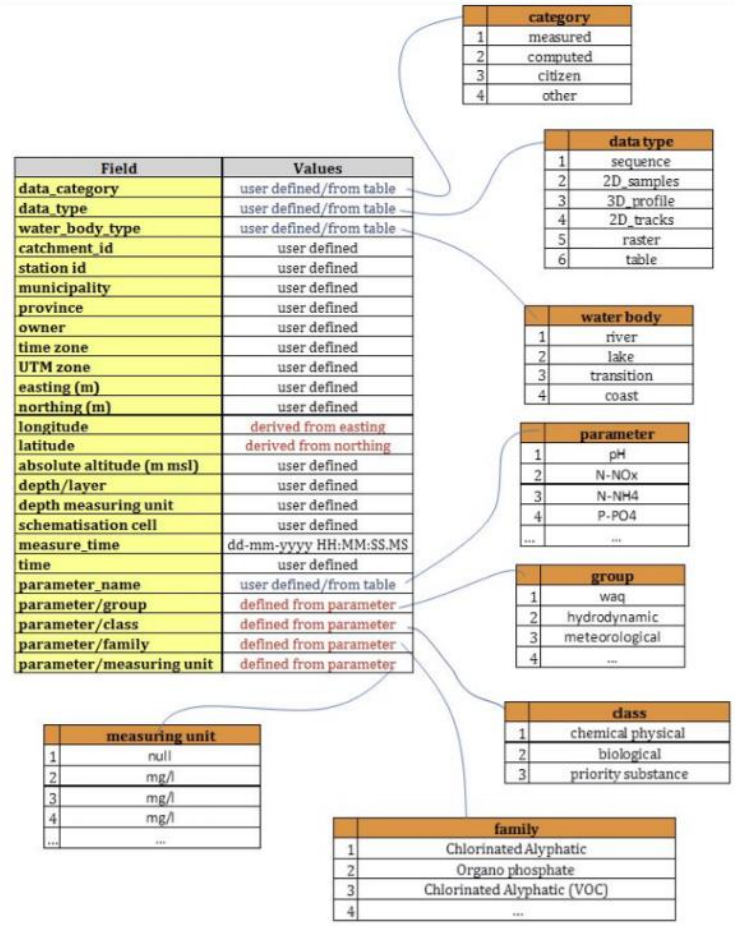
Data-opslag
Voor het opslaan van de data kan het ook zo zijn dat de leverancier hier een oplossing voor biedt. Het voordeel hiervan kan zijn dat dit zorgen over dataopslag wegneemt. Echter kan het ook nadelig zijn om data extern op te slaan. Voor bepaalde (privacy)regels of organisaties kan het namelijk van belang zijn in welk land of regio de data wordt opgeslagen (EU, USA etc.) en of de leverancier ook toegang heeft tot de data. Daarnaast is het mogelijk dat leveranciers de data maar voor een beperkte periode opslaan.
Als de data intern wordt opgeslagen is dit niet van toepassing en verder geeft het ook meer controle over de data, echter kost dit waarschijnlijk wel meer tijd en middelen. Het is dan ook van belang om na te denken over de vorm van de data opslag, enkele voorbeelden zijn een database, datawarehouse of datalake.

Data model
Bij het opslaan van data is het belangrijk om goed na te denken over het standaardiseren van de data (het datamodel). Zeker als de toepassing op grotere schaal wordt gedaan scheelt het veel tijd en moeite als alle parameters, eenheden, locaties, etc. op de zelfde manier worden genoteerd. Verder zijn er binnen het waterkwaliteitsdomein al standaarden waarbij aangesloten kan worden. De meeste gebruikte in Nederland is de AQUO-standaard die ook wordt toegepast binnen de Digitale Delta (DD). Wij adviseren om hier bij aan te sluiten. Dit scheelt veel tijd en moeite bij het bedenken van een datamodel en maakt het uitwisselen van de data ook een stuk makkelijker, wat ook externe samenwerking makkelijker maakt. Een extra punt van aandacht is het goed opslaan van de tijd en de tijdzone. De meerwaarde van hoogfrequente data is dat het de mogelijkheid bied om te vergelijken met specifieke events zoals regenval. Hierbij kan een uur tijdsverschil door een fout in de tijdzone opeens een groot verschil maken. Dit is ook belangrijk als externe labdata in het dataplatform wordt toegevoegd (bijvoorbeeld voor de vergelijking tussen sensordata met conventionele metingen).

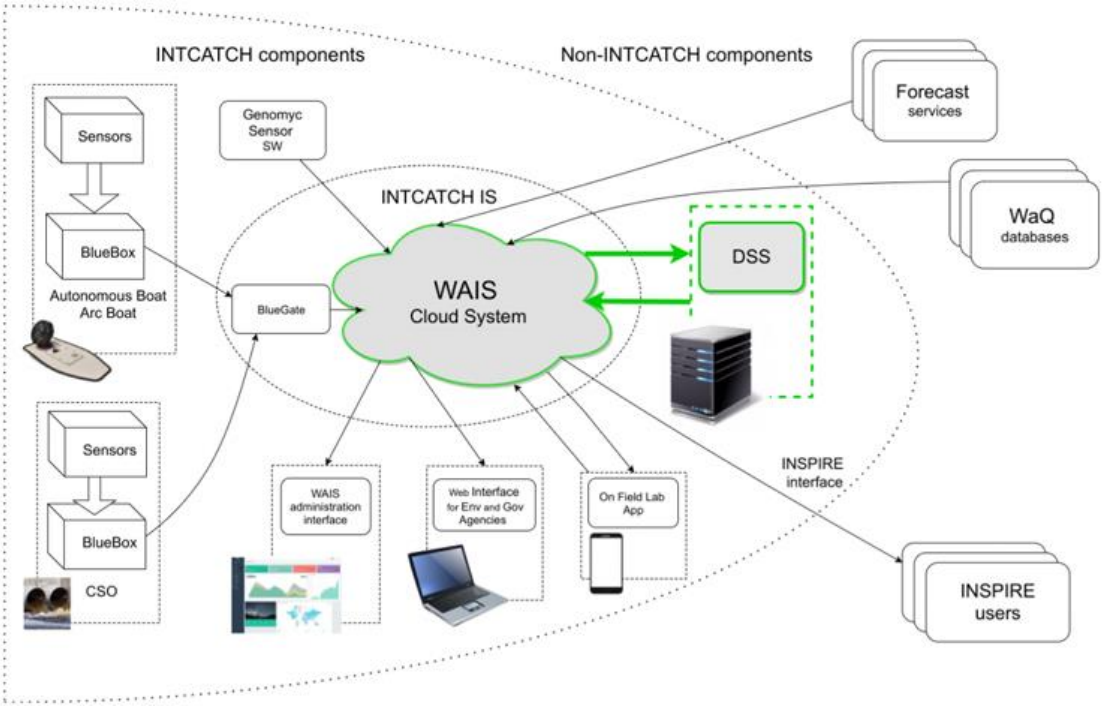
Data uitlevering
De laatste stap is de uitlevering van de data. Dit kan via verschillende manieren zoals een S-FTP, een gedeelde digitale omgeving of een API. Het is belangrijk om goed aan te sluiten bij de bestaande ICT-infrastructuur en ook hier is het verstandig om aan te sluiten bij bestaande standaarden zoals de Digitale Delta API.

Voorbeeld

Voor verschillende projecten wordt data verzameld en gecombineerd vanuit verschillende bronnen die met verschillende methoden verzameld zijn. Voor het snel kunnen raadplegen van deze gegevens, maar ook om de gegevens toegankelijk te kunnen maken voor derden, is opslag in een logische datastructuur van belang. Een voorbeeld van zo'n datastructuur is weergegeven in Figuur 4. Deze structuur uit het Horizon 2020 project INTCATCH (Di Donato et al., 2017) biedt ruimte voor verschillende datatypen, categorieën, parameters, etc. Als voorbeeld is ook de INTCATCH structuur voor de communicatie tussen sensoren, databases en gebruikers gevisualiseerd in Figuur 5. De data kan vanuit verschillende typen sensoren worden voorbewerkt en vullen de database. Daarnaast wordt de database gevoed met data uit externe databases. De cloud database wisselt gegevens uit met verschillende (online) services en apps waarin gebruikers de gegevens kunnen raadplegen en/of kunnen bewerken.



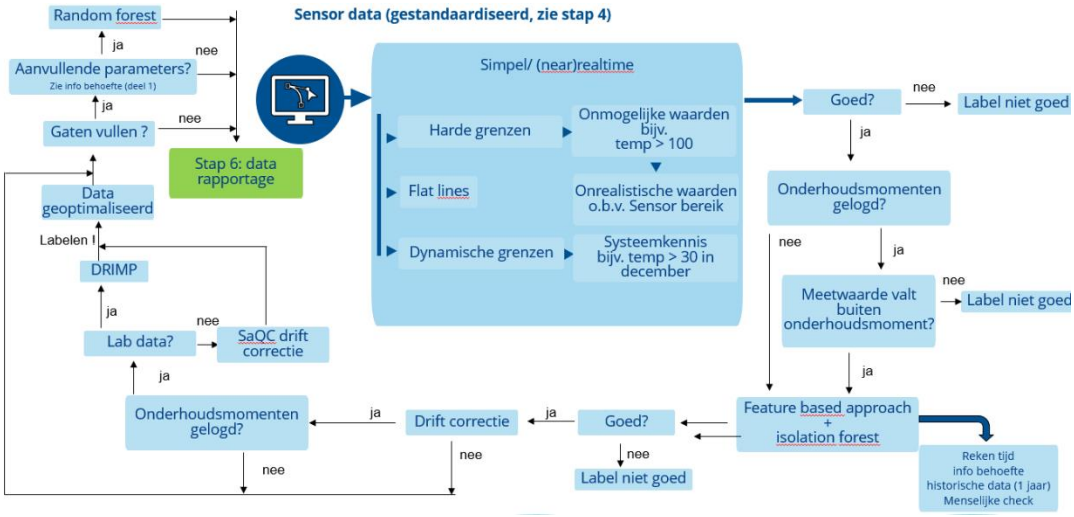
Figuur 4:Voorbeeldstructuur voor dataopslag uit het H2020 INTCATCH project (Di Donato et al., 2017)



Figuur 5:Voorbeeld van de dataverbindingen tussen sensoren, databases, services en gebruikers uit het H2020 INTCATCH project (Di Donato et al., 2017)

5 Data optimalisatie

Binnen het Optima-HWQ project is de focus vooral de optimalisatie van de gegevens die worden gegenereerd bij hoogfrequente waterkwaliteitsmetingen. Daarom is deze stap (5 data optimalisatie) in meer detail uitgewerkt dan de overige stappen.



De data optimalisatie gebeurd in meerdere stappen. Hier wordt met data optimalisatie het gehele proces bedoelt van de opwerking van ruwe data tot de uiteindelijk data die gebruikt kan worden voor de beoogde toepasing. Vaak gebeurd dit globaal in 3 stappen. Eerste wordt de ruwe data opgeslagen precies in de vorm (vaak verschillende formats) zoals deze wordt geleverd (‘as is’). Vervolgens wordt de data gestandaardiseerd naar eenzelfde format (bijv. AQUO) en worden eenheden per parameter omgerekend naar dezelfde eenheid. Vervolgens zullen er aan de hand van verschillende technieken uitschieters worden gemarkeerd en gelabeld. Verder zullen er, indien nodig, ook waarden worden aangepast door middel van driftcorrecties of het invullen van gaten. Belangrijk om hierbij op te merken is dat de data hierbij nooit volledig wordt verwijderd en er van elke stap apart een kopie wordt opgeslagen. Verder wordt er voor de verschillende toegepaste correctietechnieken aan elke meetwaarde een toepasselijk label gegeven. Dit kan zijn dat de meetwaarde als een uitschieter wordt beschouwd, maar bijvoorbeeld ook dat de meetwaarde is gecorrigeerd voor drift. De stappen van het optimaliseren van de data in dit rapport gaan van simpele naar meer complexe technieken. Deze simpele technieken worden als “simpel” beschouwd omdat ze door hun eenvoud niet tijds- en arbeidsintensief en zijn het dus altijd waard om te gebruiken en zijn ook geschikt voor geautomatiseerde (near)realtime toepassingen. Meetwaarden die eerder met de meer simpele technieken zijn gelabeld als uitschieter worden dan voor de meer complexere technieken ook buiten beschouwing gelaten.

Simpel / near real time

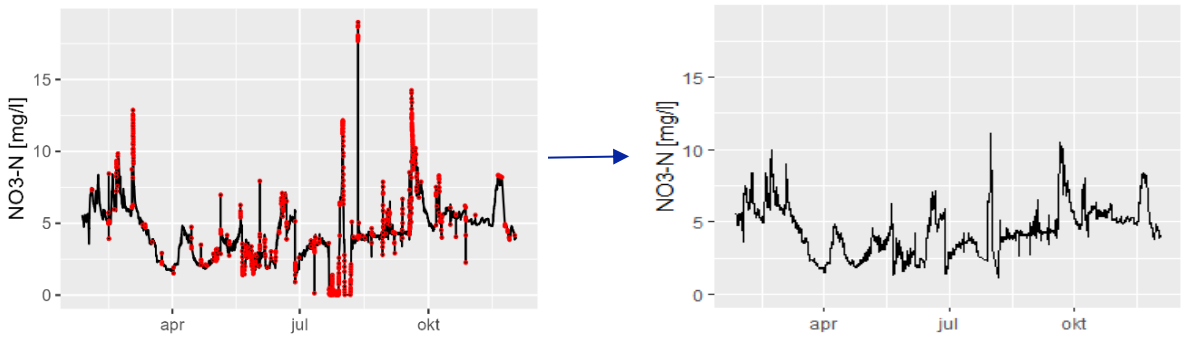
De eerste simpelere technieken die toegepast kunnen worden zijn het stellen van harde fysiek onmogelijke grenzen per parameter. Er kan hierbij gedacht worden aan een temperatuur range van bijvoorbeeld -5 tot 100 graden of een pH tussen de 0 en 14. Dit is een simpele manier om de meest opvallende uitschieters te verwijderen.
Echter wordt het stellen van harde grenzen wel ingewikkelder met het toenemen van het aantal parameters. Een alternatief dat makkelijker te automatiseren en op te schalen is, is door gebruik te maken van het aangegeven meetbereik de sensoren. In het meest gunstige

geval wordt dit gedaan in combinatie met een asset management systeem waar alle sensoren met meetbereik in zijn opgenomen.
Deze harde grenzen zijn nog verder te verfijnen door er aan de hand van systeemkennis ook dynamische grenzen aan toe te voegen, zoals bijvoorbeeld een range voor de temperatuur per maand of de pH-range op een specifieke locatie.

Een andere simpele methode is het detecteren en labelen van zogenoemde flatlines. Hierbij blijft de meetwaarde als het ware hangen op continue dezelfde waarde. Dit wordt vaak veroorzaakt door elektronische storingen. Deze flatlines zijn relatief eenvoudig te verwijderen met een algoritme (zie voorbeeld in bijlage B). Er moet wel rekening mee gehouden worden dat voor sommige parameters flatlines ook daadwerkelijk voor kunnen komen. Bij cumulatieve neerslag is het bijvoorbeeld normaal dat veel metingen dezelfde waarde hebben als het lang droog is (er komt immers geen regen bij) en ook als een parameter dicht bij de rapportagegrens zit kunnen meerdere opvolgende metingen dezelfde waarden hebben.
De combinatie van de boven genoemde technieken zijn door hun eenvoud niet tijds- en arbeidsintensief en zijn het dus altijd waard om te gebruiken en zijn ook geschikt voor geautomatiseerde (near)realtime toepassingen. Echter zal altijd aangegeven moeten worden waar en welke technieken gebruikt zijn om data aan te passen zodat het duidelijk en aantoonbaar is dat data is geoptimaliseerd.

Onderhoudsmomenten
Een andere techniek om uitschieters te markeren is door meetwaarden tijdens momenten van onderhoud te markeren als uitschieters. Met momenten van onderhoud wordt in dit geval bedoelt onderhoud aan de meetopstelling zelf, uitgevoerd door de gebruikers van de meetopstelling en niet extern onderhoud zoals bijvoorbeeld maaiwerkzaamheden. Omdat het onderhoud altijd voor verstoringen zorgt is dit een belangrijke manier om uitschieters te labelen. Echter kan dit met veel meetopstellingen arbeidsintensief zijn. Het is daarom van belang om hier van te voren over na te denken hoe dit uitgevoerd kan worden. In de meeste ideale situatie wordt dit gedaan in combinatie met een asset (of service)management systeem waarbij alle onderhoudsmomenten direct (bijvoorbeeld via het lezen van een QR code) gelogd kunnen worden. Op deze manier is dit proces makkelijk te automatiseren en kan het ook geschikt gemaakt worden voor (near)realtime toepassingen. Let op; een assetmanagement systeem is waarschijnlijk kostbaarder dan alleen een servicesysteem of service app.

Geavanceerde technieken
Een meer geavanceerde techniek voor anomalie-detectie (detecteren van uitschieters) is de ‘feature based’ aanpak (Talagala et al., 2019). Op basis van de meetreeks worden reeksen van meerdere eigenschappen (features) van de metingen berekend, zoals de log-transformatie, de eerste afgeleide en de helling, waardoor eventuele uitschieters ten opzichte van de andere metingen eerder opvallen. Waar Talagala et al. (2019) een k-nearest neighbor (kNN) algoritme gebruiken om vervolgens anomalieën te detecteren, wordt in deze studie het snellere en effectievere Isolation Forest aanbevolen. Deze methode kijkt hoe makkelijk een datapunt met gebruik van rechte lijnen is te isoleren van de rest van de wolk van datapunten in de dataset (in meerdere dimensies, afhankelijk van het aantal features), zie bijlage A.

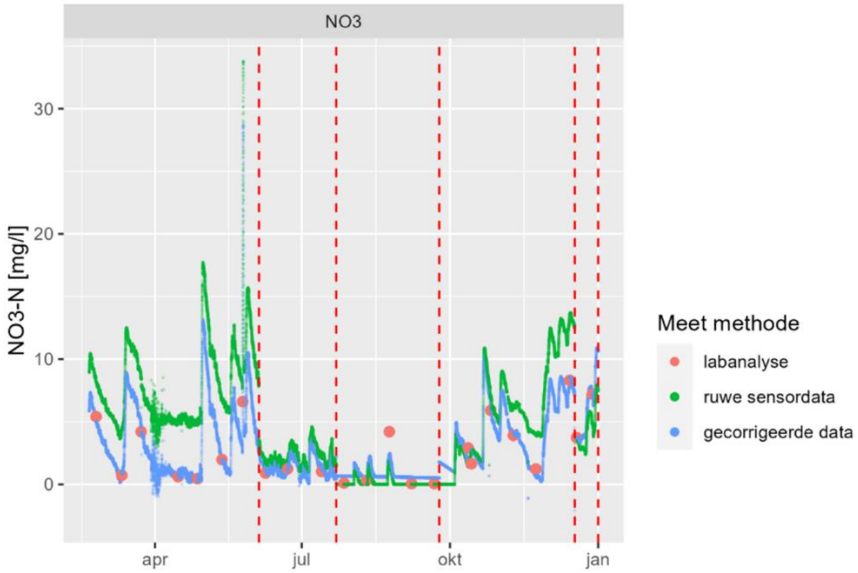


Figuur 6: Anomaliedetectie met de Feature based aanpak voor een continue meetreeks van nitraatconcentraties in oppervlaktewater (Vinkenloop, Noord Brabant). Datapunten in het rood zijn als mogelijke uitschieters gelabeld.

Figuur 6 geeft een voorbeeld van anomaliedetectie met deze methode voor een nitraatconcentratie meetreeks. Het is belangrijk te realiseren dat ook deze geavanceerde methode niet perfect werkt en het is in het begin van de toepassingsperiode ook zeker gewenst om een menselijke controle op de resultaten uit te voeren. Vooral incidentele pieken, die bijvoorbeeld bij riool overstorten of andere lozingen goed mogelijk zijn, kunnen onterecht als anomalie aangewezen worden. Bij twijfel kan een reactie van andere sensoren (andere parameters, naburige locaties, waterstanden, neerslag) duidelijk maken of er sprake was van een echte verandering in de waterkwaliteit. Met een hogere meetfrequentie (bijvoorbeeld elke 5 minuten) is vaak ook beter onderscheid te maken tussen pieken door sensor-storingen (individuele uitschieters) en echt optredende concentratiepieken met een veelal natuurlijkere verloop en meerdere meetwaarden. Verder is het ook belangrijk om op te merken dat de eerdere genoemde simpele technieken eerst zijn toegepast en de ‘feature based’ aanpak vervolgens is toegepast op de resterende data die niet als uitschieter is gemarkeerd. Deze methode vergt ook meer rekentijd en historische data dan de eerder genoemde simpeler technieken.

Drift correctie

Nadat de uitschieters zijn gemarkeerd en buiten beschouwing gelaten kunnen worden kan er een drift correctie uitgevoerd worden. Drift is een kleine continue verandering in de weergegeven meetwaarden van een meetinstrument of meetsysteem binnen een bepaald tijdsverloop .Voor de correctie van drift en jumps is een eigen methode ontwikkeld (DRIMP-correctie), waarbij we ervan uitgaan dat er bij de sensorlocatie ook reguliere waterkwaliteitsmonitoring plaatsvindt. Er bestonden al wel methoden voor driftcorrectie (bijv. Schmidt et al., 2022), maar de combinatie met conventionele controlemetingen was met openbare methoden nog niet mogelijk. De DRIMP correctie splitst de meetreeks op in stukken tussen de sprongen die veroorzaakt zijn door onderhoudsmomenten (rode stippellijnen) (zie Figuur 7). Voor elk stuk van de reeks maken we een lineair model van het verschil tussen de sensorwaarde (groene lijn) en de conventionele labanalyse (rode stippen). Met het lineaire model is vervolgens de gecorrigeerde reeks (blauwe lijn) te berekenen. Bij deze methode is het van belang dat het absolute concentratie niveau relevant is, er voldoende conventionele metingen beschikbaar zijn en dat onderhoudsmomenten goed zijn geregistreerd. Bemonsteringen tijdens concentratiepieken zijn alleen bruikbaar als het bemonsteringstijdstip nauwkeurig aan de sensorreeks te linken is.



Figuur 7: Voorbeeld van correctie voor drift en jumps in een continue meetreeks van nitraatconcentraties in oppervlaktewater (Vuursteentocht, Flevoland).de rode stippellijnen geven de onderhoudsmomenten weer.

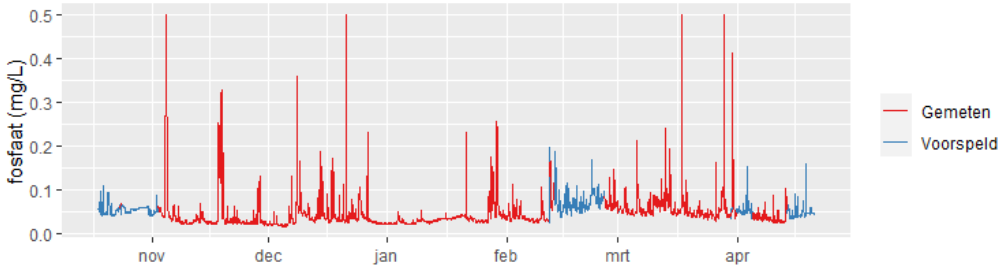
Indien er geen of niet voldoende reguliere metingen beschikbaar zijn of het absolute concentratie niveau niet belangrijk is kan er ook een driftcorrectie uitgevoerd worden zonder conventionele controlemetingen. Deze mogelijkheid is onder meer geïmplementeerd in het python pakket SaQC (Schmidt et al., 2022). Bij deze methode kan er gekozen worden tussen een lineaire of exponentiele correctie tussen de onderhoudsmomenten. Ook hier is dus een goede registratie van de onderhoudsmomenten nodig.

Opvullen gaten

Voor verschillende onderzoeksdoelen kan het nodig zijn om gaten in de meetreeks in te vullen. Dit invullen kan bijvoorbeeld nodig zijn voor een goede inschatting van verontreinigingsvrachten (concentratie maal afvoer). Eenvoudige manieren voor het opvullen van gaten in de meetreeks zijn bijvoorbeeld het door laten lopen van de laatst gemeten waarde, het lineair interpoleren tussen de laatste en de eerstvolgende meetwaarde of het invullen van het gat met de gemiddelde waarde van de reeks. Jones et al. (2022) gebruiken een tijdreeksanalysemodel (ARIMA) om datagaten in te vullen. Deze methoden werken vooral bij korte gaten. Bij grotere gaten in de meetreeks helpt het om de relatie tussen verschillende continu gemeten parameters te gebruiken. Dit kan bijvoorbeeld met een lineair regressiemodel tussen de concentraties en de afvoer. Als dit lukt zijn de concentraties voor de datagaten te berekenen op basis van de afvoer.

Random Forest (voor het opvullen van gaten)

In veel gevallen zijn de relaties echter niet lineair en ook niet constant in de tijd. Barcala et al. (2023) hebben een methode gepubliceerd waarbij de relatie tussen verschillende continu gemeten parameters gevangen wordt in een Random Forest model. Dit algoritme kan goed omgaan met niet-lineariteit kan vooral binnen de periode waarvoor trainingsdata beschikbaar zijn nauwkeurige voorspellingen doen. In Figuur 8 is een voorbeeld waar voor een meetreeks van fosfaat gaten zijn opgevuld op basis van de relatie met nitraat, troebelheid, afvoer, grondwaterstand en neerslag. Hier is voor deze methode dus van belang dat er extra parameters beschikbaar zijn om relaties me te leggen.



Figuur 8: Voorbeeld van het invullen van datagaten op basis van relaties tussen fosfaat en andere continue gemeten parameters (afvoer, grondwaterstand, neerslag, nitraat, troebelheid) die zijn vastgelegd in een Random Forest model (data uit sloot bij Winterswijk). De voorspelde waarden zijn alleen zichtbaar als er geen metingen zijn; anders ligt de blauwe lijn door de goede voorspelling (r^2=0,97) achter de rode lijn (Barcala et al., 2023).

Voor een uitgebreider overzicht van verschillende statistische methoden om uitschieters te detecteren zie bijlage A.

6 Data rapportage

De laatste stap in het proces van waterkwaliteitsbewaking is de data rapportage. Deze fase is essentieel voor het presenteren van de verzamelde en geoptimaliseerde gegevens op een manier die aansluit bij de behoeften van de eindgebruikers, zoals beleidsmakers, onderzoekers en operationele teams. Hieronder zijn een aantal opties weergeven voor drie verschillende categorieën:

- **Real time visualisatie & dashboards**:
  - Power BI
  - Tableau
  - R-shiny (ook voor Python beschikbaar)
  - Python Dash

- **Onderzoek**:
  - Rapporten (voor interne of externe toepassing)
  - (wetenschappelijke) artikelen

- **Overig gebruik**:
  - Kennis over effect van weersvariaties en -extremen op waterkwaliteit
  - Input voor modellen (aanscherpen procesparameters, kalibratie)
  - Proxy voor andere parameters (interpolatie tussen laagfrequente metingen andere parameters op dezelfde locatie) en andere locaties (beter begrip dynamiek in laagfrequente metingen op andere locaties).
  - Influent of proces controle
  - Detecteren ongewenste lozingen
  - Kwantificeren processen (zoals retentie uit dag-nachtritmes)
  - Kwantificeren bronnen (bijv. kwel versus uitspoeling)
  - Kwantificeren effect van maatregelen

Tot slot wordt hier de cirkel weer rond gemaakt door terug te gaan naar stap 1. Zijn alle onderzoeksvragen beantwoord en is er in de informatie behoefte voorzien. Indien nodig kunnen er aanpassingen gedaan worden in de relevante stappen in de cyclus.

A Literatuur studie

Memo

Datum
30 januari 2023

Aantal pagina's
1 van 40

Contactpersoon
Kevin Ouwerkerk
Joachim Rozemeijer

Doorkiesnummer
+31(0)88 335 7458
+31(0)6 22748708

E-mail
Kevin.Ouwerkerk@deltares.nl
Joachim.Rozemeijer@deltares.nl

Onderwerp
Optima-HWQ

1 Introductie

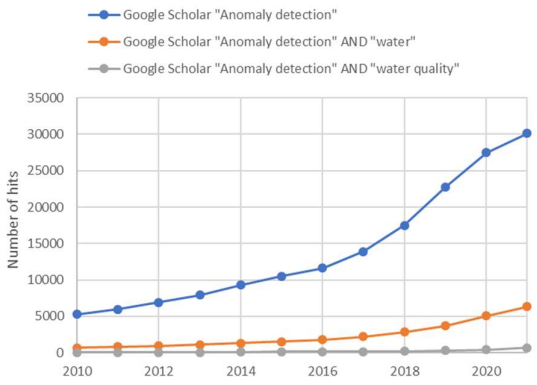
Het project Optima-HWQ (Deltares, Aquon, Waterschap Aa en Maas, MicroLan) heeft als doel om betrouwbare data te genereren uit waterkwaliteit sensoren door optimalisatie-routines te ontwikkelen. In de startfase van het project zijn internationale contacten die werken met hoge resolutie waterkwaliteitsdata geïnterviewd en is een literatuurstudie uitgevoerd. Hier presenteren we de belangrijkste bevindingen van deze verkenningsfase.

2 Algemene impressies

We hebben 3 algemene impressies, 1 algemene waarschuwing en 2 resterende ontwikkelkansen geformuleerd op basis van onze interviews en de literatuurstudie.

Algemene indruk 1
Technieken voor automatische anomaliedetectie, smoothing en ruiscorrectie zijn ruim beschikbaar, maar nog niet veel toegepast op waterkwaliteit, waarschijnlijk omdat waterkwaliteitssensoren nog maar beperkt zijn toegepast.

Figuur 1 visualiseert deze indruk voor Google Scholar hits op “Anomaly detection”, “Anomaly detection” AND “water”, en “Anomaly detection” AND “water quality”. De algemene belangstelling voor het onderwerp anomalie detectie groeit in onze digitaliserende samenleving met vooral veel toepassingen in de financiële sector, de gezondheidszorg en de industrie. Toepassingen in de waterwetenschap blijven ondertussen achter en toepassingen in de waterkwaliteitswetenschap zijn zeer beperkt.



Figuur 1: Google Scholar hits per jaar voor “Anomaly detection”, “Anomaly detection” AND “water”, en “Anomaly detection” AND “water quality”.

Algemene indruk 2
De meeste sensortoepassingen voor waterkwaliteit zijn afkomstig van wetenschappelijke groepen die hun eigen scripts gebruiken om achteraf hun gegevens te corrigeren en hun procedures voor gegevensverwerking niet rapporteren.

Publicaties met sensordata over waterkwaliteit komen steeds vaker voor in de wetenschappelijke literatuur, maar bijna geen van deze publicaties rapporteert over de procedures voor het opschonen of corrigeren van gegevens. Mogelijk is de dataverwerking ook niet op een gestandaardiseerde manier gedaan of niet goed gedocumenteerd. De publicaties met sensordata hebben meestal een andere focus en de gegevensverwerking wordt begrijpelijkerwijs niet of heel kort vermeld om het verhaal bondig en leesbaar te houden.

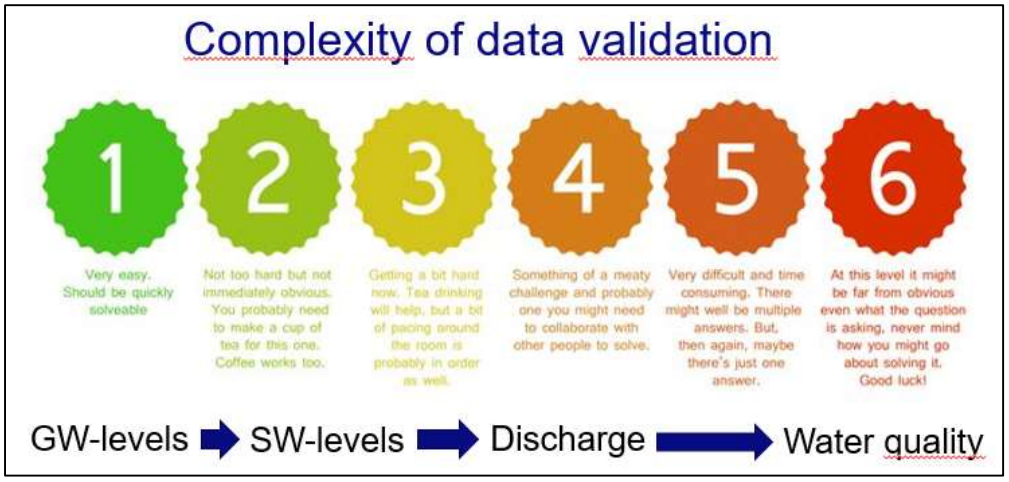
Algemene indruk 3
Toch staan we niet alleen, we kunnen leren van enkele frontrunners en we kunnen profiteren van ontwikkelingen op andere vakgebieden.

In de afgelopen jaren hebben verschillende onderzoeksgroepen open source methodebeschrijvingen en instrumenten gepubliceerd voor de verwerking van sensordata van waterkwaliteit (figuur 2). Meer informatie over hun werk staat in het volgende hoofdstuk.

Figuur 2: Verspreide onderzoeksgroepen die open source tools voor de verwerking van sensordata van waterkwaliteit hebben gepubliceerd

Algemene waarschuwing
Datavalidatie en -correctie voor sensordata van waterkwaliteit is zeer complex; foutieve metingen zijn moeilijk te onderscheiden van echte variaties in de concentraties.

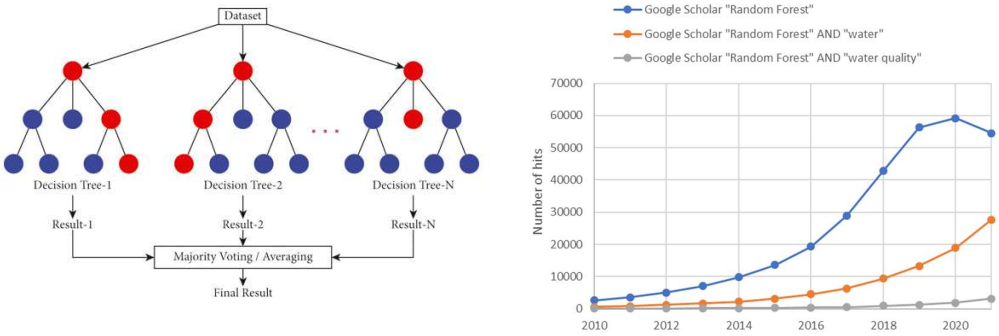
De zeer variabele en moeilijk te voorspellen concentraties in water maken de validatie van gegevens van sensoren voor waterkwaliteit complexer dan die van andere soorten hydrologische gegevens, zoals gevisualiseerd in figuur 3. Ook binnen de verschillende typen waterkwaliteitssensoren zijn er verschillen, variërend van relatief robuuste sensoren voor Electrische Conductiviteit (EC) tot de veel meer foutengevoelige Ion Selectieve Electroden.



Figuur 3: De complexiteit van gegevensvalidatie geprojecteerd op de enigma rotor scale (voor de complexiteit van puzzels, bron: https://www.gchq.gov.uk/information/enigma-rotor-scale); de complexiteit neemt toe van grondwaterstanden (gw-levels) via oppervlaktewaterpeilen (sw-levels) en afvoermetingen (Discharge) tot waterkwaliteitsgegevens (Water Quality)..

Resterende ontwikkelingskans 1
Toepassing van Random Forest voor het opsporen van uitschieters en het opvullen van gaten.

Random Forest is een veelbelovend machine-learning algoritme voor validatie van sensordata van waterkwaliteit en het opvullen van gaten in de meetreeks. De oorspronkelijke wiskundige techniek werd al in 2001 gepubliceerd (Breiman, 2001). De toepassing ervan in de waterwetenschap neemt snel toe (figuur 4), maar de toepassing in onderzoek naar waterkwaliteit is nog beperkt en spitst zich toe op data uit conventionele waterkwaliteitsmonitoring en vanuit remote sensing.



Figuur 4: Visuele uitleg van het Random Forest-algoritme op basis van meerdere beslisbomen (links) en Google Scholar-hits per jaar voor "Random Forest", "Random Forest" EN "water", en "Random Forest" EN "waterkwaliteit".

Resterende ontwikkelingskans 2
Sensorgegevens combineren met conventionele bemonsteringsgegevens

De meeste sensorstations voor waterkwaliteit worden ook nog steeds conventioneel bemonsterd en geanalyseerd in een laboratorium. Dit is zowel nodig voor de validatie van de sensorgegevens als voor de monitoring van een breder scala aan waterkwaliteitsparameters de niet met sensoren en analysers bepaald kunnen worden (denk aan opkomende stoffen, gewasbeschermingsmiddelen, etc). In de meeste gevallen ligt de kracht van conventionele

momentopname in de nauwkeurigheid van de gemeten waarden, terwijl de kracht van de sensorgegevens ligt in de hoge temporele resolutie van de concentratiemetingen. Een combinatie van deze sterke punten zou optimale tijdreeksen opleveren, maar in de wetenschappelijke waterkwaliteitsliteratuur zijn hiervoor geen procedures of voorbeelden gerapporteerd.

3 Overzicht van frontrunners

Via het literatuuronderzoek en internationale contacten hebben we enkele onderzoeksgroepen geïdentificeerd die reeds uitgebreide gegevensverwerkingsprocedures en -routines voor sensoren openbaar hebben gedocumenteerd. Voor elk van deze groepen was de reden voor hun inspanningen dat het opgeschaalde gebruik van waterkwaliteitssensoren de visuele inspectie en handmatige correcties te omslachtig, tijdrovend en lastig reproduceerbaar maakte. Hun monitoring- en gegevensverwerkingsprocedures werden openbaar gemaakt om de kennis te delen en transparant te zijn voor de stakeholders. Hier wordt informatie gepresenteerd van 4 frontrunners op het gebied van real-time gegevensverwerking voor waterkwaliteit in Utah (VS), Engeland (VK), Duitsland en Australië. Bij de meeste groepen gaat het om correcties achteraf, alleen in Australië is ook gewerkt aan real-time correcties om de online visualisatie te optimaliseren.

Een algemeen aandachtspunt is de gradatie van objectieve (ruwe) meetinformatie tot het via modellen genereren van gegevens waarbij subjectieve keuzes voor methoden en parameters belangrijk worden. Deze subjectieve keuzes beginnen al in de sensor zelf bij het omrekenen van elektrische signalen naar (ruwe) sensor-uitvoer in de vorm van concentratiewaarden. Daarbij worden vaak ook al door de fabrikant ingestelde correcties gedaan voor bijvoorbeeld de temperatuur. Vervolgens beginnen de meeste methodieken met objectieve dataverwerking, zoals het verwijderen van onmogelijke meetwaarden. Het detecteren en eventueel verwijderen van anomalieën is al onderhevig aan subjectieve keuzes en instellingen van de gebruiker. Bij het corrigeren van gegevens en het invullen van gaten in de meetreeks neemt het aantal de vrijheidsgraden verder toe. Het wordt dan belangrijker om de keuzes te documenteren en aan te tonen dat de gegenereerde data plausibel is.

3.1 Utah (VS)

De USGS en Utah State University verzamelen grote hoeveelheden aquatische sensorgegevens binnen het Logan River Observatory. De handmatige kwaliteitscontrole en correctie van de gegevens was bewerkelijk, onderhevig aan subjectiviteit, en moeilijk overdraagbaar en reproduceerbaar. Daarom werd de python-package pyhydroqc (Python hydrological quality control) ontwikkeld en openbaar gemaakt (Jones et al., 2022).

Pyhydroqc bestaat uit een set methoden voor data-gedreven anomaliedetectie en correctie voor hoogfrequente aquatische sensordata. Het instrument was niet specifiek bedoeld voor waterkwaliteitssensorgegevens, maar ook voor bijvoorbeeld waterpeilsensoren. In het Logan River Observatory werden sensorgegevens voor waterniveau, watertemperatuur, pH, opgeloste zuurstof, EC en troebelheid verzameld. De gegevens vertonen anomalieën zoals uitschieters, kunstmatige persistentie (flatline) en drift.

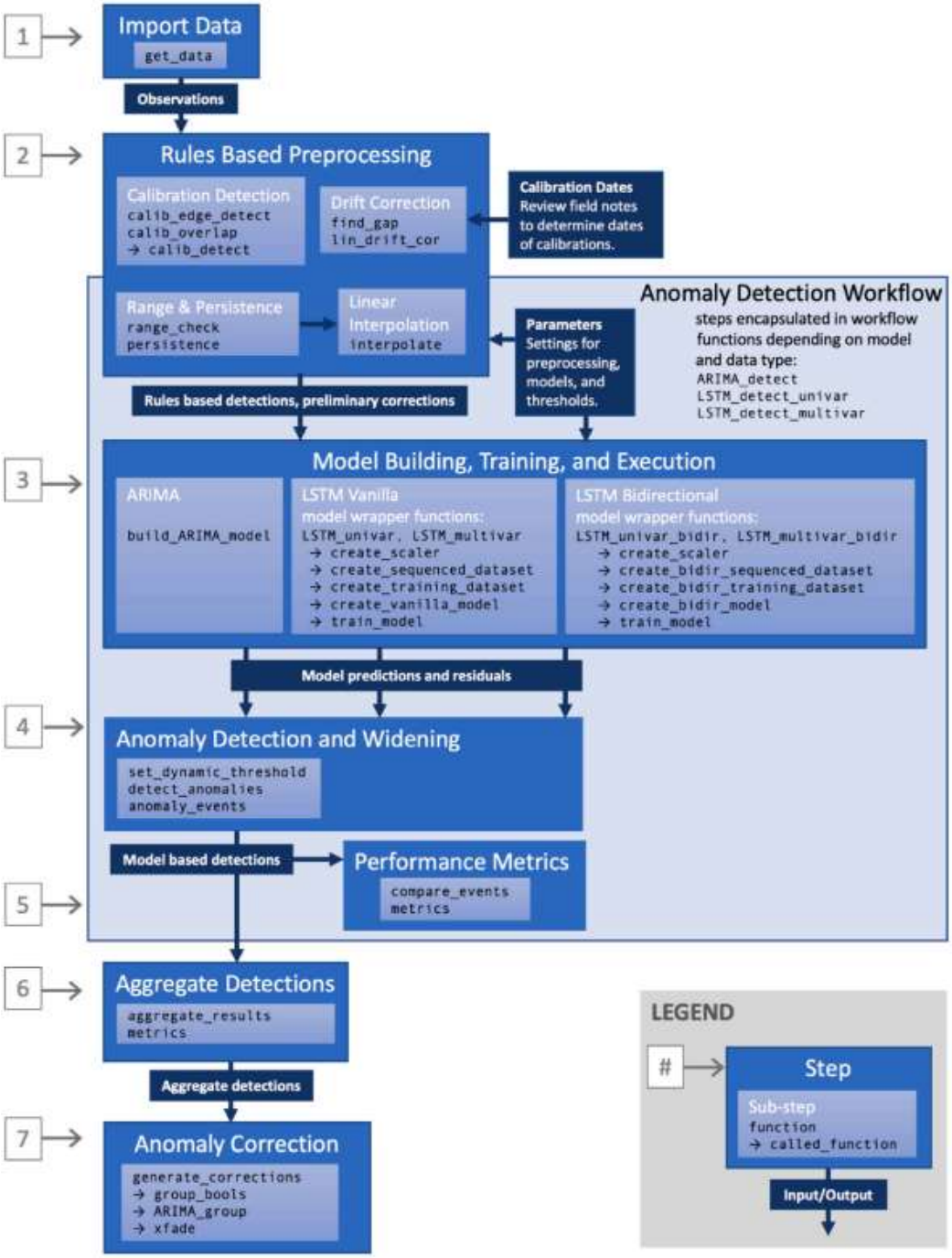
Pyhydroqc bevat zowel klassieke als deep learning tijdreeksregressiemodellen die meetwaarden voorspellen, anomalieën identificeren op basis van dynamische grenzen, en correcties uitvoeren. Het pakket bevat functies en een workflow voor anomaliedetectie en correctie. Geïmplementeerde technieken zijn auto-regressive integrated moving average (ARIMA) en twee soorten long short-term memory (LSTM). LSTM kan worden toegepast met univariate of multivariate invoergegevens. Random Forest is niet inbegrepen.

De workflow binnen Pyhydroqc is als volgt (zie ook figuur 5):

1. Importeer ruwe sensordata volgens een standaard datastructuur.
2. Voer op rules-based anomaliedetectie en correctie uit als eerstelijns kwaliteitscontrole (gegevens buiten meetbereik, persistente meetwaarden), inclusief het corrigeren van plotselinge verschuivingen van de meetwaarde door sensorkalibraties (lineaire driftcorrectie).
3. Bouw een of meer modellen voor het voorspellen van meetwaarden (ARIMA en LSTM):
    a. Bepaal de hyperparameters (instellingen) van het model.
    b. Transformeer en schaal de gegevens indien nodig.
    c. Modellen genereren en fitten.
    d. Het model gebruiken voor voorspellingen en het bepalen van residuen.
4. De modelresultaten nabewerken:
    a. Dynamische grenzen bepalen op basis van modelresiduen en door de gebruiker gedefinieerde parameters.
    b. Anomalieën opsporen waarbij de absolute waarde van het modelresidu de gedefinieerde grens overschrijdt.
    c. Herschaal en indexeer anomalieën.
5. Vergelijk handmatig gelabelde anomalieën met automatisch (met rules-based en model-based technieken) gedetecteerde anomalieën
6. Combineer anomalie-detecties door meerdere methoden voor een geaggregeerde anomaliedetectie.
7. Voer modelgebaseerde correctie uit voor punten die als afwijkend zijn geïdentificeerd (met behulp van ARIMA).

Binnen pyhydroqc kunnen meerdere anomaliedetectiemethoden worden gecombineerd en geaggregeerd. Deze aanpak is niet rechtstreeks gebaseerd op sensor-faalmechanismen, maar de aggregatie van meerdere anomaliedetectiemethoden kan nuttig zijn. Dit omvat ook het gebruik van dynamische grenzen, die gebaseerd zijn op de nauwkeurigheid van het voorspellingsmodel en enkele door de gebruiker gedefinieerde variabelen (minimumdrempel, venstergrootte).

Pyhydroqc kan automatisch draaien, maar bij een eerste toepassing moeten de gegevens (observations en calibration dates) nog in het goede format gezet worden en moeten de parameters goed ingesteld worden. Ook is het raadzaam de (tussen) resultaten goed te controleren, omdat voor nieuwe data mogelijk ook andere parameterinstellingen beter zijn.



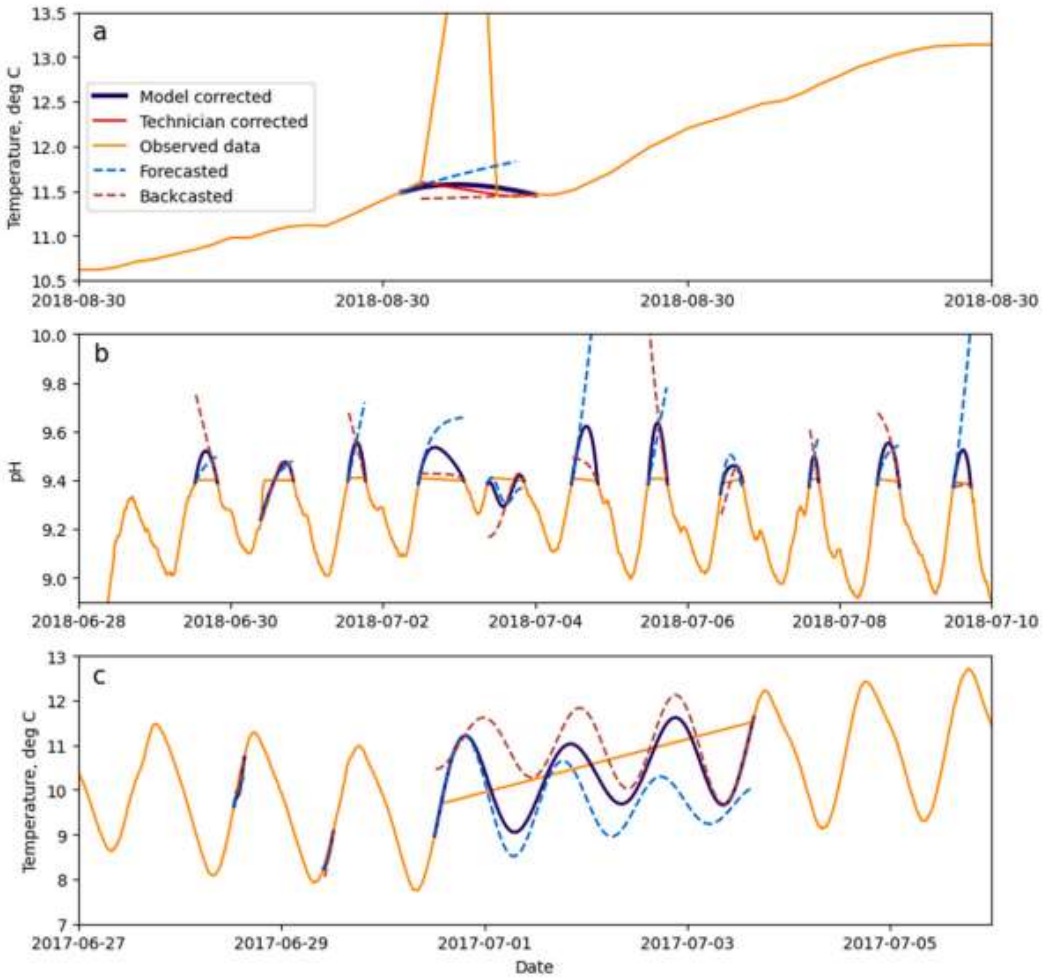
Figuur 5. Workflow van stappen en functies in pyhydroqc. Nummers links corresponderen met stappen in het proces (Uit: Jones et al., 2022).

Voor driftcorrectie kan pyhydroqc automatisch kalibratie-/onderhoudsgebeurtenissen detecteren, maar dit werkt niet altijd goed. Daarom is registratie van onderhoudsdatums nog

steeds nuttig. De tool haalt automatisch gap-waarden op (verschil voor en na onderhoud). Er is geen methode om conventionele waterkwaliteitsmetingen (laboratoriumanalyses en incidentele metingen met handheld sensoren) te betrekken bij de drift- of off-set correctie.

In stappen 1 tot en met 6 in Figuur 5 ligt de focus op de detectie van anomalieën. Hiernaast heeft Pyhydroqc functies om de gaten in de meetreeks op te vullen (stap 7 in Figuur 5). In de procedures voor het opvullen van gaten worden eerst de gaten van korte duur opgevuld, zodat de voorspellingen voor korte gaten helpen bij het opvullen van grotere gaten. Dit kan ook een nuttige aanpak zijn voor andere instrumenten, bijvoorbeeld bij het gebruik van Random Forest voor het opvullen van gaten.

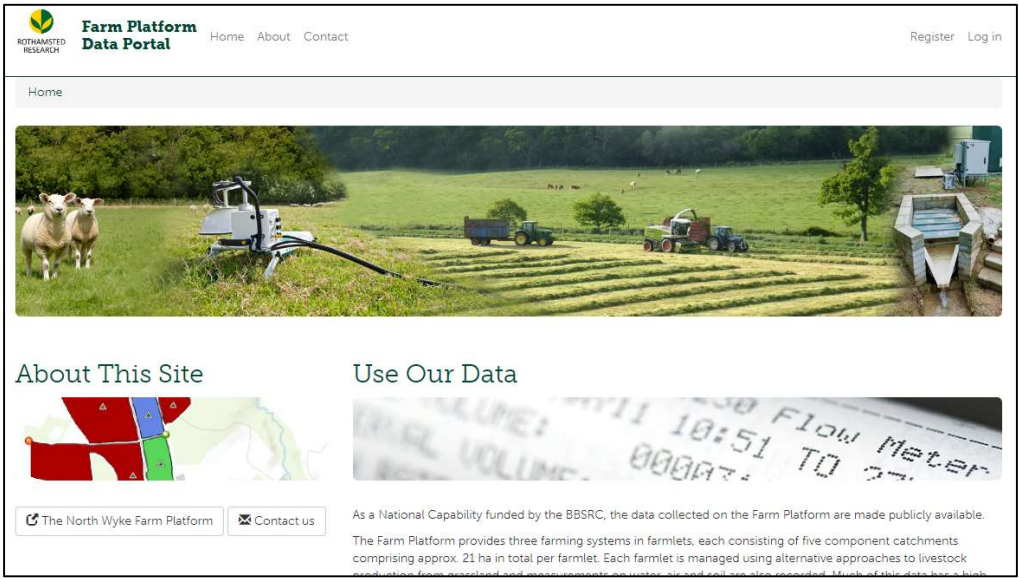
Figuur 6 geeft voorbeelden van de toepassing van het ARIMA-model, in dit geval voor het opvullen van gaten in de meetreeks. Het model maakt een voorspelling en een backcast van de ontbrekende gegevens en die worden verwerkt (via een cross-fade) tot een gecombineerde voorspelling. Deze aanpak is krachtig voor korte gaten in de meetgegevens en, in het geval van zich herhalende schommelingen (zoals dag-nachtritmes), ook voor gaten van meerdere dagen. De ARIMA-voorspellingen kunnen ook worden toegepast voor anomaliedetectie (voorspellingen met één stap vooruit, one step ahead predictions).



Figuur 6. Voorbeelden van succesvolle correctie met behulp van piecewise ARIMA-modellen en de cross-fade techniek. 6a: temperatuur bij Water Lab, 6b: pH bij Main Street, 6c: temperatuur bij Water Lab (Uit: Jones et al., 2022).

3.2 Engeland, GB

In Zuidwest-Engeland werd in 2010 het North Wyke Farm Platform[1] opgericht (figuur 7). Het doel was het effect van veehouderij op het milieu te bestuderen. Naast de waterkwaliteit worden ook de hydrologie (neerslag, afvoer, bodemvocht) en de uitstoot van broeikasgassen gemonitord. Daarnaast zijn gegevens over het landbouwkundig gebruik beschikbaar. Informatie over de monitoringprocedures is gedocumenteerd in Hawkins (2021).



Figuur 7: Screenshot van het North Wyke Farm Platform Data Portal op https://nwfp.rothamsted.ac.uk/, genomen op 27 okt. 2022.

Op 15 locaties op de boerderij pompen waterkwaliteitsstations het water vanuit de sloten door doorstroomcellen. Een YSI EXO2 sonde meet EC, temperatuur, O2, pH, NH4+, NH3 en troebelheid. Nitraat en nitriet werden gemeten met een Nitratax UV-absorptiesensor (Hach). Op vier locaties werd totaal fosfor gemeten met een Phosphax autoanalyser (Hach).

De YSI-sondes worden maandelijks vervangen en in het laboratorium gereinigd en gekalibreerd met standaardvloeistoffen. De Nitratax wordt maandelijks in het veld gereinigd en niet gekalibreerd. De Phosphax heeft een dagelijkse automatische kalibratie.

De kwaliteitscontrole wordt uitgevoerd op datasets van 4 weken in R. De onlinegegevens zijn niet live en hebben een tijdsverschil van ongeveer 2 maanden. De volgende stappen werden opgenomen in de kwaliteitscontrole:

- Dimensiecontrole van de dataset (format)
- Controle van de headers
- Controle van tijd en tijdsinterval
- Controle gegevensformat
- Sensoruitval controle (waarden worden ingesteld als "NA")
- Onmogelijke waarden - via ingestelde boven-/ondergrenzen (waarden worden ingesteld als "NA")
- Checks met alternatieve kwaliteitscontroles
- Samenvattende statistieken per dag

[1] https://www.rothamsted.ac.uk/north-wyke-farm-platform

- Samenvattende statistieken per 4 weken
- Toewijzing TLQF (Traffic Light Quality Flag) (6 niveaus: niet ingesteld, goed, aanvaardbaar, verdacht, zeer verdacht, afgewezen).
- Reden voor toekenning TLQF

De gebruikte data quality flags staan in tabel 1 en de grenzen voor uitschieters staan in tabel 2 en 3. Voor de ammoniumsensor werd ernstige drift gemeld en werden bekende problemen gelabeld, maar in sommige verdachte gevallen was er geen bewijs dat de gegevens fout waren. De nitraatgegevens waren gedurende enkele perioden zeer onregelmatig. De Phosphax-analysatoren leverden betrouwbare gegevens op, maar gaven soms onregelmatige tijdsintervallen.

Tabel 1: Quality flags voor gegevens inclusief volgorde van ernst zoals gebruikt in het North Wyke Farm Platform (Hawkins, 2021).

Description	Severity Order	Details
Not set	0	No information on quality available
Good	2	Data were checked and deemed good
Acceptable	4	Data were checked and no issues were found
Suspicious	25	Data were checked and might have been affected by an event
Highly Suspicious	95	Data were checked and have definitely been affected by an event
Reject	100	Data were rejected
High Sensor Drift	39	Instrument calibration values were high over the time period. As calibration takes place monthly, it is impossible to know if or how much the instrument drifted at the measurement timestamp (as this is not a linear relationship
Missing Sensor Drift	40	Missing instrument calibration information, this level of instrument drift during period is unknown
Outlier	20	The value falls outside 'regular' limits but within the extreme limits, therefore could still be fine
Level Reset	14	Level pressure sensors were reset, indicating this could result in a step in flow
Calibration	15	Calibration Datetime of the instrument
Wiper Issue	16	An issue was detected with the instrument wiper blade. This could have affected the data.

Tabel 2: Extreme grenswaarden voor de detectie van uitschieters zoals gebruikt in het North Wyke Farm Platform (Hawkins, 2021). Waarden onder of boven deze grenzen worden op NA gezet.

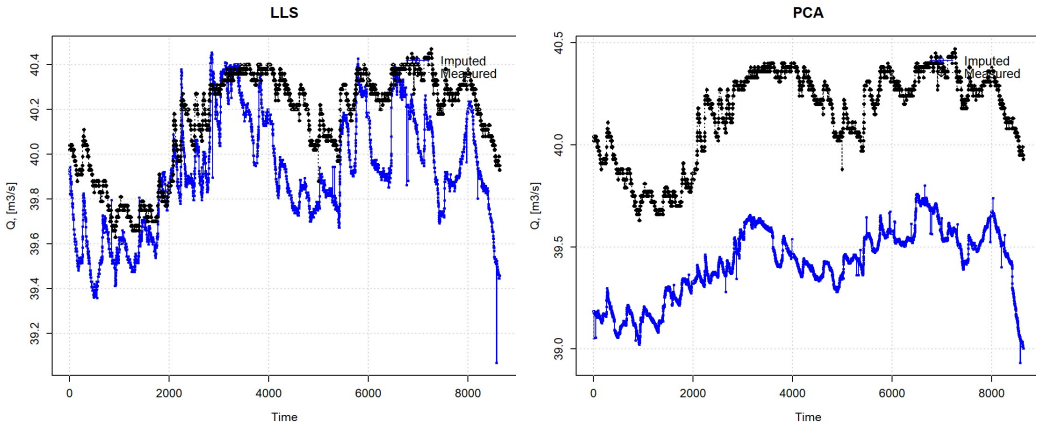
Parameter	Lower threshold	Upper threshold
Nitrate-N (mg/l)	0	48.9
Ammonium-N (mg/l)	0	200
Conductivity (uS/cm)	10	3000
Dissolved Oxygen %	5	500
pH	1	14
Turbidity (FNU)	0	5000
Total P (mg/l)	0	5
TRP (mg/l)	0	2

Tabel 3: Grenswaarden voor de detectie van uitschieters zoals gebruikt in het North Wyke Farm Platform (Hawkins, 2021). Waarden onder of boven deze grenzen worden gemarkeerd als uitschieter.

Parameter	Lower threshold	Upper threshold
Nitrate-N (mg/l)	0	20
Ammonium-N (mg/l)	0	50
Conductivity (uS/cm)	30	1600
Dissolved Oxygen %	60	105
pH	3.5	8.5
Turbidity (FNU)	0	2000
Total P (mg/l)	0	1
TRP (mg/l)	0	0.25

Driftcorrectie van de gegevens is niet opgenomen in de standaardprocedure, omdat de gebruikers daar meestal hun eigen methoden voor hebben (pers. comm. Jane Hawkins, 2022). Wel worden de driftwaarden voor de sensoren die bij ijkproeven zijn verkregen, gerapporteerd. Bovendien is gap- filling niet opgenomen, maar er worden wel aanwijzingen gegeven hoe dat kan.

R-scripts voor LLS (Local Least Square) impute en PCA (Principal Component Analysis) impute zijn beschikbaar gesteld op Rpubs door Curceac et al. (2021). LSS impute selecteert eerst de meest vergelijkbare voorspellende variabelen op basis van Pearson, Spearman en Kendall correlatiecoëfficiënten. Vervolgens worden de ontbrekende waarden voorspeld op basis van regressie met deze vergelijkbare variabelen. PCA impute maakt ook gebruik van relaties tussen variabelen en is gebaseerd op Principal Component Analysis (PCA). Dit is een techniek waarbij meerdere variabelen worden samengevat in minder dimensies (principal components), waarvan de eerste het grootste deel van de variantie dekt. De toepassing van PCA-modellen om ontbrekende waarden te imputeren is geïmplementeerd in het R-pakket missMDA (Josse en Husson, 2016). Voor het gepresenteerde voorbeeld over langetermijnvoorspellingen van bodemvochtgegevens presteerde LSS impute beter dan PCA impute (figuur 8). Voor het opvullen van kortere data-gaten presteerden beide methoden vergelijkbaar.



Figuur 8: Resultaten voor langetermijnvoorspellingen met behulp van LSS impute en PCA impute (uit Curceac et al., 2021).

3.3 Duitsland

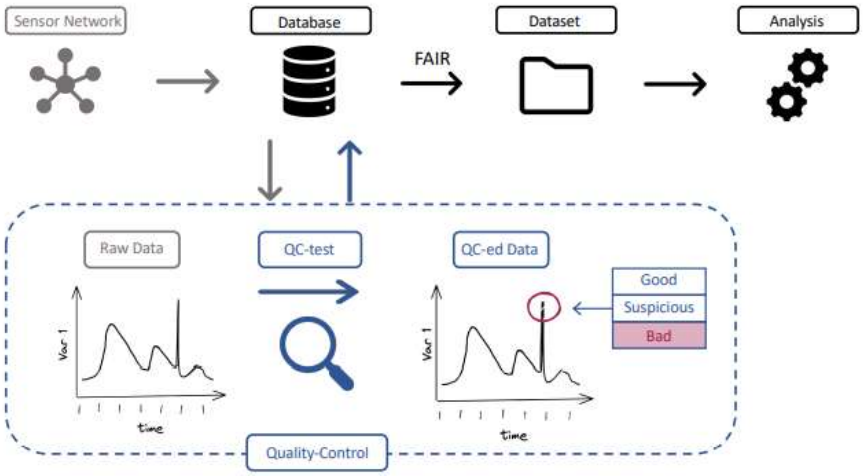
Het Duitse instituut 'Helmholtz Centre for Environmental Research GmbH - UFZ' heeft onlangs een systeem voor geautomatiseerde kwaliteitscontrole van milieugegevens gepubliceerd. De naam van de python tool is SaQC (System for automated Quality Control) en wordt gepresenteerd in een wetenschappelijke paper (Schmidt et al., 2022), een website (rdm-software.pages.ufz.de/saqc) en een github repository.

UFZ werkt met gestaag toenemende hoeveelheden sensorgegevens uit hun grootschalige milieuobservatoria. Het meest relevant voor waterkwaliteitsonderzoek zijn de vier stroomgebieden van TERENO (Terrestrial Environmental Observatories) (Zacharias et al., 2011). SaQC is niet specifiek ontwikkeld voor sensordata over waterkwaliteit, maar voor een veel breder scala aan milieusensoren. Toch kunnen verscheidene methoden en principes van SaQC nuttig zijn voor de verwerking van sensordata over waterkwaliteit.

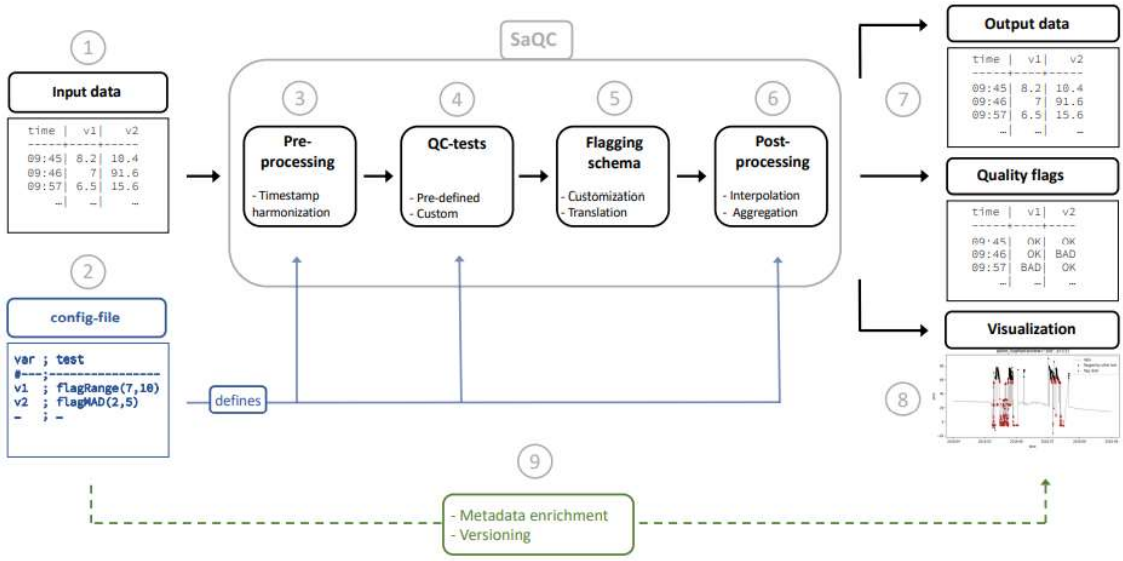
Figuur 9 laat zien waar SaQC zich bevindt in de gegevensstroom van sensor naar analyse. Zowel de ruwe sensorgegevens als de resultaten van de kwaliteitscontrole worden in de database opgeslagen. Figuur 10 geeft een meer gedetailleerd overzicht van de workflow rond de SaQC-tool. De belangrijkste stappen binnen SaQC zijn (1) voorbewerking, (2) kwaliteitscontroles, (3) markering en (4) nabewerking. De postprocessingstap kan ook het genereren van nieuwe, gecorrigeerde tijdreeksen inhouden.

Een overzicht van de functies binnen SaQC wordt gegeven in tabel 4, verdeeld in de hoofdgroepen (1) verwerking, (2) basis-kwaliteitscontrole en (3) geavanceerde kwaliteitscontrole. De meeste functies kunnen met verschillende algoritmen worden uitgevoerd. Tot de algoritmen voor de opsporing van anomalieën behoren bijvoorbeeld het Stray-algoritme, de Modified Z-score (MAD) outlier detection methode en de Grubbs-test. De geavanceerde kwaliteitscontrole omvat functies voor ruis- en driftcorrecties en voor het opvullen van gaten.

Figuur 11 geeft een voorbeeld van ruwe en verbeterde gegevens voor sac254 (spectrale absorptiecoëfficiënt bij 254 nm, een indicatie voor DOC) inclusief driftcorrectie en verwijdering van uitschieters met behulp van SaQC. Eerst werd een eenvoudige meetrange-test uitgevoerd. Na een timestamp harmonisatiestap werd de driftcorrectiefunctie toegepast. Deze functie gebruikt het begin- en eindmoment van het veldonderhoud en corrigeert in dit geval voor exponentieel driftverloop tussen onderhoudsbeurten. Als laatste stap werd een multivariate outlierdetectie gebruikt op basis van k-Nearest Neighbors (kNN) en het STRAY-algoritme (Search and TRace AnomalY). Bij deze multivariate outlier-detectiemethoden zijn ook waterstands- en temperatuurgegevens van dezelfde locatie gebruikt als input.



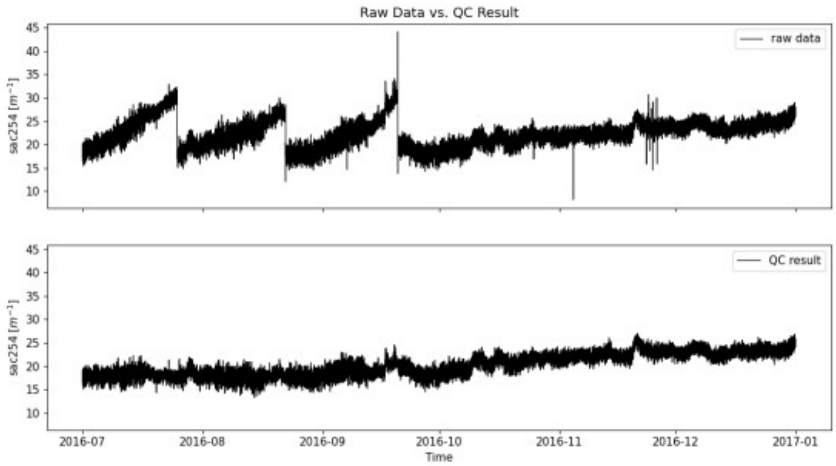
Figuur 9: Illustratie van een algemene gegevensstroom van sensor tot analyse, inclusief geautomatiseerde kwaliteitscontrole: De ruwe gegevens van de sensoren stromen naar de database en vervolgens naar de QC-test. Hier wordt een piek in de gegevens (rode cirkel) geïdentificeerd als foutief en gemarkeerd als "Slecht". De resulterende QC-gegevens, samen met een kwaliteitsvlag voor elk gegevenspunt, worden teruggestuurd naar de database, van waaruit ze als dataset kunnen worden gepubliceerd en later voor analyse kunnen worden gebruikt (Uit Schmidt et al., 2022).



Figuur 10: Algemene QC-workflow met SaQC: Invoergegevens en een config-bestand worden doorgegeven aan SaQC, dat kwaliteitsgecontroleerde uitvoergegevens retourneert, samen met kwaliteitsvlaggen en visualisaties van de QC-resultaten. De kleuren in de figuur onderscheiden vier categorieën: gegevensstromen (zwart), configuratie (blauw), metadatastromen (groen) en annotaties (grijs). De nummers komen overeen met de workflow (uit Schmidt et al., 2022).

Tabel 4: Functies opgenomen in SaQC. Elke categorie bestaat uit functionele groepen die een bepaald aantal (#) functies omvatten die hetzelfde doel dienen. (Uit Schmidt et al., 2022).

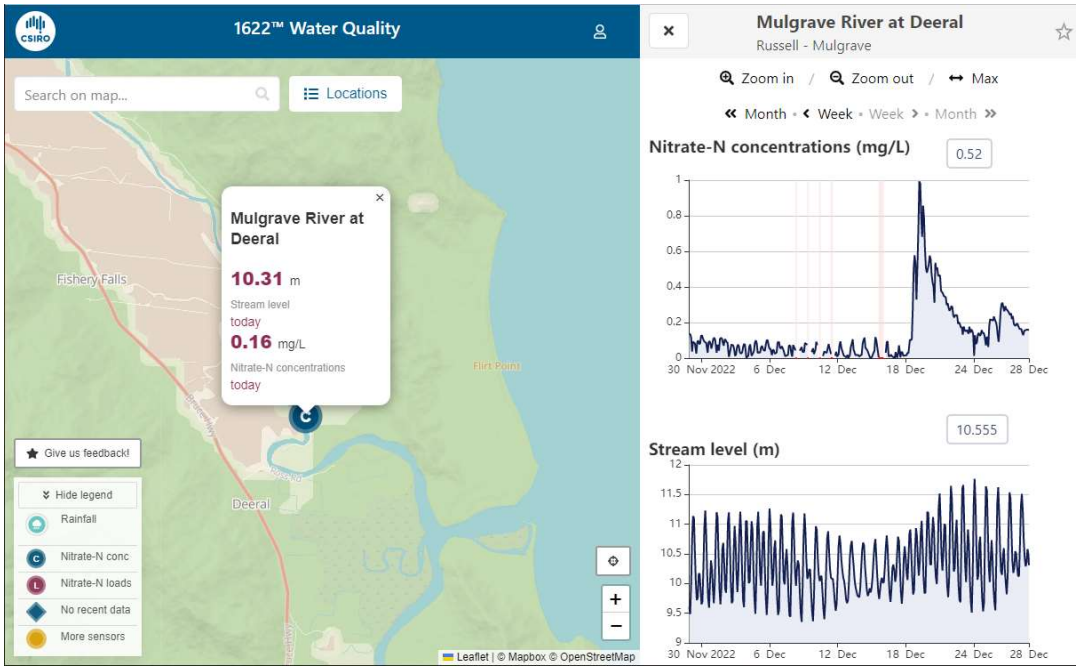
Group	Objective	#
Processing	*Processing steps prior to (pre-processing) or after QC-testing (post-processing)*	
Resampling	Aligning data to equi-distant timestamps by shifting, resampling or interpolation	3
Smoothing	Curve smoothing using parametric and non-parametric methods	2
Transformation	Derive new variables by transformation or using custom functions	2
Flag projection	Project flags from resampled data back onto original data	1
Basic QC	*Low algorithmic complexity and parametrization effort*	
Constants	Deterministic and variance-based detection of undesired stationary behaviour	2
Breaks	Detection of missing/isolated values or jumps in the data	2
Outliers	Detection of outliers and spikes, both deterministic and statistical methods	6
Manual flagging	Integration of precedent manual QC by experts from auxiliary files	1
Advanced QC	*Higher algorithmic complexity and parametrization effort*	
Noise	Separate noise from true signal using low pass filters	1
Changepoints	Detection of points of undesired system state transitions	2
Drift	Detection and correction of sensor drift based on deviation from reference system state	3
Pattern recognition	Detection of undesired patterns in the data based on Dynamic time Warping and Wavelets	2
Custom functions	Combination of existing/integration of custom QC-tests inside config-file or source code	1
Machine learning	Training of machine learning models for flagging and data imputation	4



Figuur 11: Voorbeeld van ruwe (boven) en verbeterde (onder) data met driftcorrectie met behulp van SaQC. (Uit Schmidt et al., 2022).

3.4 North Queensland (Australië)

Langs de noordoostkust van Australië (North-Queensland) wordt een grootschalig waterkwaliteitsmeetnet uitgevoerd om het Great Barrier Reef te kunnen beschermen. Dit Great Barrier Reef Catchment Loads Monitoring Program (GBRCLMP) bestaat uit 11 real-time waterstands- en/of waterkwaliteitsmeetstations. 7 andere stations met nitraatsensoren worden beheerd vanuit verschillende andere projecten in dezelfde regio. De real-time gegevens zijn beschikbaar via de 1622WQ app (https://wq.1622.farm/; zie figuur 12), waarvan de naam refereert naar de hoogte van Mount Bartle Frere, met 1622m de hoogste berg in North Queensland. De 1622WQ-app is ontwikkeld om landbouwers bewust te maken van de waterkwaliteitsproblematiek en de toepassing van Best Management Practices om vermindering van nutriëntenverliezen te bevorderen. De ontwikkeling van de app is beschreven door Vilas et al. (2020).



Figuur 12: Screenshot van de WQ1622 app ingezoomd op een van de waterkwaliteitsstations (datum genomen: 28-12-2022).

Gezien de hoeveelheid data uit de waterkwaliteitsensoren, de afgelegen ligging van de meetlocaties en de doelstellingen van real-time visualisatie en kwantificering van de belasting, hebben verschillende groepen in Noord Queensland gewerkt aan geautomatiseerde verwerkingsroutines voor sensordata. Hieronder geven we een samenvatting van een cloud-based data imputation systeem door CSIRO (Zhang & Thorburn, 2022) en een framework voor het detecteren van outliers door Monash university (Talagala et al., 2019).

3.4.1 Cloud-based data imputation systeem

Zhang & Thorburn (2022) presenteerden een overzicht van imputatietechnieken voor ontbrekende gegevens, toegepast op sensordata van waterkwaliteit. Er werd een cloud-based data imputation systeem ontwikkeld met verschillende imputatiemethoden. De imputatiemethoden omvatten:

- Statistische methoden:
  - gemiddelde imputatie
  - last observation carried forward (LOCF)
  - lineaire imputatie)
- Modelgebaseerde methoden:
  - Expectation Maximization (EM, schatting van de maximale waarschijnlijkheid)
  - Multiple imputations by chained equations (MICE)
  - K-nearest neighbour (KNN).
- Op neuraal netwerk gebaseerde methoden:
  - sequence-to-sequence imputation model (SSIM)
  - Dual SSIM
  - BRITS
  - Multidirectioneel recurrent neuraal netwerk (M-RNN)

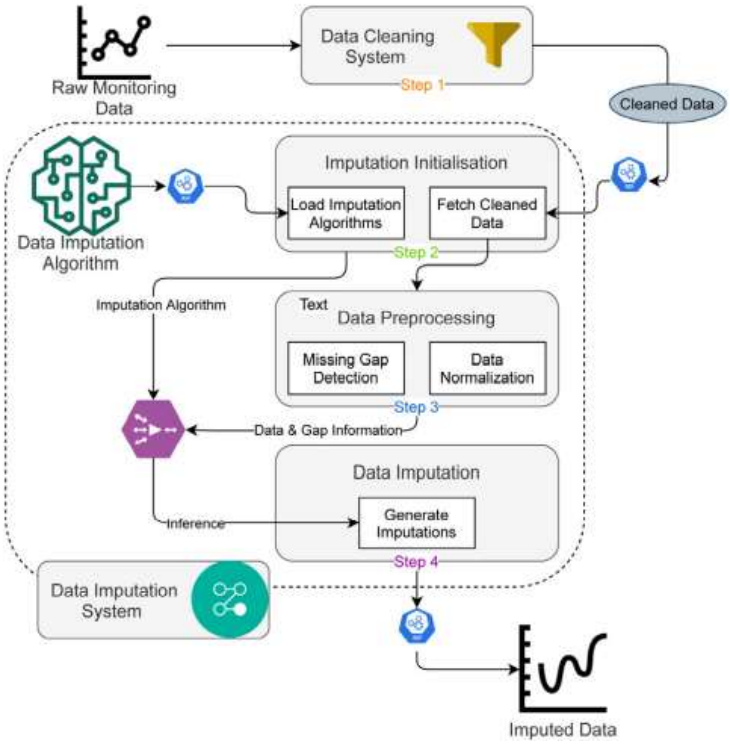
De sterke punten en beperkingen van de verschillende methoden worden in onderstaande tabel opgesomd.

Tabel 5: Sterke punten en beperkingen de groepen van toegepaste methoden voor imputatie van ontbrekende gegevens (uit: Zhang & Thorburn, 2022)

Method	Strength	Limitation
Mean Imputation		The variability in the data is reduced;the standard deviations and the variance estimates may get underestimated [35],
LOCF	Easy to understand, Efficient to apply	The assumption is mostly unrealistic [36],
Linear Imputation		There should be a linear relationship between the predictor and response variables
EM	Good Interpretability,	High risk to overfit the training data
MICE	Lazy Learning, do not build	Computationally expensive for large datasets
KNN	models from training data	
SSIM	Can capture and use the	Black box method,
Dual-SSIM	temporal information,	Performance heavily relies on
BRITS	Deep architecture brings	hyperparameters tuning,
M-RNN	strong representation learning	High computational cost to build the model

De stappen van het data imputatie systeem zijn weergegeven Figuur 13:

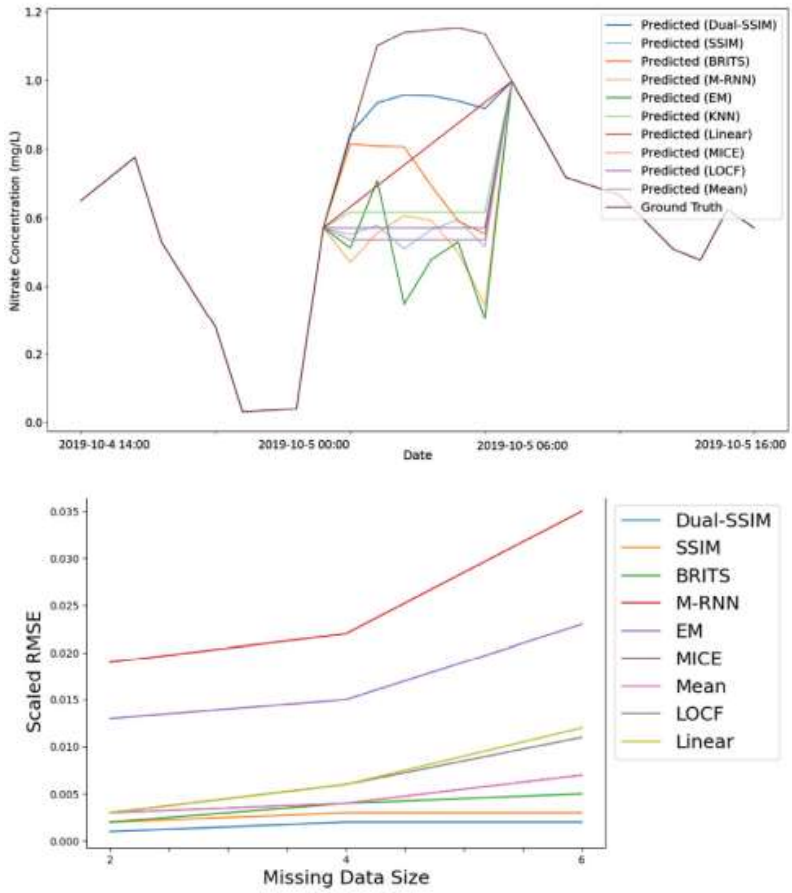
- Opschonen waterkwaliteitsmonitoringgegevens (verwijderen ongeldige, buiten bereik liggende waarden)
- Initialisatie imputatiesysteem (selecteren en laden van het imputatiealgoritme)
- Voorbewerking van gegevens (normalisatie, herschaling, identificatie van datagaten)
- Genereren van de imputatie data



Figuur 13: Overzicht van en stappen binnen het dataimputatiesysteem (uit: Zhang & Thorburne, 2022)

De prestaties van de verschillende imputatiemethoden voor het invullen van ontbrekende gegevens worden vergeleken in figuur 14. In de grafieken worden de prestaties van de methoden voor het reproduceren van een nitraatconcentratiepiek vergeleken (links), en wordt het effect van de grootte van het datagat op de prestaties getoond (rechts). Dual-SSIM presteerde het best bij de reproductie van de voorbeeldpiek voor de nitraatconcentratie (bovenste grafiek figuur 14). Zowel Dual-SSIM als SSIM presteren beter dan andere methoden, ook bij grotere datagaten (onderste grafiek figuur 14). Merk op dat een gat van 6 datapunten (zoals in figuur 14) nog steeds relatief kort is, gezien de gebruikelijke reactietijden van onderhoudspersoneel om sensoren ter plaatse te repareren (vaak > 1 dag). Dual SSIM en SSIM zijn beide ontwikkeld door CSIRO en worden beschreven in Zhang & Thorburn (2021) en in Zhang et al. (2019).

SSIM (sequence-to-sequence imputation model) is een nieuwe methode voor het invullen van ontbrekende gegevens in hoogfrequente tijdreeksen van waterkwaliteit. De methode is ontworpen voor situaties waarin alle gemeten parameters een gelijktijdig datagat hebben (en de correlaties tussen de parameters niet kunnen worden gebruikt). SSIM maakt gebruik van sequence-to-sequence deep learning architectuur en de techniek Long Short-Term Memory (LSTM) Network is gekozen om zowel de informatie uit het verleden als die uit de toekomst te benutten voor variërende tijdsperioden. Het Variable-Length Sliding Window Algorithm (VLSW) selecteert verschillende stukken meetreeks van variabele lengte en vergroot daarmee de hoeveelheid trainingssamples wat de resultaten van het deep learning-algoritme verbeterd (Zhang et al., 2019). Een uitbreiding van SSIM is de Dual SSIM-methode (Zhang & Thorburn, 2021), waarbij de gegevens aan beide zijden van het gegevensgat afzonderlijk worden verwerkt.



Figuur 14: Prestaties van verschillende imputatiemethoden voor de imputatie van een ontbrekende piek in NO3-concentraties (boven) en prestaties van de verschillende methoden als functie van het aantal ontbrekende datapunten (onder) (uit: Zhang & Thorburne, 2022).

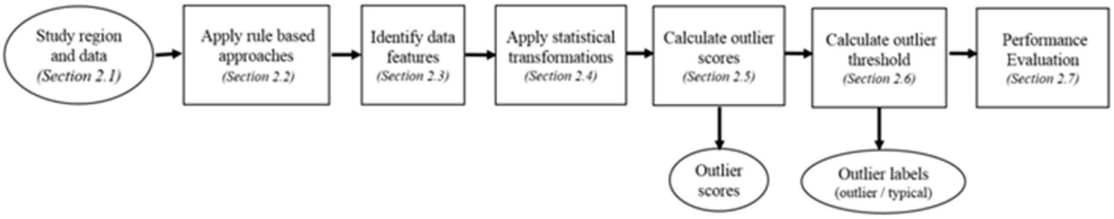
3.4.2 Framework voor outlier detectie

Het project "Revolutionising water quality monitoring in the information age" onder leiding van het Centre for Data Science van de Queensland University of Technology heeft relevante publicaties en R-pakketten opgeleverd. Veel van deze producten zijn beschikbaar gesteld op de projectwebsite[2] . Een paper van Talagala et al. (2019) beschrijft een framwork voor het detecteren van uitschieters dat is geïmplementeerd in het R-pakket *oddwater*. De voorbeelden richten zich op turbiditeit, geleidbaarheid en rivierpeilgegevens van het sensornetwerk in rivieren die uitmonden in het Great Barrier Reef (Australië). De nadruk ligt op uitschieters met abrupte waardeveranderingen, waaronder plotselinge pieken, plotselinge geïsoleerde dalingen en niveauverschuivingen in de meetreeks. De methoden kunnen worden toegepast in de meerdimensionale ruimte, zodat de correlatie tussen parameters kan worden meegenomen in de opsporing van uitschieters.

De opgenomen technieken zijn unsupervised. Dit betekent dat zij geen trainingsgegevens nodig hebben met vooraf bekende uitbijters. In de praktijk zijn niet alle potentiële uitschieters vooraf bekend, en zijn trainingsgegevens met bekende uitbijters vaak niet beschikbaar, wat de toepassing van (semi-)supervised methoden bemoeilijkt. De workflow van het voorgestelde framework is weergegeven in figuur 15 en omvat de volgende stappen:

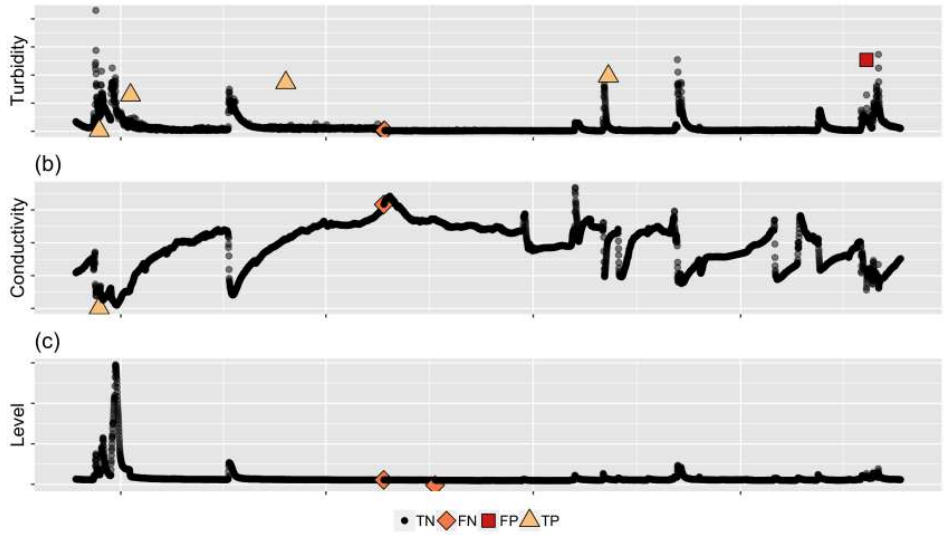
[2] https://research.qut.edu.au/qutcds/projects/revolutionising-water-quality-monitoring/

- Met rule based benaderingen worden uitschieters, onmogelijke en ontbrekende waarden opgespoord.
- De soorten uitschieters worden geïdentificeerd (gegevenskenmerken).
- Statistische transformaties helpen vaak om de verschillende soorten uitschieters naar voren te halen. Geïmplementeerde transformaties zijn log transformation, first difference, time gap, first derivative, one sided derivative, rate of change, en relative difference.
- De scores voor uitschieters worden berekend met behulp van acht technieken voor het unsupervised scoren van uitschieters (KNN-SUM, KNN-AGG, HDoutliers, LOF, COF, INFLO, LDOF, RKOF).
- De extreme value theory (EVT) is gebruikt om de grenswaarden voor de uitschieterscore te berekenen en uitschieters te onderscheiden van normale datapunten.



Figuur 15: Workflow van het framework voor het detecteren van uitschieters (uit: Talagala et al., 2019).

De voorbeeldevaluatie in Figuur 16 laat zien dat er nog steeds uitschieters kunnen zijn die niet worden gedetecteerd (False Negatives) of normale datapunten die verkeerd worden geclassificeerd als uitschieters (False Positives). In vervolgonderzoek werd ook de informatie van nabijgelegen sensoren gebruikt bij de detectie van uitschieters. Deze methode is opgenomen in het R-pakket *conduits*.



Figuur 16: Voorbeeld van de evaluatie van de methode voor de detectie van uitschieters (met KNN-SUM als scoringstechniek voor uitschieters). TN=True Negative, TP=True Positive, FN=False Negative, TP=False Positive. (uit: Talagala et al., 2019)

4 Literatuur

Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.
http://dx.doi.org/10.1023/A:1010933404324

Curceac, S. Hawkins, J., Harris, P., 2021.Advanced Quality Control Report 1: Missing value imputation of the 15-minute soil moisture data.
https://rpubs.com/North_Wyke_Farm_Platform/765406

Hawkins, 2021. User guide to fine resolution (15 minute) data. Version 1.10.
http://resources.rothamsted.ac.uk/sites/default/files/groups/North_Wyke_Farm_Platform/FP_UG.Doc_.002_15MinData_ver1.10.pdf

Josse J. and F. Husson. 2016. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. Journal of Statistical Software, 70(1), 1-31.

Schmidt, L., Schaefer, D., Geller, J., Lünenschloss, P., Palm, B. , Rinke, K., and Bumberger, J. System for Automated Quality Control (Saqc) to Enable Traceable and Reproducible Data Streams in Environmental Science. SSRN Electronic Journal.
http://dx.doi.org/10.2139/ssrn.4173698.

Spackman Jones, A., T.L. Jones, J. S. Horsburgh, 2022. Toward automating post processing of aquatic sensor data. Environmental Modelling and Software 151.

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

Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., Frenzel, M., Schwank, M., Baessler, C., Butterbach-Bahl, K., et al., 2011. A network of terrestrial environmental observatories in germany. Vadose zone journal 10, 955–973. doi:10.2136/vzj2010.0139.

Zhang, Y. F., Thorburn, P. J., Xiang, W., & Fitch, P. (2019). SSIM—A deep learning approach for recovering missing time series sensor data. IEEE Internet of Things Journal, 6(4), 6618-6628.

Zhang, Y. & P.J. Thorburn, 2021. A dual-head attention model for time series data imputation, Comput. Electron. Agric. 189, http://dx.doi.org/10.1016/j.compag.2021.106377.

Zhang, Y. & P. J. Thorburn, 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Generation Computer Systems 128.

Vilas, M. P., P. J. Thorburn, S. Fielke, T. Webster, M. Mooij, J. S. Biggs, Y.F. Zhang, A. Adham, A. Davis, B. Dungan, R. Butler, P. Fitch, 2020. 1622WQ: A web-based application to increase farmer awareness of the impact of agriculture on water quality. Environmental Modelling & Software 132. https://doi.org/10.1016/j.envsoft.2020.104816

5 Appendix A: Samenvattingen artikelen

Reference Arriagada, P., B. Karelovic, O. Link, 2021. Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. Journal of Hydrology 598.https://doi.org/10.1016/j.jhydrol.2021.126454
Field of Science Quantitative hydrology
Summary techniques and results The MissForest algorithm is introduced for automatic gap-filling of daily streamflow time series. Stekhoven and Bühlmann (2012) extended the RF to the MissForest algorithm (MF) for missing value imputation in mixed-type data. Implemented in R package MissForest. The algorithm is applied in 122 gauges in a data-scarce region with multiple climates. Results show that MissForest is a precise and reliable tool for simulation of the missing data in an automatic way. MF failed in cases of severe alterations of the flow regime.
Link to Optima HWQ • MissForest (R package) may also be useful for gap-filling water quality sensor data • Performance also checked through artificial data gaps (both single data points and longer periods)

Reference Stekhoven, D.J., Bühlmann, P., 2012. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597
Field of Science Mathematics
Summary techniques and results A new non-parametric missing value imputation method was developed which can cope with different types of variables (continuous and categorial) simultaneously. No unkown tuning parameter is used and the method is flexible in the type of data. This MissForest (MF) algorithm was tested against four other data imputation methods: **KNNimpute,** K Nearest Neighbours impute (Troyanskaya et al., 2001): A missing value variable X_i is imputed by finding its *k* nearest observed variables and taking a weighted mean of these *k* variables for imputation. Thereby, the weights depend on the distance of the variable X_j. The tuning parameter *k* has a large effect, but is not known beforehand. Method from gene expression analysis. **MissPALasso**, Missingness Pattern Alternating Imputation and l_1-penalty algorithm (Städler and Bühlmann, 2010): Missing variables are regressed on the observed ones using the lasso penalty by (Tibshirani, 1996). In the following E step, the obtained regression coefficients are used to partially update the latent distribution. The tuning parameter λ has a large effect, but is not known beforehand.

MICE, Multivariate Imputation by Chained Equations (Van Buuren and Oudshoorn, 1999): In the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types.

Link to Optima HWQ

- MissForest (R package) may also be useful for gap-filling water quality sensor data
- Other methods (KNNimpute, MissPALasso, MICE) may also be interesting

Reference

Spackman Jones, A., T.L. Jones, J. S. Horsburgh, 2022. Toward automating post processing of aquatic sensor data. Environmental Modelling and Software 151. https://doi.org/10.1016/j.envsoft.2022.105364

Field of Science

Environmental Sciences, water quality

Summary techniques and results

This study applied both classical and deep learning time series regression models that estimate values, identify anomalies based on dynamic thresholds, and offer correction estimates. The Python package pyhydroqc includes custom functions and a workflow for anomaly detection and correction. Implemented techniques are auto-regressive integrated moving average (ARIMA) and two types of long short-term memory (LSTM). The package is applied to a large-scale dataset (Logan River, northern Utah, USA); several result examples and experiences also from manual data processing were presented.

Workflow:

1. Import raw sensor data into a memory-resident data structure.
2. Perform rules-based anomaly detection and correction as a first pass at quality control (out-of-range data, persistent data), including addressing sensor calibration shifts (linear drift correction).
3. Build one or more models for predicting observed values (ARIMA and LSTM):
   a. Determine model hyperparameters.
   b. Transform and scale data if necessary.
   c. Build and fit models.
   d. Execute the model to determine model predictions and residuals.
4. Post-process model results:
   a. Determine dynamic thresholds based on model residuals and user-defined parameters.
   b. Detect anomalies where the absolute value of the model residual exceeds the defined threshold.
   c. Widen and index anomalous events.
5. Compare technician labeled and detected anomalous events (rules-based and model-based detections, inclusive) to assign confusion matrix categories and report metrics.
6. Combine detections identified by multiple models for an aggregate anomaly detection
7. Perform model-based correction for points identified as anomalous (using ARIMA).

Link to Optima HWQ

- pyhydroqc (Python package) may be useful for drift correction, anomaly detection and correction, maybe also for gap filling (using ARIMA).
- multiple anomaly detection methods combined and aggregated (although not directly based on sensor failure mechanisms)
- During gap-filling, the short duration gaps are filled in first, so that the predictions help with the filling of larger gaps. This may be a useful approach also for other tools. In addition, sensor cleaning without calibration can cause data shifts as well, this is not accounted for.
- Automatically detecting calibration events does not always work very well; including field notes and calibration dates is useful here. Gap values (between before and after calibration) can be automatically retrieved from the data by the tool. However, sample data are not applied for correcting, which would be useful (both for drift and off-set).
- Rather than using constant thresholds for anomaly detection, dynamic thresholds allowed for responsiveness to data variability. The dynamic thresholds are based on the prediction model accuracy and some user-defined variables (minimum threshold, window size).

Reference
Tyrali, H. and G. Papacharalampous, 2017. Variable Selection in Time Series Forecasting Using Random Forests. Algorithms 2017, 10, 114.
https://doi.org/10.3390/a10040114

Field of Science
Environmental Sciences, global temperature data

Summary techniques and results
This study focused on assessing the performance of random forests in one-step forecasting using two large datasets of short yearly average temperature time series with the aim to suggest an optimal set of predictor variables. The performance was compared to benchmarking methods. The primary aim was to investigate how the performance of RF is related to the variable selection in one-step (101st value) forecasting of short time series. RF was proven to be a competent one-step forecasting algorithm. The RF methods performed better when using fewer predictor variables (because more predictor variables reduce the length of the training dataset and reduce the exploitation of information from the original time series).

Techniques used for one-step forecasts:

ARMA, autoregressive moving average; provides a parsimonious description of a (weakly) stationary stochastic process in terms of two polynomials: The AR part involves regressing the variable on its own lagged (i.e., past) values. The MA part involves modeling the error term as a linear combination of error terms.
ARFIMA, Autoregressive Fractionally Integrated Moving Average; generalization of the ARMA method, better suited for forecasting of systems with long memory.
Theta method, predicts based on two or more auxiliary time series with different (modified) local curvatures with respect to the original, namely the “Theta-lines”.
Random Forest
Naïve methods (for benchmarking): the prediction is equal to the last value and to the average value

Link to Optima HWQ

- One-step forecasting may be interesting for (real time) anomaly detection
- Differences between performance of RF functions were small, RF models with fewer prediction variables performed better. RF predictions sometimes better, sometimes worse than other methods.
- Benchmark comparisons: the last value is also the predicted value, and the average value is the predicted value.
- R-function arima.sim was used for ARMA (forecast through 'predict' built in R function) and R package fracdiff for ARFIMA (forecast trough forecast function of forecast R package).
- Theta method through R package forecast, thetaf function
- Random Forest modeling through randomForest R package, training with train function of caret R package, forecasting using predict function of caret R package
- Codes available via Supplementary material for the paper "Variable selection in time series forecasting using random forests" - Mendeley Data

Reference

Kang M, Ichii K, Kim J, Indrawati YM, Park J, Moon M, Lim J-H, Chun J-H, 2019. New Gap-Filling Strategies for Long-Period Flux Data Gaps Using a Data-Driven Approach. *Atmosphere* 10(10):568. https://doi.org/10.3390/atmos10100568

Field of Science

Environmental Sciences, atmospheric research, evaporation/CO2 emissions

Summary techniques and results

The study used a data-driven machine learning method for filling long (>30 days) gaps in eddy covariance carbon/water/energy flux measurements. Many reported methods are not useful for such long gaps.

SVR (support vector regression) was used for gap-filling: a machine learning technique that transforms nonlinear regressions into linear regressions, including a data classification process that maps the original low-dimensional input space to a higher-dimensional feature space for classifying the data linearly.
Comparison to random forest (RF) and artificial neural network (ANN), which produced similar accuracies.

The study concludes that long term gap-filling can best be performed using in-situ prediction variables, long training periods, but separated models in case of system state changes.

Link to Optima HWQ

- The study found that in-situ measurements are preferentially used as input for the machine learning (rather than remote sensing or modeled data in this case), also in water quality: prefer in-situ data over regional data / modeled data?
- Long training periods give better gap-filling results; in case of ecosystem state changes separate models should be used for the different state periods

Reference
Caquilpán P V, Aros G G, Elgueta A S, Díaz S R, Sepúlveda K G, Sierralta J C., 2019. Advantages and challenges of the implementation of a low-cost particulate matter monitoring system as a decision-making tool. Environ Monit Assess. 191(11):667. https://doi.org/10.1007/s10661-019-7875-4

Field of Science
Environmental Sciences, atmospheric research, air quality (particulate matter)

Summary techniques and results
As part of a low-cost air quality monitoring system, machine learning was applied in the calibration process. First outliers and data collected at humidity values >95% were removed. Then, multiple linear regression (LR) and random forest (RF) algorithms were applied to improve the monitoring stations data (with data loss rates between 25 and 70%), using temperature and relative humidity as prediction variables. RF lead to better results than LR.

Link to Optima HWQ
- Study showed that random forest performed good at predicting air quality data, better than multiple linear regression
- In this case, the original data was totally replaced by the RF modeled data. This may be applicable for water quality, also in combination with sampling data.

Reference
María Castrillo, Álvaro López García, 2020. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods, Water Research 172. https://doi.org/10.1016/j.watres.2020.115490

Field of Science
Environmental Sciences, water quality

Summary techniques and results
Open source hourly water quality data from two tributaries of the Thames (Enborne and Cut) were used to predict nutrient concentrations (TRP, TP, NO3, NH4) based on seven physico-chemical predictors (EC, Turb, Temp, DO, pH, Chl, flow) using machine learning, especially random forest. Multiple Linear Regression and Random Forest was used. Random Forest performed better, except with only one predictive variable.

Link to Optima HWQ
- Interesting way for stepwise selection of predicting variables: For the Multiple Linear Regression: iteratively adding the variable adding most to the R^2. For Random Forest, start with all available variables; the variables with the lowest prediction scores (in this case RMSE) are dropped stepwise. Also, the highest number of trees was applied that still reduced the error with more than 5%. Reducing the number of variables and trees improves the stability of the model (reduces overfitting) and reduces the computational demand.
- Example of prediction of nutrient concentrations (TRP, TP, NO3, NH4) from EC, Discharge, Temp, Turbidity, pH, Chl.

Reference
Shen, L.Q., Amatulli, G., Sethi, T., P. Raymond, S. Domisch, 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci Data 7, 161. https://doi.org/10.1038/s41597-020-0478-7

Field of Science
Environmental Sciences, water quality

Summary techniques and results
In this study, models were built using Random Forest (RF) that regressed the seasonally measured N and P concentrations (TN, TP, TDN, TDP, and NO3) collected at 62,495 stations across the US streams for the period of 1994–2018 onto a set of 47 in-house built environmental variables that are available at a near-global extent. The predictive powers measured by Pearson Coefficients reached approximately 0.66 on average.

Data cleaning by removing extreme values, Box–Cox power transformation to get normal distributions, and selection of locations with minimal 3 measurements per month and with a Coefficient of Variation lower than two. 47 predictors were used belonging to four categories: topography, soil, land cover and climate. The RF regression algorithm implemented in the R-package randomForestSRC was used. The scripting procedure is available at the GitLab repository (https://gitlab.com/Ferdinand18/np_us_streams).

Link to Optima HWQ
- Example of RF application to water quality data, in this case using catchment characteristics as predictors.
- R package randomForestSRC may be useful

Reference
Dastorani, M.T., A. Moghadamnia, J. Piri, M. Rico-Ramirez, 2010. Application of ANN and ANFIS models for reconstructing missing flow data. Environ Monit Assess 166, 421–434. https://doi.org/10.1007/s10661-009-1012-8

Field of Science
Environmental Sciences, water quantity (discharge data)

Summary techniques and results
This study tried to predict the missing data of discharge gauging stations in Iran using data from neighboring sites and a relevant architecture of artificial neural networks (**ANN**) as well as adaptive neuro-fuzzy inference system (**ANFIS**). To be able to evaluate the results produced by these new techniques, two traditionally used methods including the **normal ratio method** (between two stations) and the **correlation method** were also employed. All methods had acceptable results, ANFIS was best and ANN was also efficient. ANFIS aims at systematically generating unknown fuzzy rules connecting an input and output dataset.

Link to Optima HWQ
- Study is relatively old, ANFIS may be useful.

Reference
Ha, N.T., Nguyen, H.Q., Truong N.C.Q., Le T.L., Thai, V.N., Pham, T.L., 2020. Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam. Environ Monit Assess. 192(12):789. https://doi.org/10.1007/s10661-020-08731-2

Field of Science
Environmental Sciences, water quality

Summary techniques and results
This study aimed at testing random forest (RF) to predict nutrient concentrations for the Tri An Reservoir (TAR). Nitrite, nitrate, and phosphate were empirically estimated using the field observation dataset (2009-2014) of six surrogates of total suspended solids (TSS), total dissolved solids (TDS), turbidity, electrical conductivity (EC), chemical oxygen demand (COD), and biochemical oxygen demand (BOD5). The RF regression model was reliable for N-NO2-, N-NO3-, and P-PO43- prediction with a high R2 of 0.812-0.844 for the training phase (2009-2012) and 0.888-0.903 for the validation phase (2013-2014). EC, TSS and turbidity were the most powerful predictors.
RF results were compared with Multiple Linear Regression (R^2 and RMSE). The Pyton scikit-learn library was used.

Link to Optima HWQ
- Example of RF application to water quality data, in this case using physio-chemical parameters as predictors.
- Pyton scikit-learn library may be useful

Reference
Rozema P.D., Kulk G., Veldhuis M.P., Buma A. G. J., Meredith M.P., Van de Poll W.H., 2017. Assessing Drivers of Coastal Primary Production in Northern Marguerite Bay, Antarctica. Frontiers in Marine Science 4. https://doi.org/10.3389/fmars.2017.00184

Field of Science
Environmental Sciences, marine water quality / ecology

Summary techniques and results
A non-assumptive random forest model (RF) was used to predict primary production (PP) around the Western Antarctic Peninsula. Variability in PP was best explained by light availability and chlorophyll a followed by physical (temperature, mixed layer depth, and salinity) and chemical (phosphate, total nitrogen, and silicate) water column properties. A reduced model showed how light availability, chlorophyll a, and total nitrogen concentrations can be used to obtain the best estimate of PP (R2 = 0.93).
The log of primary production was modeled; 2 models: 1 maximum (2000 trees) and 1 reduced, optimized. Variables least important in MRFmax were removed individually until the fit of the new model no longer improved and this became MRFmin. R package randomForest was used.

Link to Optima HWQ
- Interesting way from complicated RF model (with many predictors) towards less complicated (stepwise deletion of least important predictor).
- Example of efficient RF application for ecological water quality: primary production

Reference
Parkhurst DF, Brenner KP, Dufour AP, Wymer LJ., 2005. Indicator bacteria at five swimming beaches-analysis using random forests. Water Res. 39 (7) https://doi.org/10.1016/j.watres.2005.01.001

Field of Science
Environmental Sciences, marine water quality / health

Summary techniques and results
Random Forest was used to study relationships of indicator bacteria density at five beaches in the USA to numerous other variables. The best predictors were day of the week, indicator density 24h earlier, water depth at the sampling point, and cloud cover. Using data from the first 52 days of measurement allowed predicting indicator densities in the following 10 days to order of magnitude at some of the beaches. The predictions worked poorly for raw densities, but performed well for part of the beaches in predicting orders of magnitude (logarithms) of the densities.

Link to Optima HWQ
- Relatively early RF application in water quality (without very good results)

Reference
Qiao Z, Sun S, Jiang Q, Xiao L, Wang Y, Yan H, 2021. Retrieval of Total Phosphorus Concentration in the Surface Water of Miyun Reservoir Based on Remote Sensing Data and Machine Learning Algorithms. Remote Sensing 13(22):4662. https://doi.org/10.3390/rs13224662

Field of Science
Environmental Sciences, water quality

Summary techniques and results
This study models total phosphorus (TP) concentration in the Chinese Miyun Reservoir by comparing 12 machine learning algorithms, including support vector machine (SVM), artificial neural network (ANN), Bayesian ridge regression (BRR), lasso regression (Lasso), elastic net (EN), linear regression (LR), decision tree regressor (DTR), K neighbor regressor (KNR), random forest regressor (RFR), extra trees regressor (ETR), AdaBoost regressor (ABR) and gradient boosting regressor (GBR). Predictors were the 9 spectral bands of Landsat. No information was reported about what spectral bands were the best predictors.

The model was built in the Python environment: arcpy library function (format conversion and mask extraction on the remote sensing data); numpy library (convertion into a one-dimensional array; sklearn library (machine learning); pickle library (retrieval results).

Comparison table of results:

Algorithm	Mean Absolute Error (mg/L)	Mean Square Error (mg/L)	Explained Variance Score	R^2
Linear Regression	0.001747	0.000007	0.598713	0.598713
Bayesian Ridge Regression	0.001608	0.000008	0.579374	0.579374
Lasso Regression	0.001723	0.000007	0.596967	0.596967
K Neighbor Regressor	0.001735	0.000007	0.598132	0.598132
Elastic Net	0.001447	0.000005	0.724383	0.724263
Decision Tree Regressor	0.000421	0.000003	0.850468	0.897365
Support Vector Machine	0.001953	0.00001	0.44061	0.432786
Artificial Neural Network	0.003344	0.000022	0	0
AdaBoost Regressor	0.001415	0.000005	0.739572	0.738588
Random Forest Regressor	0.000935	0.000003	0.814934	0.814851
Extra Trees Regressor	0.000433	0.000003	0.850468	0.850468
Gradient Boosting Regressor	0.000636	0.000003	0.844646	0.844646

Link to Optima HWQ

- Beside RF, some other algorithms perform very well: Extra Trees Regressor, Decision Tree Regressor, Gradient Boosting Regressor.
- Useful Python libraries: sklearn
- Application of remote sensing data may be possible for larger water surfaces, Sentinel partly has 10 m2 resolution.

Reference

Olson, J. R., and C. P. Hawkins (2012), Predicting natural base-flow stream water chemistry in the western United States, Water Resour. Res. 48. https://doi.org/10.1029/2011WR011088

Field of Science

Environmental Sciences, water quality

Summary techniques and results

Linear regression and Random Forest were used to predict base flow streamwater chemistry: EC, acid neutralization capacity (ANC), Ca, Mg, and SO4 based on properties of the catchment. RF models were superior to LR models, explaining 71% of the variance in EC,
61% in ANC, 92% in Ca, 58% in Mg, and 74% in SO4. The relative importance of different environmental factors in predicting stream chemistry varied among models, but on average rock chemistry > temperature > precipitation > soil = atmospheric deposition > vegetation > amount of rock/water contact > topography.

Link to Optima HWQ

- Example of successful water quality application of RF

Reference

Singh, N. K., Emanuel, R. E., Nippgen, F., McGlynn, B. L., & Miniat, C. F., 2018. The relative influence of storm and landscape characteristics on shallow groundwater responses in forested headwater catchments. Water Resources Research 54, 9883–9900. https://doi.org/10.1029/2018WR022681

Field of Science

Environmental Sciences, groundwater quantity

Summary techniques and results

Field observations of shallow groundwater levels combined with random forest modeling were used to identify the factors that affect shallow groundwater responses and the relative influences of key response drivers. Event response metrics (e.g. initial response timing, time to peak) were best predicted with storm depth, antecedent

groundwater level, profile curvature, and mean intensity (3).The model performance varied from 32 to 75% and were consistently high (>60%) for initial response timing and relatively low (<40%) for time to peak.

The RF model parameters included the number of trees generated (ntree = 1,000; default = 500) and mtry = 5 (default) as these parameters have minimal effect on RF model outcomes. The RF modeling was conducted with “randomForest 4.6–12” in R.

Link to Optima HWQ

- Example of RF application for groundwater-rainfall relation, yet not with a very high performance
- randomForest 4.6–12 package in R used

Reference

Bhattarai, A., Dhakal, S., Gautam, Y., Bhattarai, R., 2021. Prediction of Nitrate and Phosphorus Concentrations Using Machine Learning Algorithms in Watersheds with Different Landuse. Water 13, 3096. https://doi.org/10.3390/w13213096

Field of Science

Environmental Sciences, water quality

Summary techniques and results

The performance of nine different ML algorithms is evaluated to predict nitrate and phosphorus concentration for five different watersheds around lake Eerie (USA) using descriptive statistics for the watersheds.
MATLAB R2020a is used to implement all ML algorithms: Linear Regression (LR), k Nearest Neighbors (kNN), Regression Tree (RT), Ensemble (Bagging and Least-Squares Boosting (LSBoost)), Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), Gaussian Process Regression (GPR). Bayesian Optimization (BO) was used for the hyperparameter settings (such as tree size) for RT, ensemble, and GPR: BO is a hyperparameter search method applied in ML problems by minimizing a particular objective function (in this study mean squared error). Based on the past evaluation results of the objective function, an alternate function is established with BO to minimize its value.

Comparison tables for nitrate and phosphorus predictions:

Table 3. Comparison of R^2, RMSE, and MAE using different ML algorithms for nitrate prediction on test dataset.

Watershed	Parameter	LR	F-SVM	M-SVM	kNN	RF	ANN	RT-BO	Ensemble-BO	GPR-BO
Cuyahoga	R^2	0.404	0.751	0.745	0.690	0.689	0.754	0.745	0.749	0.752
	RMSE	1.0083	0.6748	0.6873	0.7300	0.7290	0.6670	0.6749	0.6702	0.6688
	MAE	0.7806	0.4836	0.4839	0.5360	0.5360	0.4860	0.4933	0.4900	0.4886
Grand River	R^2	0.039	0.152	0.188	0.023	0.079	0.093	0.214	0.173	0.145
	RMSE	0.3592	0.3479	0.3446	0.3840	0.3930	0.3740	0.3236	0.3335	0.3433
	MAE	0.2475	0.2334	0.2287	0.2740	0.2710	0.2570	0.2263	0.2345	0.2415
Maumee	R^2	0.160	0.463	0.462	0.282	0.387	0.479	0.466	0.470	0.477
	RMSE	2.5797	2.0970	2.0362	2.5140	2.4870	2.1780	2.2078	2.1863	2.1726
	MAE	2.1279	1.5809	1.5620	1.9500	1.8470	1.6880	1.7092	1.6886	1.6815
Raisin	R^2	0.251	0.466	0.476	0.409	0.406	0.485	0.468	0.485	0.482
	RMSE	1.8689	1.6611	1.5797	1.8310	1.9320	1.7130	1.7328	1.6797	1.7082
	MAE	1.4777	1.1395	1.0997	1.3080	1.3250	1.2310	1.2558	1.2159	1.2280
Sandusky	R^2	0.147	0.544	0.492	0.351	0.431	0.544	0.533	0.5395	0.542
	RMSE	2.6995	1.9639	2.0125	2.5060	2.5390	2.1590	2.1973	2.1850	2.1553
	MAE	2.1899	1.4242	1.4762	1.8670	1.8320	1.6340	1.6615	1.6662	1.6367

Table 6. Comparison of R^2, RMSE, and MAE using different ML algorithms for phosphorus prediction on test dataset.

Watershed	Parameter	LR	F-SVM	M-SVM	kNN	RF	ANN	RT-BO	Ensemble-BO	GPR-BO
Cuyahoga	R^2	0.800	0.574	0.778	0.720	0.754	0.829	0.778	0.808	0.820
	RMSE	0.1007	0.1013	0.0766	0.1000	0.1034	0.0901	0.0975	0.0941	0.0926
	MAE	0.0873	0.0618	0.0511	0.0765	0.0787	0.0737	0.0770	0.0745	0.0746
Grand River	R^2	0.665	0.483	0.633	0.590	0.676	0.665	0.720	0.733	0.718
	RMSE	0.0414	0.0505	0.0429	0.0294	0.0264	0.0264	0.0381	0.0369	0.0375
	MAE	0.0251	0.0266	0.0240	0.0111	0.0093	0.0097	0.0215	0.0204	0.0216
Maumee	R^2	0.800	0.800	0.850	0.708	0.812	0.846	0.822	0.838	0.842
	RMSE	0.0640	0.0629	0.0568	0.0743	0.0603	0.0556	0.0592	0.0587	0.0563
	MAE	0.0466	0.0414	0.0390	0.0514	0.0427	0.0390	0.0417	0.0407	0.0391
Raisin	R^2	0.652	0.688	0.762	0.561	0.618	0.680	0.671	0.709	0.697
	RMSE	0.0572	0.0507	0.0459	0.0578	0.0539	0.0505	0.0504	0.0488	0.0487
	MAE	0.0389	0.0264	0.0236	0.0341	0.0311	0.0308	0.0296	0.0285	0.0283
Sandusky	R^2	0.817	0.816	0.876	0.808	0.857	0.877	0.868	0.878	0.865
	RMSE	0.0895	0.0885	0.0752	0.0879	0.0774	0.0729	0.0758	0.0737	0.0751
	MAE	0.0567	0.0426	0.0395	0.0493	0.0416	0.0391	0.0403	0.0394	0.0399

Model performances and best method depended on land use (urban, agricultural, forested), probably due to different nutrient inputs and processes. Maumee, Raisin, and Sandusky are agricultural, Grand is forested, Cuyahoga is urban. Overall, the performance of ML methods was quite comparable. For nitrate, LR and kNN were outperformed by the other methods. For P, LR worked comparable with other methods. RF did not outperform other methods, number of trees was set quite low (10) which may explain that.

Parameter settings for the ML algoritms (or search spaces for BO) are available in the supplements https://www.mdpi.com/article/10.3390/w13213096/s1

Link to Optima HWQ

- Good model comparison for NO3 and P predictions in 5 watersheds
- Bayesian Optimization of hyperparameters (e.g. number of trees) useful?

Reference

Astuti, A.D., Aris, A.B., Salim, M.R., Azman, S., & Said, M.I., 2020. Artificial Intelligence Approach to Predicting River Water Quality: A Review.
https://doi.org/10.1155/2020/6659314

Field of Science

Environmental Sciences, water quality

Summary techniques and results

The objectives of this review were 1) to categorise AI methods comprehensively and 2) to discuss their advanced application to water quality modelling and prediction.

The authors observed increased used of support vectore machine (SVM) in water quality predictions; as well as potential of hybrid approaches (ANN-ARIMA, wavelet-ANFIS, wavelet-ANN).
The paper also states that ML algorithms should make the step from science into practice, although software is not yet user friendly enough.
Random Forest is mentioned, but not really covered in this review paper.

Link to Optima HWQ

- Application of wavelet transformation (daily/seasonal fluctuations) may be useful in some cases?
- Support for Optima-HWQ objectives
- Case for applying SVM for water quality predictions

Reference
Tyralis H, Papacharalampous G, Langousis A., 2019. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water 11(5). https://doi.org/10.3390/w11050910

Field of Science
Environmental Sciences, hydrology

Summary techniques and results
The purpose of this review study is to: (a) provide a comprehensive review of random forests and their software implementation for the practicing water scientist, (b) introduce their variants for possible use in water resources problems, and (c) familiarize the reader with the use of RF algorithms in water science, providing appropriate guidelines for full exploitation of their merits according to the broader literature.

Random Forests is a machine learning algorithm that combines the concepts of: classification and regression trees (using qualitative and quantitative predictors respectively), and bagging (bootstrap aggregation) with some additional degree of randomization (by randomly selecting candidate predictors for splitting rules).

RF is good at forecasting and big data, but the models are (together with SVM) along the least interpretable ones (opposite to linear models).

The default (hyper)parameter values in randomForest R package are satisfactory:

- ntree: number of trained trees, default set to 500. More trees, more accuracy and computation time.
- mtry: number of randomly selected predictor variables, default $p^{1/2}$ (root of number of parameters). Lower mtry, lower computation time. Default may be too small, optimization may be useful.
- nodesize (or maxnodes) gives the maximum nodes of the tree (and prevents further tree expansion). Default is 1 for classification tasks and 5 for regression tasks. Compared to ntree and mtry, this parameter has less impact on the performance.
- sampsize: number of observations sampsize used in each tree, default at n (total number of observations). Lower values (subsampling) may be faster, but value must be tuned.

Prediction variable selection can be based on variable importance metrics (VIMs). The two major VIMs used in RF applications are: the mean decrease in node impurities resulting from splitting, and the more advanced permutation VIM (based on reduction in accuracy in prediction of an out of bag (OOB) sample when permuting a predictor variable).
Excluding variables with VIMs that fluctuate around zero (low reduction in accuracy) is a reasonable assumption. This can be done stepwise either by progressively removing or adding predictors.

Arguments for using RF (quote from Efron and Hastie, 2016): "*Random forests and boosting live at the cutting edge of modern prediction methodology. They fit models of breathtaking complexity compared with classical linear regression, or even with standard GLM modeling as practiced in the late twentieth century. They are routinely used as prediction engines in a wide variety of industrial and scientific applications. For the more cautious, they provide a terrific benchmark for how well a traditional*

parameterized model is performing: if the random forests does much better, you probably have some work to do, by including some important interactions and the like".

Advantages of RF (references given in the paper):

1. They demonstrate increased predictive performance, as verified in competitions
2. They can capture non-linear dependencies between predictor and dependent variables
3. They are non-parametric, i.e., no parametric statistical model needs to be defined for their use
4. They are fast compared to other machine learning algorithms and, also, they can operate in parallel computing mode.
5. They can be applied to large-scale problems
6. They are straightforward to use
7. They do not overfit
8. They are stable
9. The number of model parameters is small, and the default values in corresponding software implementations are properly set and the algorithm is robust to changes of the parameters
10. They are robust to the inclusion of noisy predictor variables
11. They can handle highly correlated predictor variables
12. They can operate successfully when interactions are present
13. They are flexible (i.e., there is a large potential for modifications) while there is a large number of variants of random forests designed to perform different tasks
14. They permit ranking of the relative significance of predictor variables, through variable importance metrics (VIMs)
15. Variable selection procedures, based on VIMs, can be combined with other machine learning algorithms
16. They can effectively handle small sample sizes
17. They are suitable for coping with high dimensional data (i.e., of the form n << p (much more predictors than samples)
18. They can simultaneously incorporate continuous and categorical variables
19. They can be used to solve problems with many classes of the response variable
20. They are invariant to monotone transformations of the predictor variables
21. They can effectively handle missing data
22. There exist free software implementations of RF algorithms, with most variants and extensions been available as contributed packages in the R programming language

However (disadvantages):

1. The theoretical properties of random forests are not fully understood, and they are usually interpreted based on simplified/stylized versions of the algorithm
2. Random forests cannot extrapolate outside the training range
3. Variable importance metrics (VIMs) are not always reliable, as they are affected by high correlations and interactions
4. Random forests are harder to interpret/understand compared to single trees
5. The automation of random forests may result in a slight decrease of their predictive performance compared to e.g., highly parameterized tree-based
6. They cannot adequately model datasets with imbalanced data (i.e., datasets in which the number of observations of the response variable belonging to one class differs significantly compared to other classes
7. Their original version is not suited for causal inference

The paper gives a list of RF variants and a list of related R packages (among which CALIBERrfimpute as describes by Shah et al., 2014). Also, R packages for variable selection / variable importance.

Link to Optima HWQ

- Very good overview and explanations around Random Forests
- Links to relevant r packages for parameter selection and missing data imputation (CALIBERrfimpute).
- CALIBERrfimpute may be an alternative for the missForest package. Developers (Shah et al., 2014) claim that it performs better, although it was tested on a different type of dataset (patient database).
-

Reference

Zhang, Y. F., Thorburn, P. J., Xiang, W., & Fitch, P. (2019). SSIM—A deep learning approach for recovering missing time series sensor data. IEEE Internet of Things Journal, 6(4), 6618-6628.

and

Y. Zhang, P.J. Thorburn, 2021. A dual-head attention model for time series data imputation, Comput. Electron. Agric. 189, http://dx.doi.org/10.1016/j.compag.2021.106377.

Field of Science

Environmental Sciences, water quality

Summary techniques and results

Zhang et al. (2019) propose a new method for filling in missing data in high frequency water quality time series: the sequence-to-sequence imputation model (SSIM). The method is designed for situations where all measured parameters have a simultaneous gap. SSIM uses sequence-to-sequence deep learning architecture, and the Long Short-Term Memory Network is chosen to utilize both the past and future information for variable time periods. The Variable-Length Sliding Window Algorithm (VLSW) increase the amount of training samples and thereby improves the results from the deep learning algorithm.

The method is evaluated using water quality data from two monitoring sites in the Mulgrave-Russell catchment in the Great Barrier Reef, Australia. The parameters include temperature, rainfall, evaporation, radiation, vapour pressure, electrical conductivity, water discharge, water level, turbidity and nitrate. SSIM outperformed the other methods ARIMA, SARIMA, Matrix Factorization (MF), Multiple Imputation by Chained Equations (MICE) and Expectation Maximization (EM).

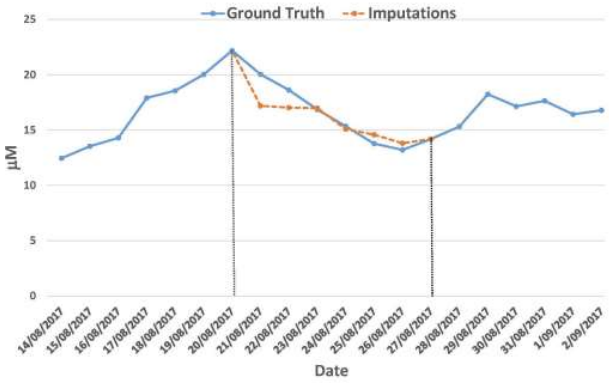


Figure 3: Recovering 6 missing nitrogen data from 21/8/2017 to 26/8/2017

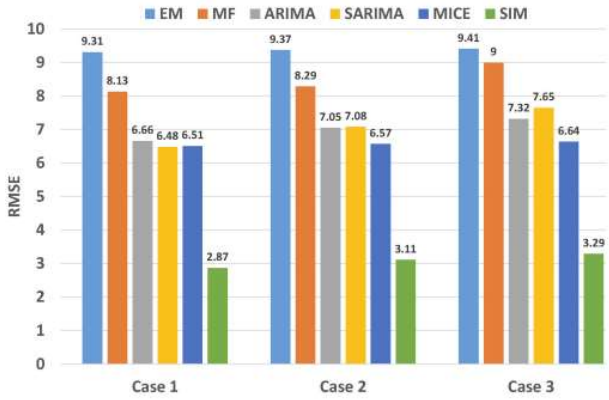
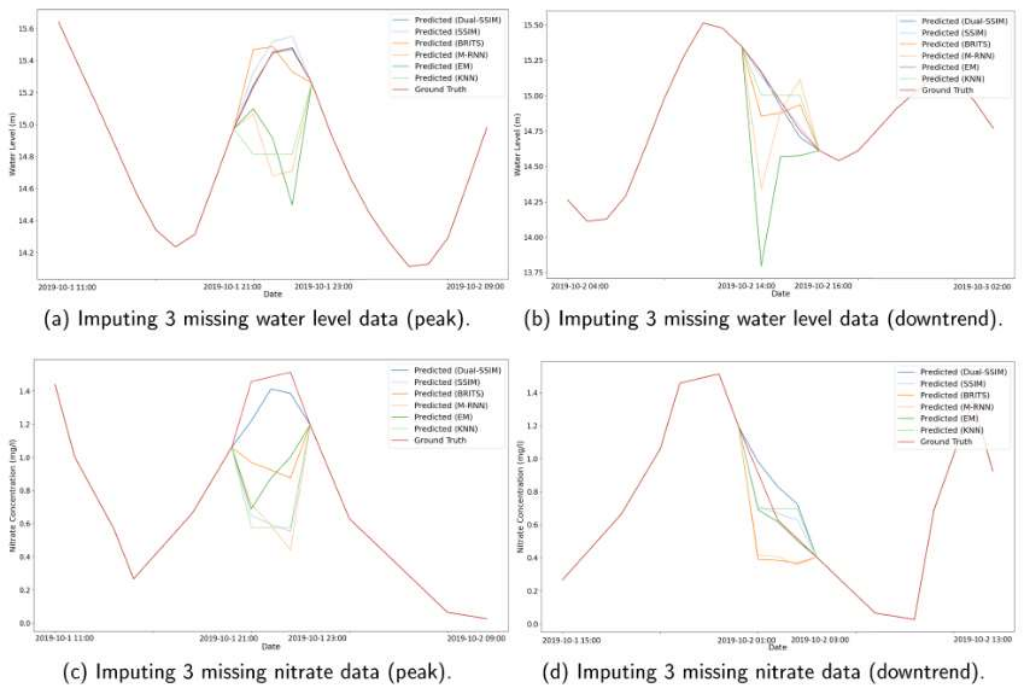


Figure 4: Evaluation of missing data imputation using the RMSE.

Zhang & Thorburn (2021) propose the Dual SSIM method, in which the data on both sides of the data gap are processed separately. Experimental results demonstrated that Dual-SSIM outperforms other SSIM and the other methods EM, KNN, BRITS and M-RNN.



(a) Imputing 3 missing water level data (peak).

(b) Imputing 3 missing water level data (downtrend).

(c) Imputing 3 missing nitrate data (peak).

(d) Imputing 3 missing nitrate data (downtrend).

Link to Optima HWQ

- SSIM and Dual-SSIM could be useful in cases where no predictor data is available (because there is only 1 parameter, or because all sensors give simultaneous data gaps

Reference

Zhang, Y. & P. J. Thorburn, 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Generation Computer Systems 128. https://doi.org/10.1016/j.future.2021.09.033

Field of Science

Environmental Sciences, water quality

Summary techniques and results

This work presents an overview of missing data imputation techniques applied to water quality sensor data. A cloud-based data imputation system was developed including several imputation methods. This includes:

Statistical methods:

- mean imputation
- last observation carried forward (LOCF)
- linear imputation)

Model based methods:

- Expectation Maximization (EM, maximum likelihood estimation)
- Multiple imputations by chained equations (MICE)
- K-nearest neighbour (KNN)

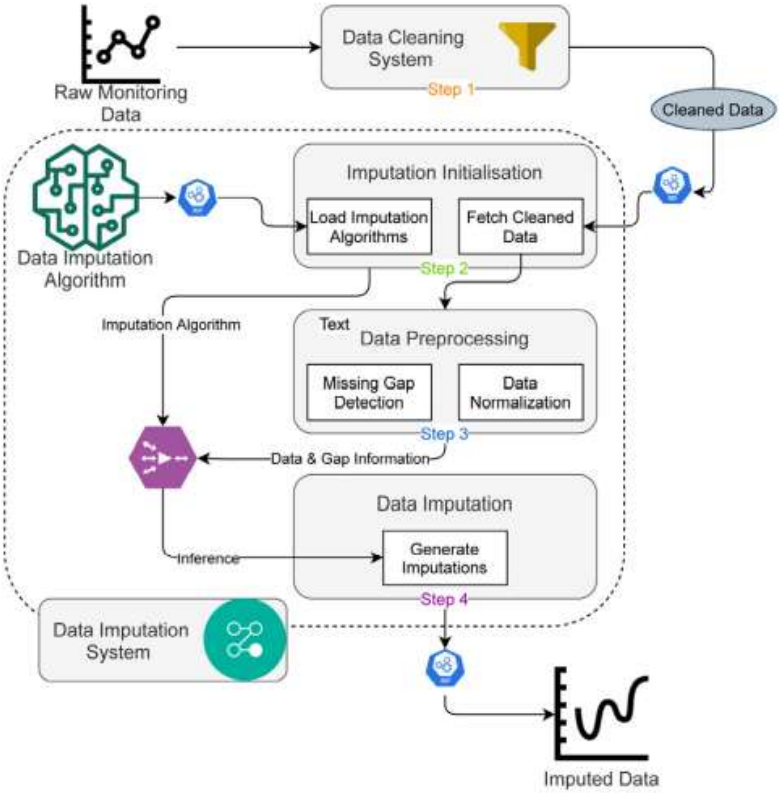
Neural-network based methods:

- sequence-to-sequence imputation model (SSIM)

- Dual SSIM
- BRITS
- Multi-directional Recurrent Neural Network (M-RNN)

The steps of the data imputation system are shown in the figure below:

- Water Quality Monitoring Data Cleaning (remove invalid, out of range values)
- Imputation System Initialization (select and load the data imputation algorithm)
- Data Preprocessing (normalization, rescaling, gap identification)
- Imputation Data Generation



The strengths and limitations of the different methods are listed in the table below. In the graphs, the performance of the methods for reproducing a nitrate concentration peak are compared, and the effect of gap sizes on the performance is shown.

Method	Strength	Limitation
Mean Imputation		The variability in the data is reduced;the standard deviations and the variance estimates may get underestimated [35],
LOCF	Easy to understand, Efficient to apply	The assumption is mostly unrealistic [36],
Linear Imputation		There should be a linear relationship between the predictor and response variables
EM MICE KNN	Good Interpretability, Lazy Learning, do not build models from training data	High risk to overfit the training data Computationally expensive for large datasets
SSIM Dual-SSIM BRITS M-RNN	Can capture and use the temporal information, Deep architecture brings strong representation learning	Black box method, Performance heavily relies on hyperparameters tuning, High computational cost to build the model

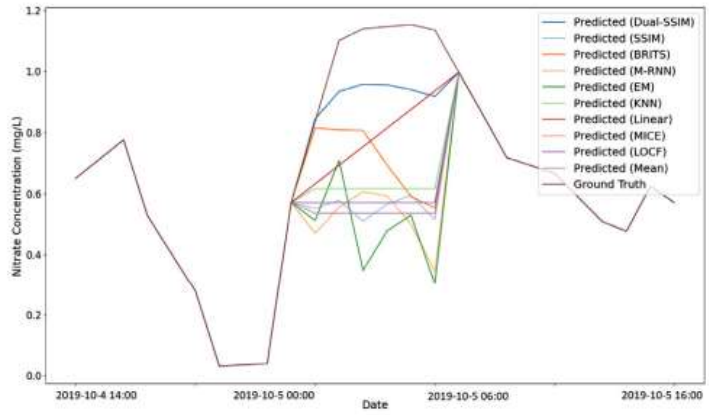


Fig. 6. Model outputs in imputing 6 consecutive missing values for nitrate concentration from GBR monitoring network. The solid red-brown line represents the ground truth data. Other lines represent the imputation results generated by different models. 20 available data before and after the gap are used as the model's input.

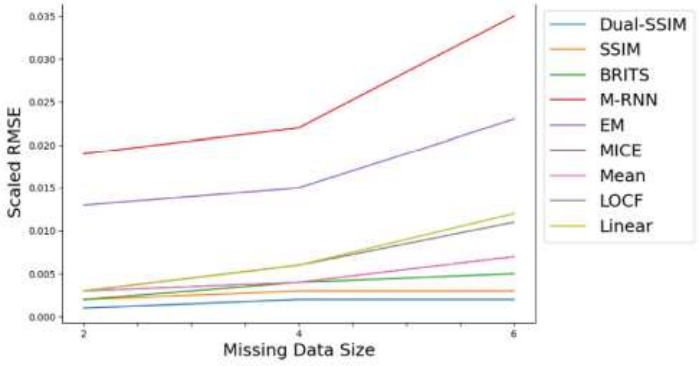


Fig. 7. Imputation accuracy for estimating data with different size by different methods.

Link to Optima HWQ

- In case this cloud-based system is open, it may be useful for Optima-HWQ. The focus is on filling relatively short gaps without using predictor variables.

Reference

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

Field of Science

Environmental Sciences, water quality

Summary techniques and results

This paper summarizes a framework for detecting outliers which is implemented in the R-package *oddwater*. The examples focus on turbidity, conductivity, and river level data from the sensor network in rivers flowing into the Great Barrier Reef (Australia). The focus is on outliers involving abrupt changes in value, including sudden spikes,

sudden isolated drops and level shifts. The methods can be applied in the high dimensional space, including the correlation between parameters for outlier detection.

The included techniques are unsupervised. This means that they do not need training data with pre-labeled known outliers. In practice, not all potential outlier types are known beforehand, and training data are often not available, which complicates the application of (semi-)supervised methods.

The workflow of the proposed framework is shown in the Figure below.

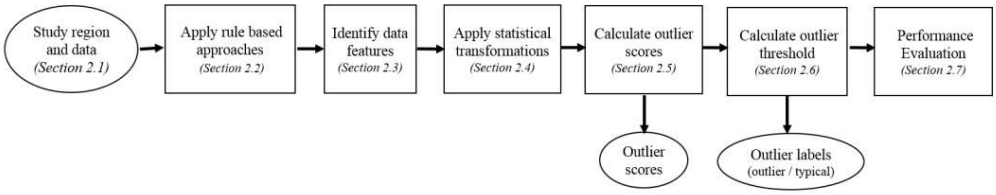


Figure 1: *The proposed framework for outlier detection in water quality data from in situ sensors. Squares represents the main steps involved. Circles correspond to input and output.*

- The rule-based approaches involve detecting out-of-range, impossible and missing values.
- The types of outliers are identified (data features).
- Statistical transformations often help to highlight different types of outliers. Transformations included are log transformation, first difference, time gap, first derivative, one sided derivative, rate of change, and relative difference.
- Outlier scores are calculated using eight unsupervised outlier scoring techniques (KNN-SUM, KNN-AGG, HDoutliers, LOF, COF, INFLO, LDOF, RKOF).
- The extreme value theory (EVT) was used to calculate the outlier score threshold for distinguishing outliers from typical data points.

The example below shows that there can still be outliers that are not detected (False Negatives) or typical data points that are misclassified as outliers (False Positives).

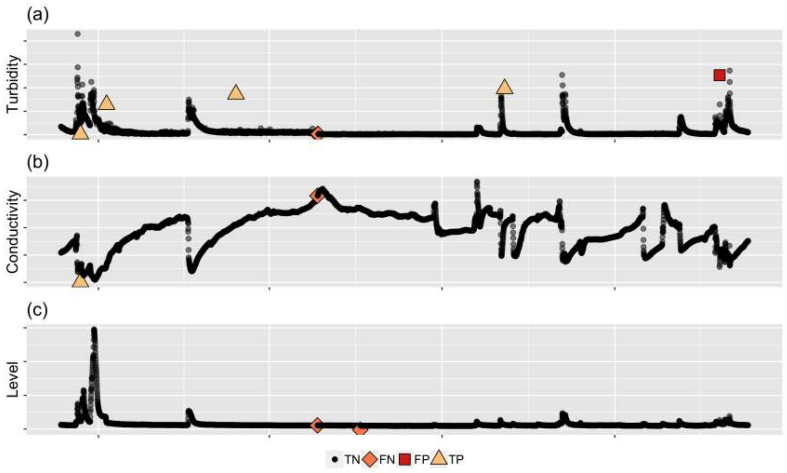


Figure 7: *Classification of turbidity (T), conductivity (C), level (L) observations measured by in situ sensors at Sandy Creek by KNN-SUM algorithm as true negatives (TN), true positives (TP), false negatives (FN), false positives (FP) when applied to the TCL combination with one sided derivative transformation.*

A publication list, R packages and presentations from the project 'Revolutionising water quality monitoring in the information age' of the Centre for Data Science of Queensland University of technology are available at: https://research.qut.edu.au/qutcds/projects/revolutionising-water-quality-monitoring/#Videos

Link to Optima HWQ

- The R-package oddwater may be useful
- The paper shows that statistical transformations can be useful for outlier detection
- The paper also shows that taking account of the correlation structure between parameters can help to detect outliers
- The approach with multiple statistical transformations and outlier scoring techniques makes the framework flexible.
- Follow-up work in the R package conduits includes information from nearby upstream sensors (predict downstream concentrations based on upstream sensor data, deviations are anomalies).

B Voorbeeld data optimalisatie

voorbereiding workshop

Voorbereiding workshop

```
library(tidyverse)
library(leaflet)
library(sf)
library(plotly)
# library(anomalize)
library(slider)
library(anomaly)

# read custom functions
source('../scripts/test_get_tsfeatures.R')
```

Data inlezen

```
locs <- read_csv2("../data/raw/locatie_mapping.csv")
sensor_lims <- read_csv2("../data/raw/meet_range_sensoren.csv") %>%
  filter(path_length %in% c(1, NA))
par_ranges <- readxl::read_excel('../data/raw/validation_master_v1.xlsx', sheet = 'realistic_
df <-
  read_csv("../data/intermediate/data_voorbereidingworkshop.csv") %>%
   mutate(
    sensor = case_when(
      sensor == 'C4E' ~ 'Ponsel C4E',
      sensor == 'PHEHT' ~ 'Ponsel PHEHT',
      TRUE ~ sensor
    ),
    datetime = ymd_hms(datetime, tz = 'CET')
  )
```

```r
df_extra <- read_csv("../data/intermediate/data_voorbereidingworkshop_extra.csv") %>%
   mutate(
    sensor = case_when(
      sensor == 'C4E' ~ 'Ponsel C4E',
      sensor == 'PHEHT' ~ 'Ponsel PHEHT',
      TRUE ~ sensor
    ),
    datetime = ymd_hms(datetime, tz = 'CET')
  )

lims <- c(0.0001, 50)

oh <- read_csv('../data/raw/data_oh_voor_workshop.csv') %>%
  mutate(datum = ymd_hms(datum, tz = 'CET'))

oh_saqc <- oh %>% select(datum_start = datum) %>%
  mutate(maint = datum_start + hours(3))

df_lab_cor <- read_csv('../data/raw/lab_cor_data_workshop.csv') %>%
  mutate(datetime_compare = ymd_hms(datetime_compare, tz = 'CET'))
```

criteria tabellen

```r
sensor_lims %>%
  knitr::kable(caption = "Sensor ranges, dit geval OTT EcoN NO3 Nnf")
```

Table 1: Sensor ranges, dit geval OTT EcoN NO3 Nnf

sensor	parameter_code	hoedanigheid_code	eenheid_code	ondergrens	bovengrens	path_length
Ponsel PHEHT	pH	NVT	DIMSLS	0	14	NA
Ponsel PHEHT	T	NVT	oC	-5	50	NA
Ponsel C4E	T	NVT	oC	-5	50	NA
Ponsel C4E	GELDHD	NVT	uS/cm	0	2000	NA
Ponsel OPTOD	O2	NVT	mg/l	0	20	NA
Ponsel OPTOD	O2	NVT	%	0	200	NA

sensor	parameter_code	hoedanigheid_code	eenheid_code	ondergrens	bovengrens	path_length
Ponsel OPTOD	T	NVT	oC	-5	50	NA
Eureka Manta +35	pH	NVT	DIMSLS	0	14	NA
Eureka Manta +35	GELDHD	NVT	uS/cm	0	5000	NA
Eureka Manta +35	O2	NVT	mg/l	0	50	NA
Eureka Manta +35	O2	NVT	%	0	500	NA
Eureka Manta +35	T	NVT	oC	-5	50	NA
YSI EXO 3	pH	NVT	DIMSLS	0	14	NA
YSI EXO 3	GELDHD	NVT	mS/cm	0	100	NA
YSI EXO 3	GELDHD	NVT	uS/cm	0	100000	NA
YSI EXO 3	O2	NVT	mg/l	0	50	NA
YSI EXO 3	O2	NVT	%	0	500	NA
YSI EXO 3	T	NVT	oC	-5	50	NA
Trios NICO	NO3	nf	mg/l	0	266	1
Trios NICO	NO3	Nnf	mg/l	0	60	1
OTT EcoN	NO3	nf	mg/l	0	266	1
OTT EcoN	NO3	Nnf	mg/l	0	60	1
YSI EXO 2	pH	NVT	DIMSLS	0	14	NA
YSI EXO 2	GELDHD	NVT	uS/cm	0	100000	NA
YSI EXO 2	O2	NVT	mg/l	0	50	NA
YSI EXO 2	O2	NVT	%	0	500	NA
YSI EXO 2	T	NVT	oC	-5	50	NA

```
par_ranges %>% select(parameter_id, jaar_min, jaar_max) %>% knitr::kable(caption = "Realistis
```

Table 2: Realistische ranges per parameter, dit geval Nitraat_N_mg/l

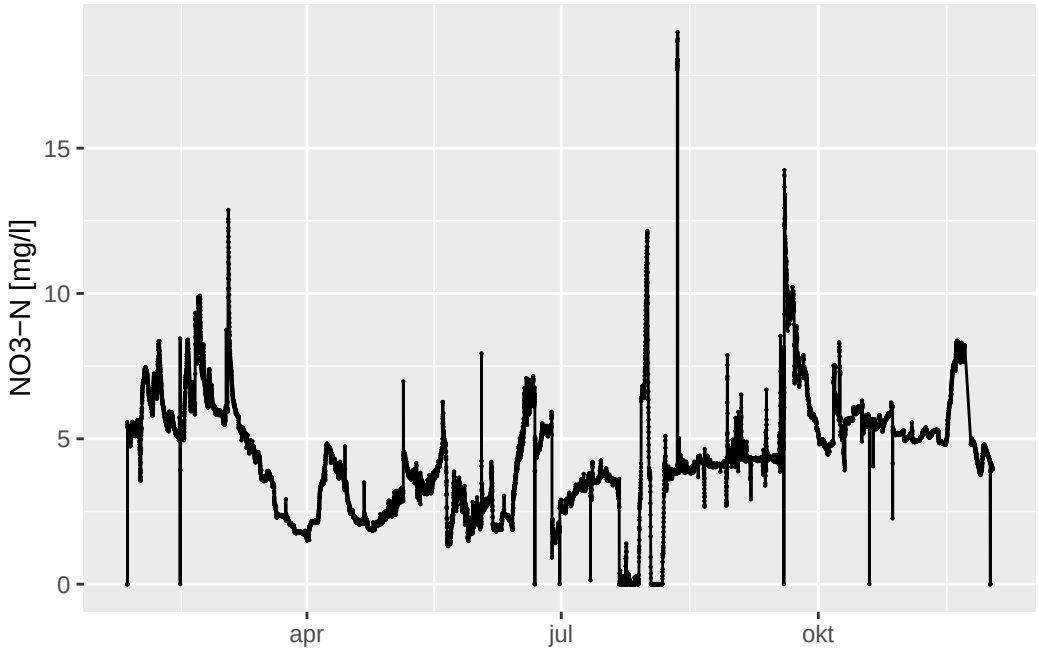
parameter_id	jaar_min	jaar_max
Ammonium_mg/l	0.00	1000.0
Ammonium_N_mg/l_AW	0.00	1000.0
Ammonium_N_mg/l_OW	0.00	1000.0

parameter_id	jaar_min	jaar_max
Chlorofyl-a_blauwalg_ug/l	0.00	5000.0
Chlorofyl-a_groenalg_ug/l	0.00	5000.0
Doorzicht_dm	0.00	617.0
Geleidendheid_25oC_uS/cm	100.00	1000.0
Geleidendheid_25oC_mS/m	10.00	100.0
Lichtintensiteit_1/m	0.00	28000.0
Lichtintensiteit_ref_A_212_nm_SNR	0.00	28000.0
Lichtintensiteit_ref_B_254_nm_SNR	0.00	28000.0
Lichtintensiteit_ref_C_360_nm_SNR	0.00	28000.0
Lichtintensiteit_ref_D_reference_diode_SNR	0.00	28000.0
Nitraat_mg/l	0.00	50.0
Nitraat_N_mg/l	0.00	50.0
pH	2.00	13.0
Ptot_P_mg/l	0.00	4.0
Redoxpotentiaal_mV	-400.00	800.0
Relative_fluoresence_units_blauwalg	0.00	100.0
Relative_fluoresence_units_groenalg	0.00	100.0
SAC_254_nm	0.00	1500.0
Saliniteit_PSU	0.05	0.5
SQI	0.00	1.0
Temperatuur_oC_OW	NA	NA
Totaal_opgeloste_bestanddelen_mg/l	0.00	500.0
Troebelheid_NTU	0.00	10000.0
TRP_mg/l	0.00	1.0
Zuurstof_verzadiging_%	0.00	300.0
Zuurstof_opgelost_mg/l	NA	NA
Chlorofyl-a_blauwalg_ug/l	0.00	400.0
Chlorofyl-a_blauwalg_ug/l	0.00	400.0
Chlorofyl-a_groenalg_ug/l	0.00	400.0
Chlorofyl-a_groenalg_ug/l	0.00	400.0
Geleidendheid_25oC_mS/m	25.00	100.0
Geleidendheid_25oC_mS/m	25.00	250.0
Geleidendheid_25oC_mS/m	30.00	260.0
Geleidendheid_25oC_mS/m	43.00	54.0
Geleidendheid_25oC_mS/m	20.00	83.0
Temperatuur_oC_OW	1.00	28.0
Temperatuur_oC_OW	6.00	28.0
Temperatuur_oC_OW	9.00	30.0
Temperatuur_oC_OW	4.00	28.0
Temperatuur_oC_OW	4.00	28.0
pH	6.00	11.0

parameter_id	jaar_min	jaar_max
pH	6.00	11.0
pH	6.00	9.0
pH	7.00	11.0
pH	7.00	11.0
Zuurstof_verzadiging_%	0.00	180.0
Zuurstof_verzadiging_%	0.00	130.0
Zuurstof_verzadiging_%	0.00	100.0
Zuurstof_verzadiging_%	0.00	250.0
Zuurstof_verzadiging_%	0.00	250.0
Zuurstof_opgelost_mg/l	NA	NA
Zuurstof_opgelost_mg/l	NA	NA
Zuurstof_opgelost_mg/l	NA	NA
Zuurstof_opgelost_mg/l	NA	NA
Zuurstof_opgelost_mg/l	NA	NA

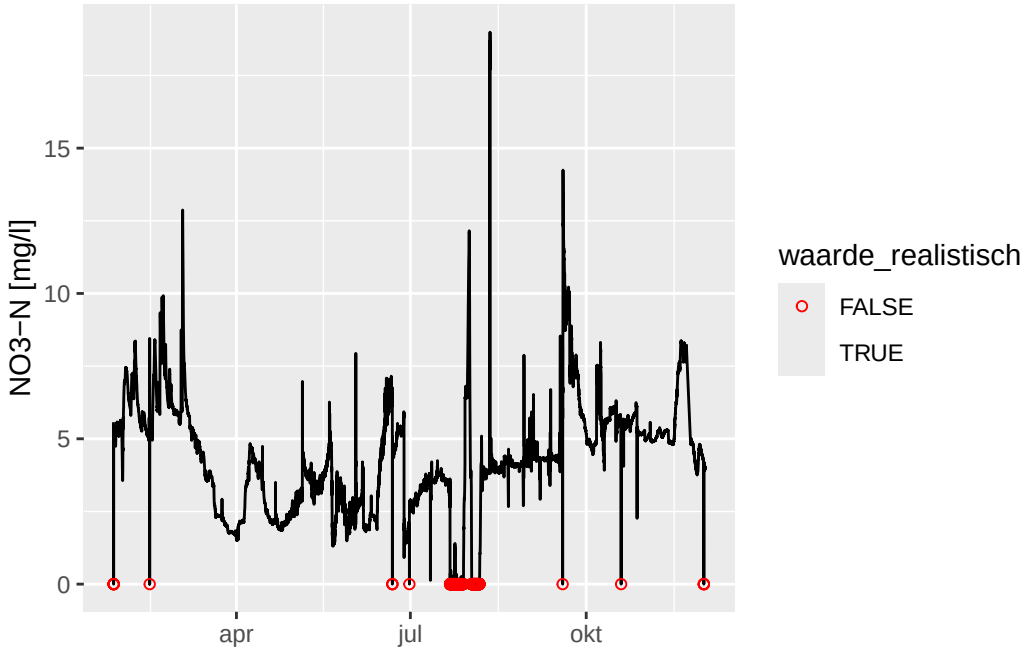
Visualiseer originele data serie

```
p <- ggplot(df, aes(x = datetime, y = waarde)) +
  geom_point(size = 0.1, shape = 1) +
  geom_line() +
  labs(x = NULL, y = "NO3-N [mg/l]")

#ggplotly(p)
p
```



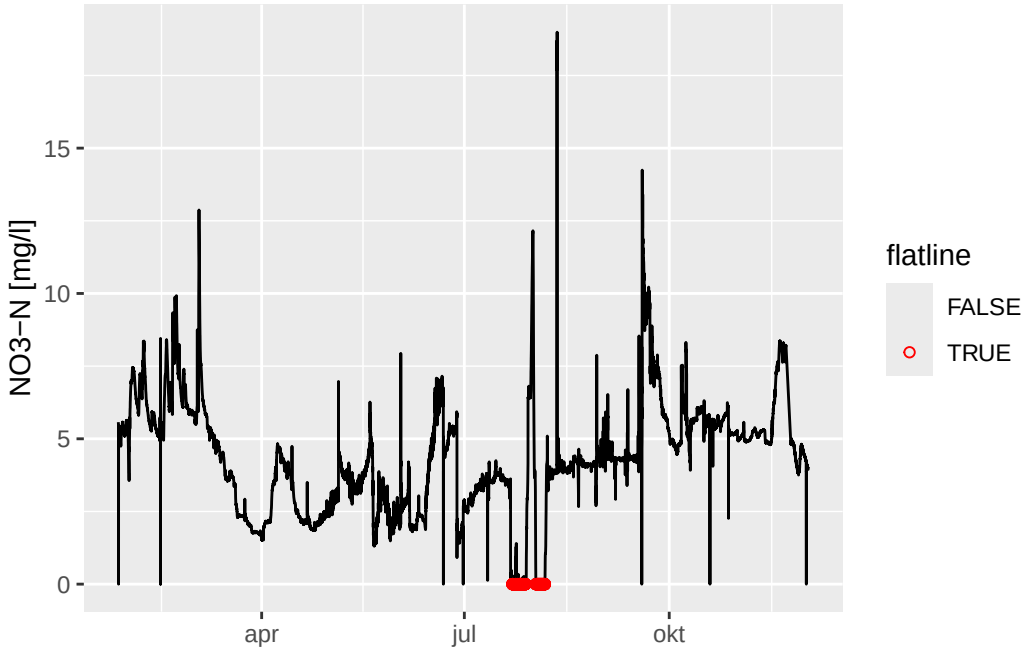
Check op onrealistische waarden

```
df <- df %>% mutate(waarde_realistisch = between(waarde, left = lims[1], right = lims[2]))

ggplot(df, aes(x = datetime, y = waarde)) +
  geom_line() +
      geom_point(aes(col = waarde_realistisch, shape = waarde_realistisch)) +
      labs(x = NULL, y = "NO3-N [mg/l]") +
      scale_color_manual(values = c("red", "black")) +
      scale_shape_manual(values = c(1, NA))
```



Check op flatlines

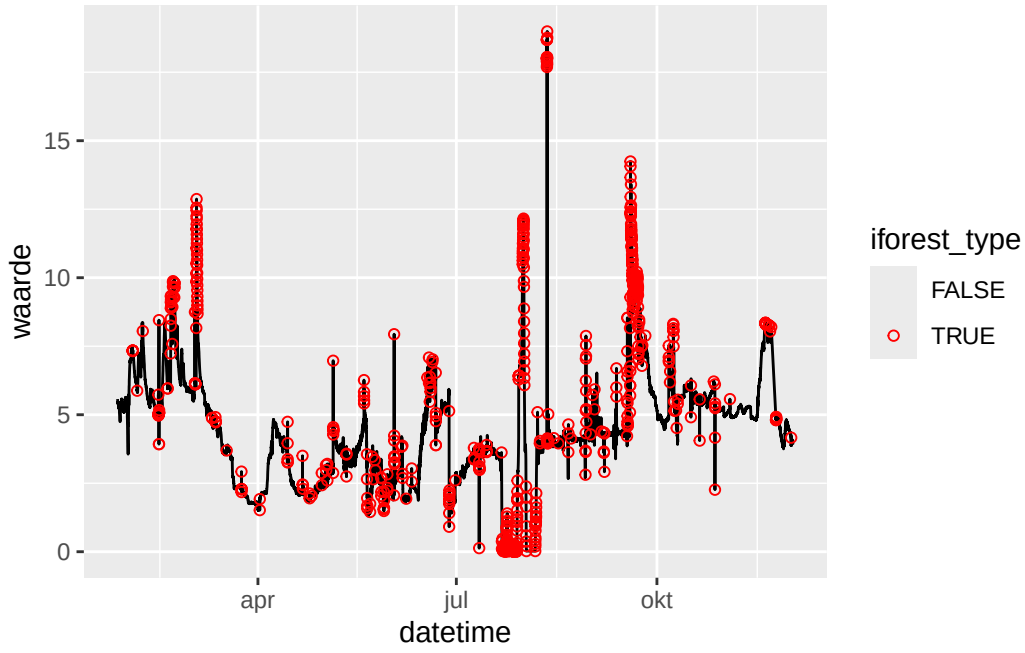
```
df <- get_flatline(dataset = df, value_col = "waarde", threshold = 0, n_measurements = 10)


ggplot(filter(df, !is.na(flatline)), aes(x = datetime, y = waarde)) +
  geom_line() +
      geom_point(aes(col = flatline, shape = flatline)) +
      labs(x = NULL, y = "NO3-N [mg/l]") +
      scale_color_manual(values = c("black", "red")) +
      scale_shape_manual(values = c(NA, 1))
```



Check overige outliers

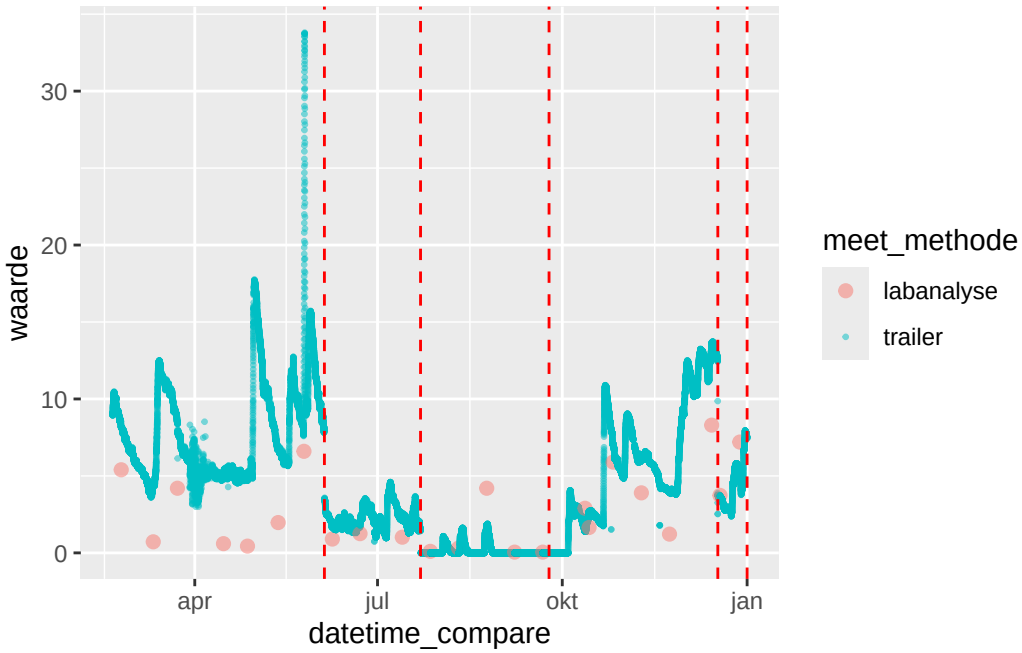
```
df_trans <- filter(df, flatline == FALSE, waarde_realistisch == TRUE) %>%
  distinct(datetime, waarde) %>%
    group_by(datetime) %>%
    summarise(waarde = mean(waarde, na.rm = TRUE)) %>%
    ungroup() %>%
    transform_data_uv(time_col = 'datetime')

  df_trans$features <- slide(df_trans, ~get_tsfeatures(.x$waarde), .before = 9, .complete = I
  df_trans <- df_trans %>% unnest_wider(features)

col_keep <- is.na(df_trans) %>% colMeans()
col_keep <- names(col_keep[col_keep < 1])
col_keep <- col_keep[!col_keep %in%  c('datetime', 'waarde', 'time')]
df_testout <- df_trans[, col_keep]

out_forest <- isotree::isolation.forest(data = df_testout, ntrees = 500)

df_testout$datetime <- df_trans$datetime
df_testout$waarde <- df_trans$waarde
```

```r
df_testout$iforest_score <- predict(out_forest, df_testout)
iforest_th <- quantile(df_testout$iforest_score, probs = 0.95, na.rm = TRUE, names = FALSE)
df_testout$iforest_type <- df_testout$iforest_score >= iforest_th

#plot
ggplot(df_testout, aes(x = datetime, y = waarde)) +
  geom_line() +
      geom_point(aes(col = iforest_type, shape = iforest_type)) +
      # labs(title = par) +
      scale_color_manual(values = c("black", "red")) +
      scale_shape_manual(values = c(NA, 1))
```



Lab correctie

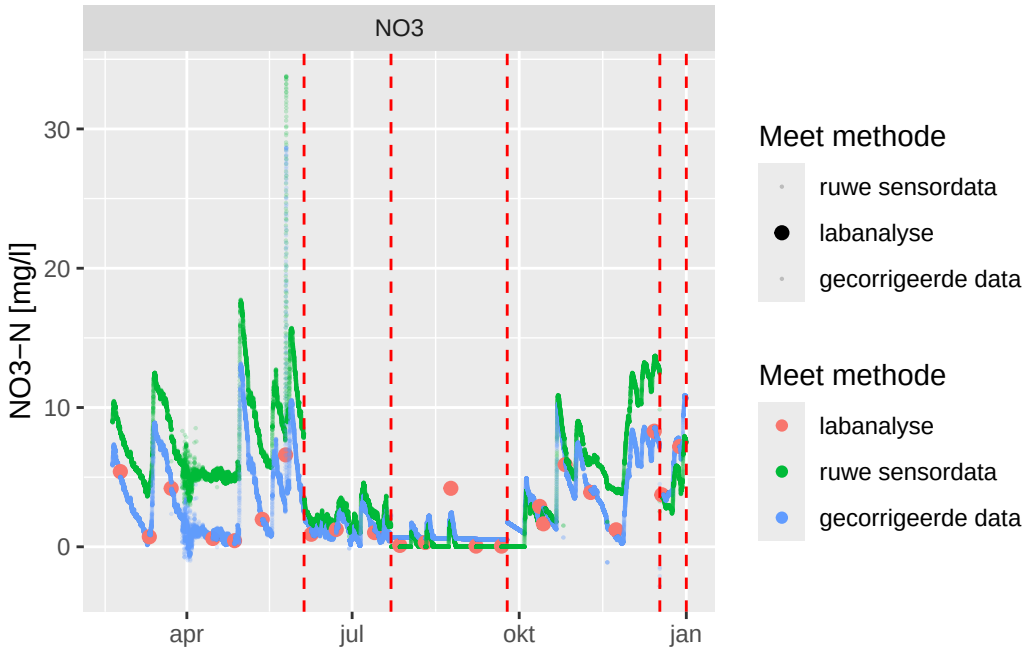
```r
#plot
ggplot(data = df_lab_cor, aes(x = datetime_compare, y = waarde, size = meet_methode, col = me
  geom_point(alpha = 0.5) +
  scale_size_manual(values = c(2, 0.5)) +
  geom_vline(data = oh, aes(xintercept = datum), linetype = "dashed", col = "red")
```




```
df_lab_cor_wide <- pivot_wider(df_lab_cor, names_from = meet_methode, values_from = waarde, v
    mutate(verschil = trailer - labanalyse)

df_lab_cor_wide_saqc <- df_lab_cor_wide %>% select(datetime = datetime_compare, NO3 = trailer

lab_corrected_data_mntnc <- NULL
# date <- "2021-06-04 13:00:00 CEST"
for (date in unique(oh$datum)) {

    date_idx <- which(date == unique(oh$datum)) -1
    prev_date <- ifelse(date_idx == 0, ymd_hms("2021-01-01 08:00:00", tz = "CET"), unique(oh$
    df_mntnc <- filter(df_lab_cor_wide, datetime_compare > prev_date, datetime_compare < date

    # test plot #
    ggplot(df_mntnc, aes(x = trailer, y = labanalyse)) +
      geom_point() +
      geom_smooth(method = "lm", se = FALSE) +
      geom_abline(col = "red", linetype = "dashed")

    # models #
    # normal lm model
    lm_mod <- lm(labanalyse ~ trailer, data = filter(df_mntnc, !is.na(labanalyse)))
```

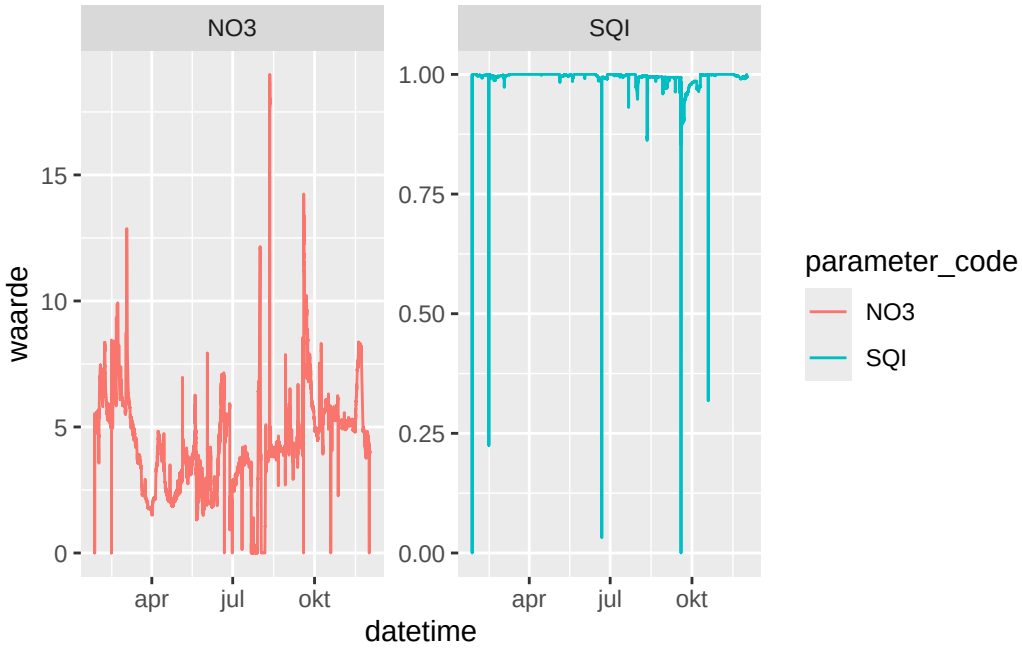
```
    # diff lm model #
    lm_mod_diff <- lm(verschil ~ datetime_compare, data = filter(df_mntnc, !is.na(labanalyse)

    df_mntnc <-
      mutate(
        df_mntnc,
        trailer_correctie_lm = predict(lm_mod, newdata = df_mntnc),
        trailer_correctie_lm_diff_raw = predict(lm_mod_diff, newdata = df_mntnc)
      ) %>%
      mutate(trailer_correctie_lm_diff = trailer - trailer_correctie_lm_diff_raw,) %>%
      select(-trailer_correctie_lm_diff_raw,-verschil)

    df_mntnc_corrected <-
      pivot_longer(
        df_mntnc,
        cols = trailer:trailer_correctie_lm_diff,
        names_to = "meet_methode",
        values_to = "waarde"
      )


    #plot#
    ggplot(df_mntnc_corrected,
           aes(x = datetime_compare, y = waarde, col = meet_methode)) +
      geom_point(aes(size = meet_methode, alpha = meet_methode)) +
      scale_alpha_manual(
        values = c(0.2, 1, 0.2, 0.2),
        breaks = unique(df_mntnc_corrected$meet_methode)
      ) +
      scale_size_manual(
        values = c(0.1,  2, 0.1, 0.1),
        breaks = unique(df_mntnc_corrected$meet_methode)
      ) +
      labs(x = NULL)

    lab_corrected_data_mntnc <-
      bind_rows(lab_corrected_data_mntnc, df_mntnc_corrected)

}

# plot
```

```r
ggplot(filter(lab_corrected_data_mntnc, meet_methode != 'trailer_correctie_lm'),
       aes(x = datetime_compare, y = waarde, col = meet_methode)) +
  geom_point(aes(size = meet_methode, alpha = meet_methode)) +
  scale_color_discrete(labels = c('labanalyse', 'ruwe sensordata', 'gecorrigeerde data'), nam
  scale_alpha_manual(
    values = c(0.2, 1, 0.2, 0.2),
    breaks = unique(df_mntnc_corrected$meet_methode),
    labels = c('ruwe sensordata', 'labanalyse', "trailer_correctie_lm", 'gecorrigeerde data')
    name =  'Meet methode'
  ) +
  scale_size_manual(
    values = c(0.1,  2, 0.1, 0.1),
    breaks = unique(df_mntnc_corrected$meet_methode),
    labels = c('ruwe sensordata', 'labanalyse', "trailer_correctie_lm", 'gecorrigeerde data')
    name =  'Meet methode'
  ) +
  geom_vline(
    data = oh,
    aes(xintercept = datum),
    linetype = "dashed",
    col = "red"
  ) +
  facet_wrap( ~ parameter_code, scales = "free", nrow = 3) +
  labs(x = NULL, y = 'NO3-N [mg/l]')
```



Twee variabelen

```
df_extra2 <- select(df_extra, datetime, parameter_code, waarde) %>%
  arrange(parameter_code, datetime)

ggplot(df_extra2, aes(x = datetime, y = waarde, col = parameter_code)) +
  geom_line() +
    facet_wrap(~parameter_code, scales = 'free')
```




```
df_extra2$waarde[df_extra2$parameter_code == 'SQI' & df_extra2$waarde == 0] <-  0.001

df_extra2 <- mutate(df_extra2, waarde_realistisch = between(waarde, left = lims[1], right = l
  filter(waarde_realistisch == TRUE)

df_etrxa_wide <- select(df_extra2, datetime:waarde) %>% pivot_wider( names_from = 'parameter_

df_extra_trans <- oddwater::transform_data(df_etrxa_wide, time_col = 'datetime')
col_keep <- is.na(df_extra_trans) %>% colMeans()
col_keep <- names(col_keep[col_keep < 1])
col_keep <- col_keep[!col_keep %in%  c(colnames(df_etrxa_wide), 'time')]

nn_hd <- oddwater::NN_HD(na.omit(df_extra_trans[, col_keep]))
hd_out <- stray::find_HDoutliers(df_extra_trans[, col_keep])
out_forest_extra <- isotree::isolation.forest(df_extra_trans[, col_keep])
```

SaQC

```
# import saqc
# import pandas as pd
# import matplotlib.pyplot as plt
```

```
#
# data = pd.read_csv('../data/raw/hydro_data.csv', index_col=0)
# maint = pd.read_csv('../data/raw/hydro_maint.csv', index_col=0)
# maint.index = pd.DatetimeIndex(maint.index)
# data.index = pd.DatetimeIndex(data.index)
# qc = saqc.SaQC([data, maint])  # dataframes "data" and "maint" are integrated internally
#
# qc = qc.flagManual('sac254_raw', mdata='maint', method='closed', label='Maintenance')
# qc = qc.flagRange('level_raw', min=0)
# qc = qc.flagRange('water_temp_raw', min=-1, max=40)
# qc = qc.flagRange('sac254_raw', min=0, max=60)
#
# # qc.plot('sac254_raw')
# qc = qc.align(['sac254_raw', 'level_raw', 'water_temp_raw'], freq='15min')
#
# qc = qc.correctDrift('sac254_raw', target= 'sac254_corrected', maintenance_field='maint', m
#
#
# plt.plot(qc.data['sac254_raw']['2016'], alpha=.5, color='black', label='original')
# plt.plot(qc.data['sac254_corrected']['2016'], color='black', label='corrected')
# plt.show()
```

SaQC eigen data

```
# import saqc
# import pandas as pd
# import matplotlib.pyplot as plt
#
# data_own = r.df_lab_cor_wide_saqc
# maint_own = r.oh_saqc
# data_own = data_own.set_index('datetime')
# maint_own = maint_own.set_index('datum_start')
# qc_own = saqc.SaQC([data_own, maint_own])
#
# qc_own = qc_own.flagManual('NO3', mdata='maint', method='closed', label='Maintenance')
# qc_own = qc_own.flagRange('NO3', min = 0.0001, max = 50)
# # qc_own.plot('NO3')
#
# qc_own = qc_own.align(['NO3'], freq = '10min')
#
```

```
# qc_own = qc_own.correctDrift('NO3', target='NO3_corrected', maintenance_field='maint', mode
#
# plt.plot(qc_own.data['NO3'], alpha=.5, color='grey', label='original')
# plt.show()
# plt.plot(qc_own.data['NO3_corrected'], color='red', label='corrected')
# plt.show()
```