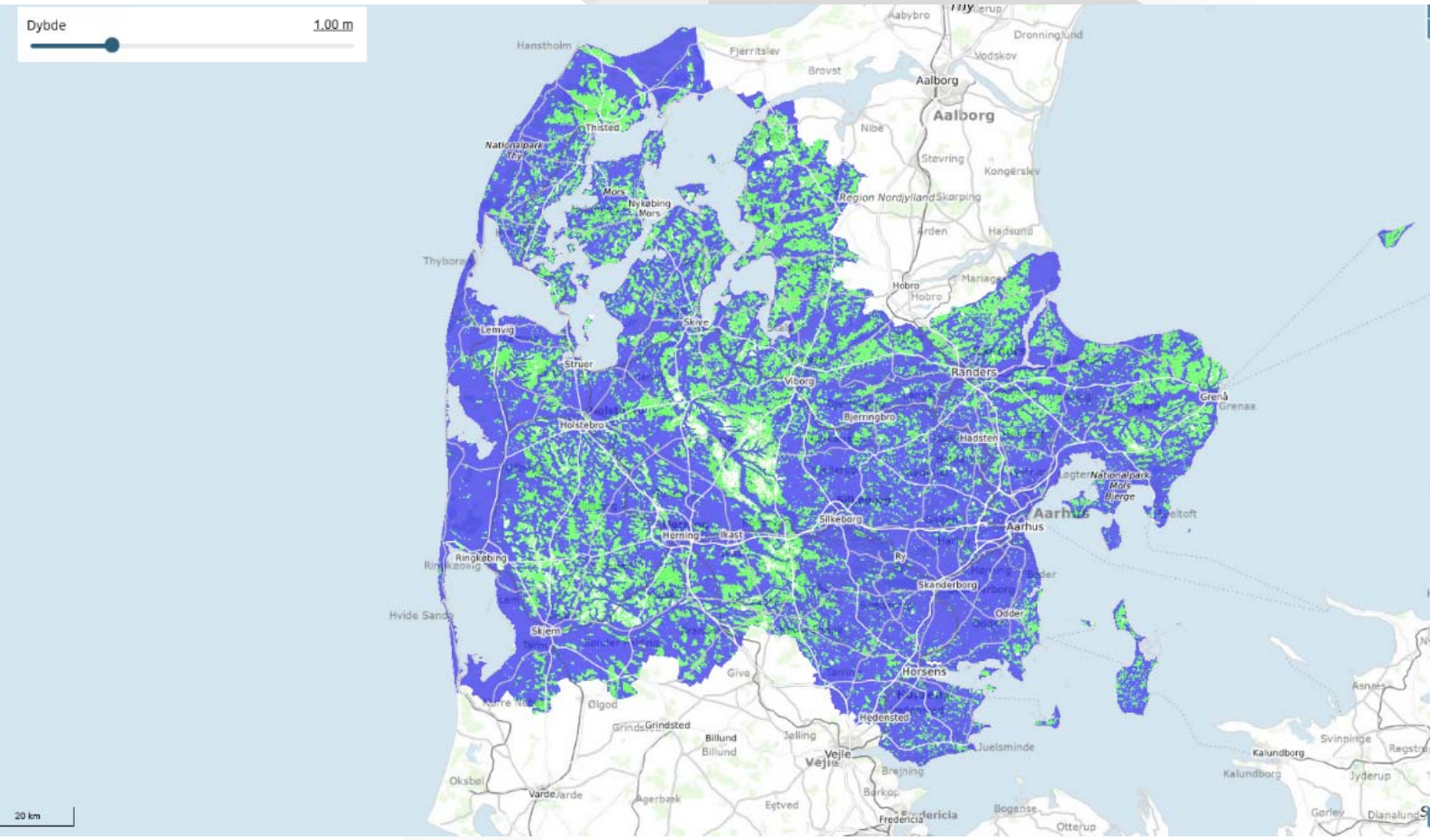
FEBRUAR 2019
REGION MIDT, THISTED, MORSØ OG VESTHIMMERLAND KOMMUNER

PLANLÆGNINGSVÆRKTØJ TIL TERRÆNNÆRT GRUNDVAND

BESKRIVELSE OG DOKUMENTATION



GEUS

SCALGO

Coast to Coast Climate Challenge

COWI

ADRESSE COWI A/S
Parallelvej 2
2800 Kongens Lyngby

TLF +45 56 40 00 00
FAX +45 56 40 99 99
WWW cowi.dk

FEBRUAR 2019
C2C CC / REGION MIDTJYLLAND, THISTED, MORSØ OG VESTHIMMERLAND KOMMUNER

PLANLÆGNINGSVÆRKTØJ TIL TERRÆNNÆRT GRUNDVAND

BESKRIVELSE OG DOKUMENTATION

PROJEKTNR.	DOKUMENTNR.
A108785	A108785-001-01

VERSION	UDGIVELSESDATO	BESKRIVELSE	UDARBEJDET	KONTROLLERET	GODKENDT
0.1	04.02.2019		COWI/GEUS/ SCALGO	CAFK	CAFK

INDHOLD

1 Indledning og formål

I Coast to Coast Climate Challenge regi er udviklet et planlægningsværktøj til visualisering og registrering af terrænnært grundvand i Region Midtjylland samt Thisted, Morsø og Vesthimmerland kommuner.

Formålet med værktøjet er anvendelse i forbindelse med planlægning af byudvikling og klimatilpasning.

Værktøjet kan på planlægnings- og screeningsniveau bl.a.:

- Skabe indsigt og overblik over, hvor i et projektområde der er eller kan opstå problemer med det terrænnære grundvand (vinter max. situation).
- Give konkret og stedspecifikt svar på, hvor dybden til det terrænnære grundvand står mindre end 1 m under terræn (ca. 1 gang om året typisk for en vinter max. situation), og angive en estimeret usikkerhed på beregningen.
- Give et bud på det fremtidige grundvandsspejl for det terrænnære grundvand, under hensyntagen til de forventede ændringer i nedbør og temperatur i et varmere og vådere klima med en tidshorisont på ca. 50 år (2021-2050 i forhold til 1961-1990).
- Bruges til at indføre nye vandstandspejlinger i det interpolerede potentialekort og lokalt opkvalificere vidensniveauet for et givet projektområde.

I nærværende rapport beskrives den udviklede applikation, som er integrereret i SCALGO Live samt baggrunden og datagrundlaget for de beregnede typiske høje terrænnære grundvandsstande.

Formålet med nærværende dokument er at beskrive og dokumentere det udviklede planlægningsværktøj, herunder det indsamlede og processerede datagrundlag, der er anvendt, samt beregningsmetoden Machine Learning/Random Forest, som er anvendt til beregning af den typisk høje terrænnære grundvandsstand.

Manualer og forklaringer til planlægningsværktøjet ligger online indbygget i selve applikationen i SCALGO Live.

English summary
The purpose of the developed planning tool is to be able to visualize the typically high groundwater level for use in connection with urban development and climate adaptation.
It has been chosen to calculate a typically high groundwater level. It is very limited what is available of longer time series for groundwater levels close to the ground. The analysis is based on all available measurements of the groundwater level from boreholes. The starting point is data from the period 1998-2017, which represents a climate period without prolonged periods with very dry years. The selected period is considered to be representative of the current conditions.

The typical high water levels are projected to 2050 based on three climate scenarios (wet, medium and dry) calculated with the DK model. Based on the existing high groundwater levels, model calculations have been performed in a 50m * 50m grid with a calculation method called Machine Learning / Random Forrest. The method uses measured water levels as well as a large number of significant parameters as model input. The model calculations have been performed for the current situation and for 2050 on the basis of the latest nationwide data for projected groundwater levels, cf. The A1B scenario. The projection is therefore a rough estimate of the development approx. 50 years ahead.

In the developed application, the typically high groundwater level near the ground is visualized in the form of the selected point measurements as well as the model-calculated water level maps as both raster and isolation lines. With a slider it is possible to show areas with groundwater level closer to terrain than a given depth to the groundwater. Data can also be extracted and used in a special GIS environment.

It is also possible to record longitudinal profiles, where terrain, calculated groundwater levels and boreholes with measurement data are displayed on the same profile.

Finally, a workspace function has been developed, where it is possible to load new data as well as delete existing basic data and to calculate updated potential maps for the typically high groundwater level.
If new bearing observations have been collected or errors have been found in bearing observations and / or corrections to drilling information, these must be reported to Jupitor. This new or corrected data can then be included in later updates of the data base.
In the following, the application itself, the data basis, the background for the calculated typically high groundwater levels as well as the model calculations of the near-field potential map with Machine Learning based on terrain data and explained variables are described.

2 Baggrund og resumé

Formålet med det udviklede planlægningsværktøj er at kunne visualisere den typisk høje terrænnære grundvandsstand til brug i forbindelse med byudvikling og klimatilpasning.

Der er valgt at beregne en typisk høj grundvandsstand (svarende til en vintervandstand der forekommer ca. en gang om året). Det er meget begrænset, hvad der foreligger af længere tidsserier for terrænnær grundvandsstand. Der er i analysen taget udgangspunkt i alle tilgængelige målinger af grundvandsstanden fra boringer filtersat med indtag ned til 10 m.u.t. Ud fra disse er estimeret en typisk høj vintervandstand. Der er taget udgangspunkt i data for perioden 1998-2017, der repræsenterer en klimaperiode uden længerevarende perioder med meget tørre år, idet de tørre år (herunder 1992-1997 og den ekstremt tørre periode 2018-vinter 2019) ikke er indeholdt i perioden. Den valgte periode vurderes at være repræsentativ for de nuværende forhold.

De typiske høje vandstande er fremskrevet til 2050 ud fra tre klimascenarier (våd, median og tør) beregnet med DK-modellen (jf. beregnede klimagenererede ændringer se værktøjet Grundvandskort: https://www.klimatilpasning.dk/vaerktoejer/grundvand/).

Ud fra de estimerede høje terrænnære grundvandsstande er udført modelberegninger i 50 m x 50 m grid med en beregningsmetode kaldet Machine Learning/Random Forest. Metoden anvender målte vandstande samt en lang række betydende parametre/variabler som modelinput. Modelberegningerne er udført for den nuværende situation (baseret på data for perioden 1998 t.o.m. 2017) og for 2050 på baggrund af de seneste landsdækkende data for fremskrevne grundvandsvandstande (våd, median og tør) jf. A1B scenariet (2021-50 i forhold til 1961-90). Fremskrivning er derfor et groft skøn over udviklingen ca. 50 år frem.

I den udviklede applikation visualiseres den typisk høje terrænnære grundvandsstand i form af de udvalgte punktmålinger/det beregnede datagrundlag samt de modelberegnede vandstandskort som både raster og isolinjer. Med en slider er det muligt at vise områder med grundvandsstand tættere på terræn

end en given dybde til grundvandet, f.eks. områder med vandstand mindre end 1 m.u.t. Data kan desuden udtrækkes og anvendes i et sædvanligt GIS miljø.

Det er endvidere muligt at optegne længdeprofiler, hvor terræn, beregnede grundvandsstande og boringer med måledata vises på samme profil.

Endelig er der udviklet en workspace funktion, hvor det er muligt at indlæse nye data samt slette eksisterende basisdata (målte pejledata og/eller skønnede vandstande ud fra vandstande i overfladevand) samt beregne opdaterede potentialekort for det typisk høje terrænnære grundvandsspejl.

Hvis der er indsamlet nye pejleobservationer eller er konstateret fejl i pejleobservationer og/eller rettelser til boringsoplysning, skal disse indberettes til Jupiter. Disse nye eller rettede data kan herefter indgå i senere opdateringer af datagrundlaget, næste gang COWI/GEUS/SCALGO i samarbejde genererer nye vandstandskort (f.eks. en gang årligt).

I det følgende beskrives selve applikationen, datagrundlaget, baggrunden for de beregnede typisk høje vandstande samt de udførte modelberegninger af det terrænnære potentialekort med Machine Learning på basis af træningsdata og forklarende variable.

3 Beskrivelse af applikationen i SCALGO

Applikationen til præsentation af dybden til den høje terrænnære grundvandvandsstand er udviklet i SCALGO Live og dækker Region Midtjylland samt Thisted, Morsø og Vesthimmerland kommuner.

Applikationen har følgende funktionaliteter:

- Dybden til den typisk høje, terrænnære grundvandsstand kan visualiseres med dynamiske konturer og ækvidistancer. Disse temaer kan vises for de nuværende vandstandsforhold samt for klimascenarierne tør, median og våd for 2050.

- Med en "slider" er det muligt at visualisere de områder, hvor afstanden til det terrænnære grundvand er mindre end en given værdi, se Figur 3-1.

- I applikationen vises datagrundlaget dvs. oplysninger om de pejleobservationer, der er en del af grundlaget for modelberegningerne. Ved klik på et punkt vises nøgleinformationer om punktet i et pop-up vindue, se Figur 3-2.

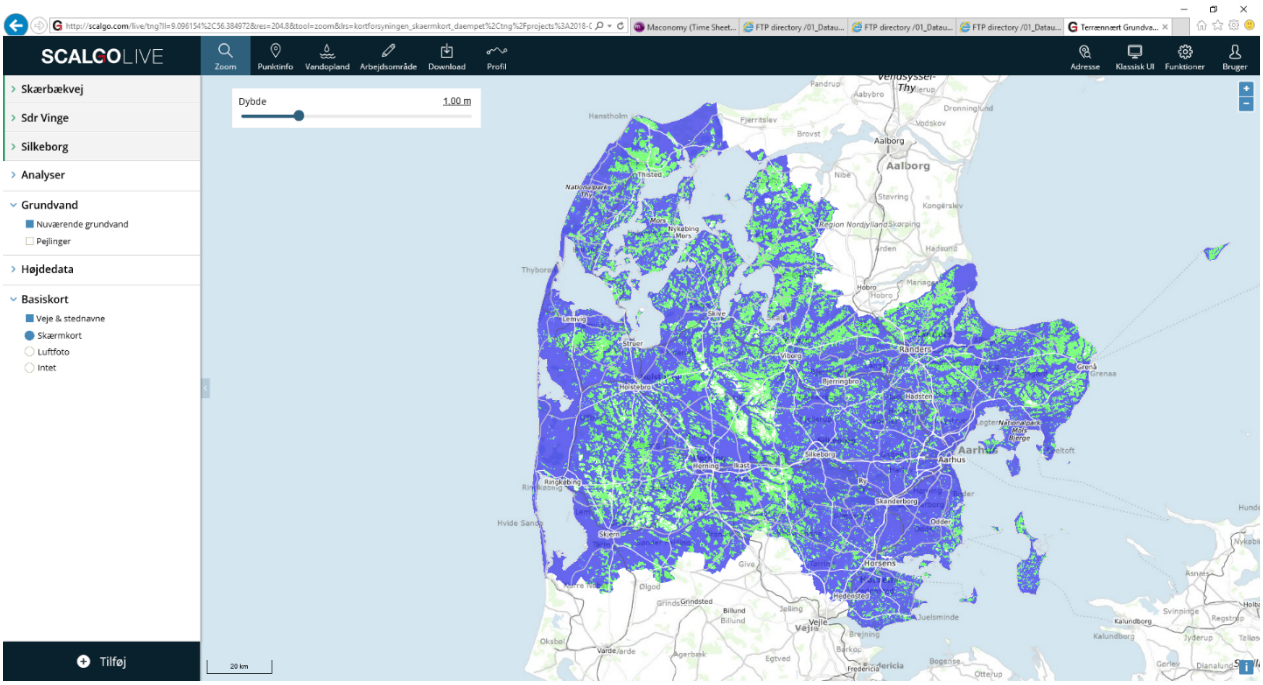
- Brugeren kan få vist et vilkårligt profil i landskabet, med angivelse af terrændata fra Danmarks højdemodel, niveauet for det typisk høje terrænnære grundvandsspejl, usikkerheden på beregningerne samt nærliggende boringer med målt vandstand eller trykniveau se Figur 3-2. Det er desuden muligt at se de rå baggrundsfiler i 50 m opløsning i profilet frem for de justerede data.

- Der kan i applikationen arbejdes med workspaces. Her kan indarbejdes lokale pejleobservationer. Brugeren kan selv slette og tilføje vandstandsværdier og udføre en ny interpolation og konturering af data se Figur 3-2.

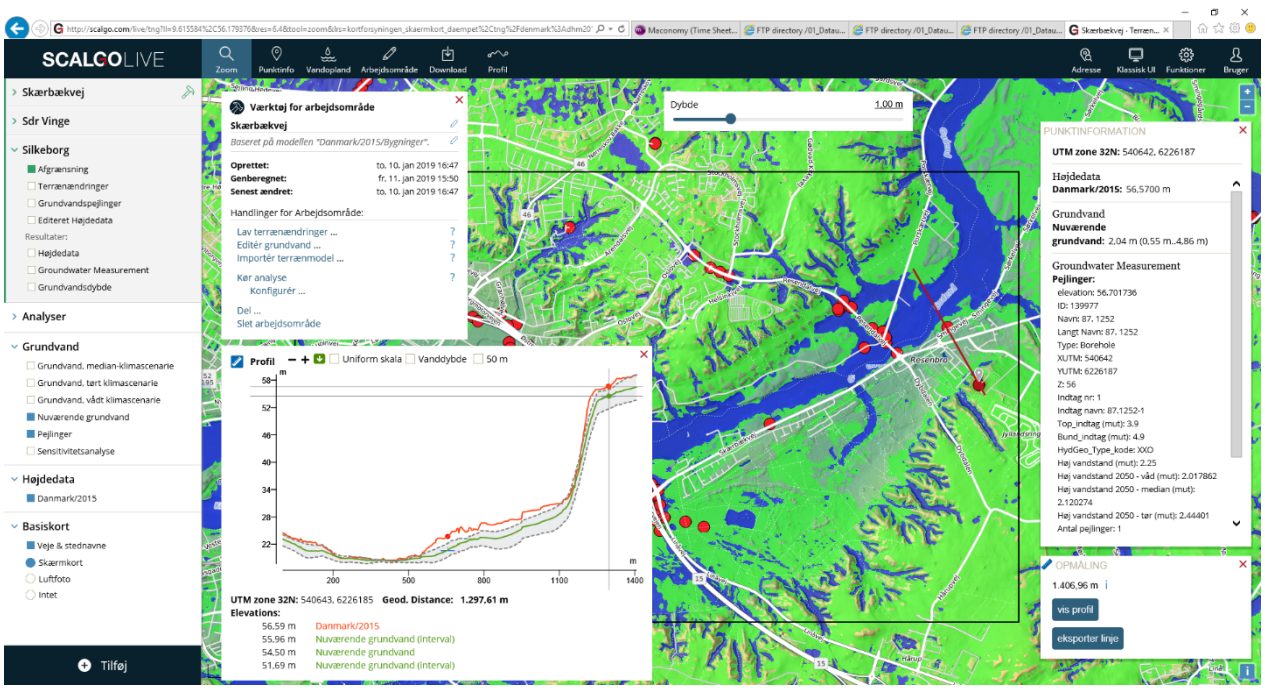
- Det er muligt at se, hvilke variable der har størst forklaringsgrad for fastsættelse af dybden til det terrænnære grundvandsspejl (sensitivitetsanalyse), se Figur 3-3.

- Brugeren kan importere temakort/GIS-data i eget workspace samt se de temaer, der allerede er en del af SCALGO Live (bl.a. offentligt tilgængelige data fra Geodatastyrelsen).

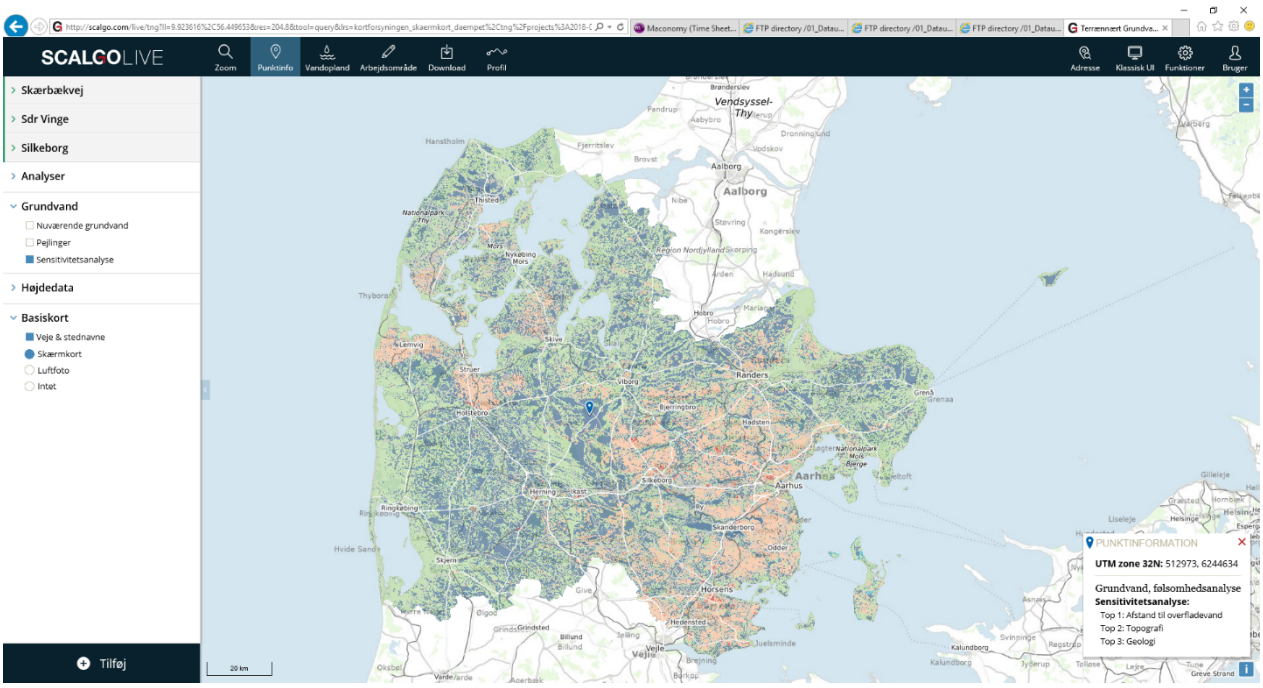
- Brugeren kan udtrække data til gængse GIS miljøer (ArcGIS, MapInfo).



Figur 3-1 *Områder hvor den typiske terrænnære grundvandsstand står højere end 1 m.u.t. vist med blåt. De grønne områder er usikkerheden på de beregnede områder*



Figur 3-2 *Områder hvor den typiske terrænnære grundvandsstand står højere end 1 m.u.t. vist med blåt. De lysegrønne områder er usikkerheden på de beregnede områder*



Figur 3-3 *Information om hvilke forklarende variable der er vigtigst for bestemmelse af det terrænnære grundvand i en given 50 m rastercelle*

SCALGO har nedskaleret beregningerne fra Machine Learning fra 50 m x 50 m grid til 40 cm grid, så vandstanden kan vises sammen med 40 cm's højdemodellen.

Baseret på højdemodellen i 40 cm's grid er beregnet en højdemodel i 50 m's grid ved at tage gennemsnitshøjden i hver 50 m celle. Denne højdemodel er sammen med en række beskrivende variable brugt i forbindelse med beregningen af grundvandsdybden med Machine Learning.

Ud fra de beregnede grundvandsdybder og 50 m's højdemodellen er grundvandskoten beregnet i 50 m opløsning ved at trække grundvandsdybden fra terrænkoten. Grundvandskoten er herefter udvidet til en blød overflade vha. Natural Neighbor Interpolation (NNI) og gemt som en raster i 40 cm opløsning.

For at tage højde for, at grundvandsspejlet delvist følger terrænet, er der foretaget en justering af den højopløselige grundvandskote. Terrænmodellen i 50 m's opløsning er udvidet til en blød overflade vha. NNI. Hvis der i en given 40 cm rastercelle er en forskel på X m imellem den tilsvarende celle i 40 cm's højdemodellen og den bløde udvidelse af 50 m terrænmodellen, så justeres den højopløselige version af grundvandskoten med C * X m i dette punkt. Der er her valgt værdien C = 0,2. Dette valg er en middelvej mellem at vælge den samme grundvandskote for alle celler i DHM (C=0) og vælge den samme grundvandsdybde for alle celler i DHM (C=1).

Yderligere er der i beregningen af den endelige grundvandsdybde og grundvandskote tillagt et fejlgrid (fejlkriging), som svarer til fejlen imellem de beregnede værdier med Machine Learning og pejleobservationerne. Dette betyder, de beregnede vandstande er justeret, så de passer med vandstanden i de enkelte pejleobservationer.

Det skal bemærkes, at den viste usikkerhed afspejler usikkerheden på modelberegningerne og ikke usikkerheden i datagrundlaget, herunder usikkerheden på de estimerede høje vandstande for pejleobservationerne.

Modelberegningernes følsomhed over for en given parameter er vist i et sensitivitetstemakort. Her er de enkelte parametre grupperet i følsomhed over for parametre relateret til følgende:

- Geologi
- Topografi
- Arealanvendelse
- Koordinater
- Nedbør
- Afstand til overfladevand

I SCALGO vises de tre mest betydende parametre for hver gridcelle.

Baggrundsfilerne, som er anvendt i forbindelse med udviklingen af SCALCO applikationen er listet i Tabel 3-1~~Tabel 3-1~~.

Tabel 3-1 Datafiler som grundlag for applikationen

Filnavn	Type	Beskrivelse
RF_training.txt (i alt 16786 punkter)	Punktfil (txt)	Indeholder datagrundlaget i form af processerede høje terrænnære grundvandsstande samt 1900 vilkårlige støttepunkter for søer, vandløb og kyst. Nuværende situation og for 2050 (dry, median, wet).
RF_out (present, dry, median, wet)	Rastergrid (tif)	Indeholder 'best estimate' for den høje terrænnære grundvandsstand i 50x50 m. Nuværende situation og for 2050 (dry, median, wet).
QRF_High1std	Rastergrid (tif)	Indeholder 'upper estimate' for den høje terrænnære grundvandsstand i 50x50 m, svarende til + 1 x standardafvigelse
QRF_Low1std	Rastergrid (tif)	Indeholder 'lower estimate' for den høje terrænnære grundvandsstand i 50x50 m, svarende til - 1 x standardafvigelse
QRF_median	Rastergrid (tif)	Medianestimat på den høje terrænnære grundvandsstand i 50x50 m. Nuværende situation (bruges ved beregning af usikkerhedsbånd)
Sensi_map (top1, top2 og top3)	Rastergrid (tif)	Filerne beskriver de 3 mest betydende parametre for hver 50x50 m gridcelle
DHM2015	Rastergrid (tif)	Danmarks Højdemodel (DHM) i 40 cm opløsning
DHM2015-50 m	Rastergrid (tif)	Danmarks Højdemodel (DHM) i 50 m opløsning
Domain_bdr.shp	ArcGIS shp-fil	Modelområde

4 Datagrundlag og database

4.1 Datagrundlag

Som datagrundlag er anvendt data fra Jupiter, Region Midts GeoGIS database samt data indhentet fra kommuner, forsyninger samt diverse anlægsprojekter og miljøsager. Alle de indsamlede data er indlæst og struktureret i en samlet SQL-database.

Der er udvalgt boringer med indtag ned til 10 m og data fra perioden 1998 t.o.m. 2017.

For en overordnet kvalitetssikring af data, er der sket en frasortering af boringer, hvor den beregnede middelvandstand ligger under bund af indtag. Yderligere er data frasorteret, hvis vandstanden ligger mere end 5 m over terræn, hvis standardafvigelsen for en tidsserie er større end 3, og hvis pejlingerne er registreret som foretaget for en boring i drift (aktive pumpe/indvindingsboringer).

Efter ovenstående udvælgelse/kvalitetssikringsprocedure, omfatter datagrundlaget 14.916 boringer med vandstandsmålinger. Dette svarer til ca. 1 pkt. pr. km². Ud af disse foreligger 392 tidsserier; her defineret som boringer med mere end 5 pejlinger.

Til beskrivelse af vandstanden langs kysten, i søer og vandløb er udtrukket støttepunkter, hvor vandstanden er sat til 0 m.u.t. Det er i forbindelse med testkørsler med Machine Learning vurderet, at ca. 1900 støttepunkter giver den bedste fordeling imellem pejledata og støttepunkter i overfladevand, og dermed de bedste modelresultater.

Yderligere er foreliggende geofysiske undersøgelser for området gennemgået, for vurdering af om disse kan sige noget om beliggenheden af det terrænnære grundvandsspejl. Det blev dog vurderet, at disse data var for usikre til at kunne indgå i datagrundlaget.

4.2 Estimering af høj terrænnær grundvandsstand

På baggrund af de indhentede data er der estimeret en typisk høj terrænnær grundvandsstand på boringsniveau.

For boringer med tidsserier med mere end 5 målinger, er der som et estimat af den typisk høje vandstand anvendt den målte maksimalvandstand i dataserien. For boringer med 5 målinger eller mindre er der udviklet modeller til beskrivelse af en typisk høj vandstand ud fra en fast årstidsvariation (sinuskurver).

De udviklede modeller svarer til sinuskurver med minimum den 15. august og maksimum den 15. februar. Kurvernes amplitude er fastsat ud fra en bestemmelse af indtagets hydrogeologiske typologi, idet der typisk er forskel på hvordan vandstanden varierer afhængig af om en boring er filtersat i eksempelvis sand eller ler, om magasinet er frit, spændt eller artesisk og om magasinet er under indflydelse af nærliggende overfladevand.

Den anvendte hydrogeologiske typologier fremgår af Tabel 4-1.

Tabel 4-1 *Anvendt hydrogeologiske typologier*

Kategori	Beskrivelse	Kode
Hydraulisk permabilitet/ ledningsevne	Højpermeabel	H
	Lavpermeabel	L
	Ukendt	X
Magasinforhold	Frit	P
	Spændt/artesisk	A
	Ukendt	X
Nærhed til overfladevand	Kystnær	C
	Vandløbsnær	S
	Andet	O

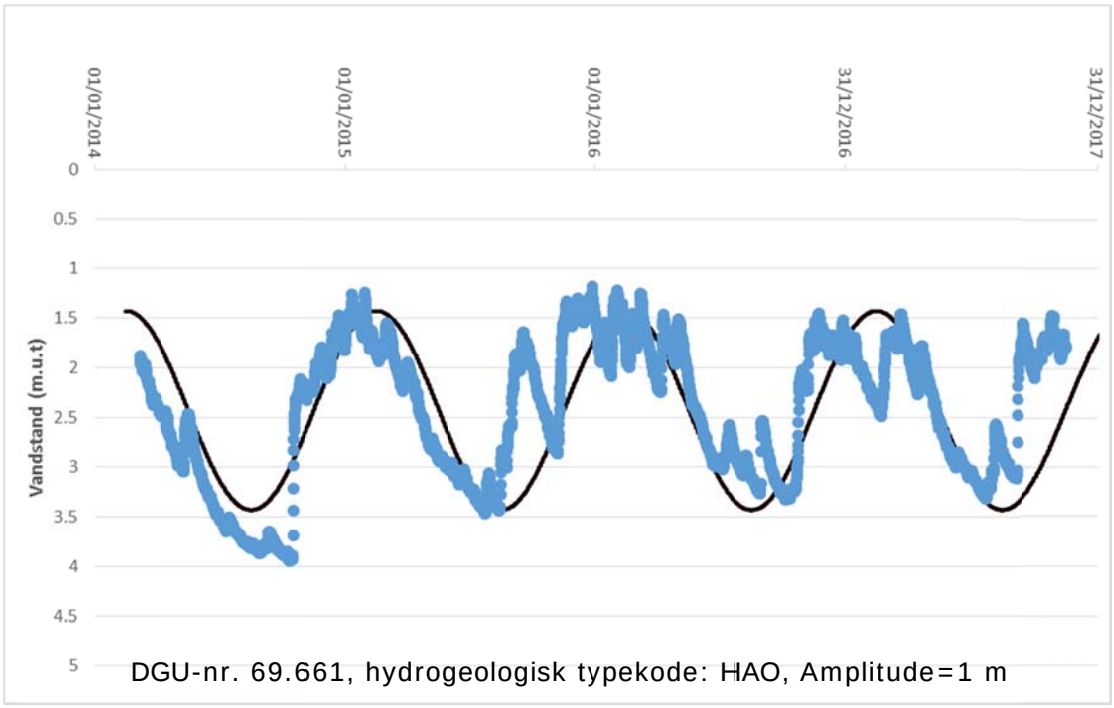
Vandløbs- og kystnære boringer, er defineret som boringer, der ligger inden for 100 m fra kysten, store søer samt vandløb.

Ovennævnte typologi giver anledning til eksempelvis typen HPC, som er en boring filtersat i en højpermeabel aflejring, i et frit magasin og mindre end 100 m fra kysten.

Boringer filtersat i lavpermeable aflejringer er dog per. definition af typen fri (P).

Til bestemmelse af typiske sæsonvariationer (amplituder) er der for de 392 tidsserier beregnet gennemsnitlige standardafvigelser for de forskellige hydrogeologiske typologier. Ud fra disse er der beregnet 99 %'s konfidensintervaller og foretaget en vurdering af typiske amplituder.

Af Tabel 4-2 fremgår det, at amplituderne er vurderet til i middel at være 0,5 til 1,5 m, svarende til variationer i vandstand på 1 til 3 m. Der ses for de terrænnære magasiner, at amplituden er størst i lavpermeable aflejringer samt langt fra overfladevand, da typerne HPC, HPS, HAC og HAS har en gennemsnitlig amplitude på 0,5 m, HPO samt HAO en gennemsnitlig amplitude på 1 m, og LPO en gennemsnitlig amplitude på 1,5 m. Et eksempel på hvordan en sinuskurve kan beskrive sæsonvariationen i vandstandsdata er givet i Figur 4-1.



Figur 4-1 *Et eksempel på hvordan en sinuskurve kan beskrive sæsonvariationen i vandstandsdata*

Det skal bemærkes, at for de dybere filtre, ses ikke i gennemsnit en mindre amplitude tæt på overfladevand.

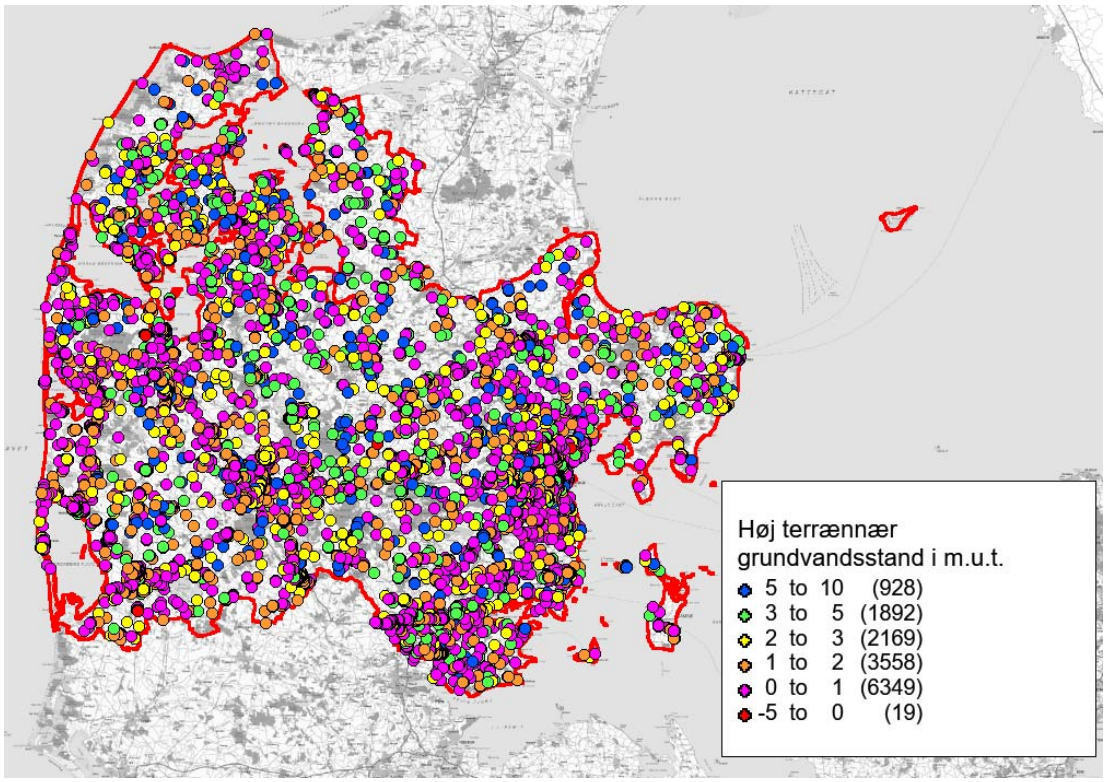
Ved fremskrivning af de målte grundvandsstande til en typisk høj grundvandsstand lander ca. 3000 vandstande over terræn. Disse er dog efterfølgende sat til 0 m.u.t., da dette vurderes som mest sandsynligt/repræsentativt, bl.a. fordi mange af de øvre magasiner er frie (ikke under tryk).

Tabel 4-2 *Vurderet sæsonvariation baseret på tidsserier med mere end 5 pejlinger*

Type	Antal	Stdev (m)	Stdev*2,576 (m)	Oversat type	Vurderet amplitude (m)
HPC	12	0,11	0,3		0,5
HPS	10	0,19	0,5		0,5
HPO	150	0,40	1,0		1
HAC	1	0,11	0,3		0,5
HAS	10	0,26	0,7		0,5
HAO	111	0,40	1,0		1
HXC	0		0,0	HPC	0,5
HXS	0		0,0	HPS	0,5
HXO	0		0,0	HPO	1
LPC	0		0,0		0,5
LPS	2	0,35	0,9		0,5
LPO	15	0,61	1,6		1,5
LXC	0		0,0	LPC	0,5
LXS	0		0,0	LPS	0,5
LXO	0		0,0	LPO	1,5
XPC	0		0,0	HPC	0,5
XPS	0		0,0	HPS	0,5

XPO	2	0,34	0,9	HPO	1
XAC	0		0,0	HAC	0,5
XAS	0		0,0	HAS	0,5
XAO	0		0,0	HAO	1
XXC	0		0,0	HPC	0,5
XXS	7	0,34	0,9	HPS	0,5
XXO	72	0,32	0,8	HPO	1

Det endelige datagrundlag med de vurderede høje terrænnære grundvandstande i m.u.t. fremgår af Figur 4-2, hvor boringerne er kategoriseret efter dybden til grundvandsspejlet.



Figur 4-2 *Det endelige datagrundlag med de vurderede typiske høje terrænnære grundvandstande i m.u.t. kategoriseret efter dybde til grundvandsspejlet*

5 Beregninger med Machine Learning

5.1 Forklarende variable

Machine Learning (ML) er en teknik og faglig disciplin, der anvendes inden for datalogi og anvendt statistik. ML er datadrevet, dvs. at ML trænes til at finde regler og sammenhænge baseret på eksisterende pejleobservationer, vandstande, kort og viden (f.eks. antagelser om repræsentativitet, støttepunkter mv.).

Ved analyser af eksisterende data kan ML analysere og lære mønstre om, hvordan de terrænnære grundvandstande hænger sammen med andre, såkaldt forklarende variabler. Dermed kan dybden til det terrænnært grundvand estimeres ud fra træningsdata og kendskab til forklarende variabler i forskellige områder, hvor der er få eller ingen boringer med information om trykniveauer. ???? viser de 26 forklarende variabler, som er vurderet til at være de mest relevante i forholdt til dybden til det terrænnære grundvand, og som er benyttet til at træne ML modellen til at finde mønstre i de ~15.000 boringer og 1900 støttepunkter.

Tabel 5-1 *Oversigt over de 26 forklarende variabler brugt til at opbygge Random Forest modellen. Den oprindelig rumlige opløsning varierer imellem lagene. Alle er dog skalerede til det samme 50 m grid.*

Variable	Gruppe
Ler indhold - a horisont	Geologi
Ler indhold - b horisont	
Ler indhold - c horisont	
Ler indhold - d horisont	
Kvartær lagtykkelse	
Top ler tykkelse	
Dræn sandsynlighed	
Dræn klasser	
Lavbund klassifikation	
Landskabstypologi	
Georegion	
Jordtype	
Højdemodel	Topografi
Højdemodel detrend	
Topographic Wetness Indeks	
Saga Wetness Indeks	
Opstrøms areal	
Hældning	
vertikal afstand til vandløb	
vertikal afstand til vandløb	Afstand til overfladvand
horisontal afstand til vandløb	
sø, vandløb, kyst klassifikation	
Nedbør	Nedbør
Befæstelsesgrad	Arealanvendelse
Arealanvendelse	
Koordinater (utmx)	Koordinater
Koordinater (utmy)	

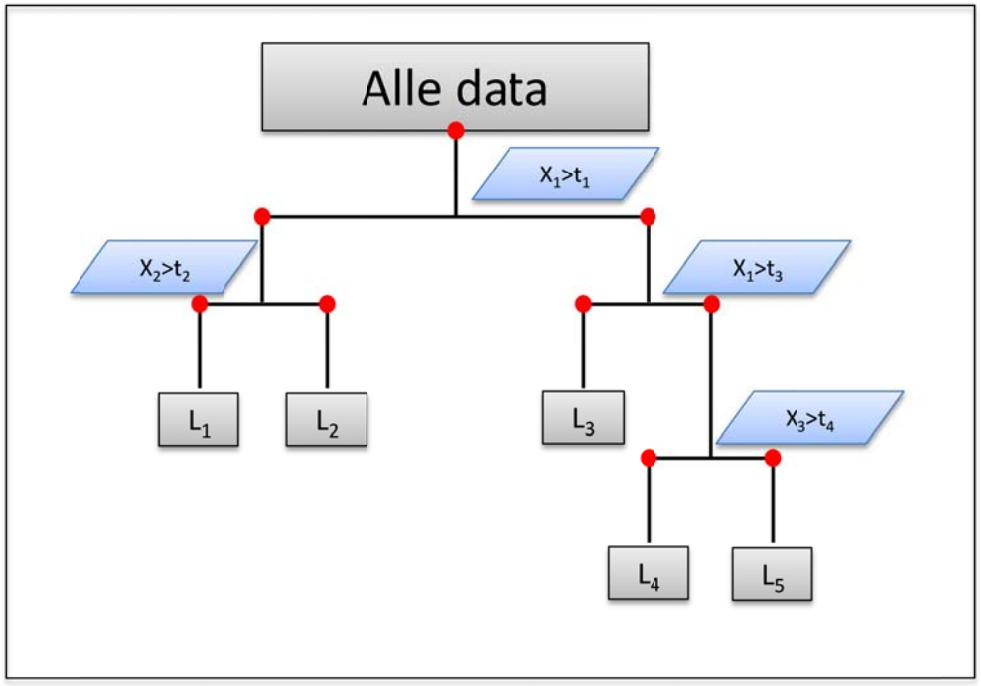
5.2 Random Forest

Der er anvendt en ML metode, som kaldes Random Forest (RF). RF er baseret på en automatiseret opbygning af et stort antal beslutningstræer, som vist i Figur 5-1~~Figur 5-1~~.

Teknikken med generering af beslutningstræer har stor fleksibilitet og kan fange relativt komplekse sammenhænge i store datasæt med en maksimal udnyttelse af informationsværdien i data. I første omgang kan træningsdata f.eks. opdeles efter, om der er tale om en lokalitet, der ligger tæt på eller langt fra et vandløb ($X_1 > t_1$, hvor X_1 er den horisontale afstand til vandløb og t_1 er en tærskelværdi, f.eks. 100 m). Efterfølgende kan hver af de resulterende opdelte grupper vide-

reopdeles på basis af viden om f.eks. geologien (og så fremdeles). I RF er hvert beslutningstræ bygget op ud fra en tilfældigt genereret udvælgelse af træningsdatasættet. Hver gang en datagruppe er delt op, benyttes et tilfældigt udvalgt sæt og rækkefølge af valgte forklarende variabler, der benyttes i underopdelingen. Den største nøjagtighed opnås, når samtlige træningsdata fordeles ud i enkelte blade på træet, og når der så at sige genereres en hel skov af træer. Metodikken med at generere en hel skov af træer er anvendt for at undgå overfitting og for at bygge en model, som generelt set er så robust som muligt.

GEUS vurderer, at RF modellen skal benytte som minimum 1000 beslutningstræer for at give et robust estimat af dybden til den typisk høje grundvandsstand. Det endelige RF resultat for et givent grid (i 50 x 50 m) er bestemt ud fra middelværdien af de i alt 1000 genererede træer, som hver for sig giver et muligt estimat.



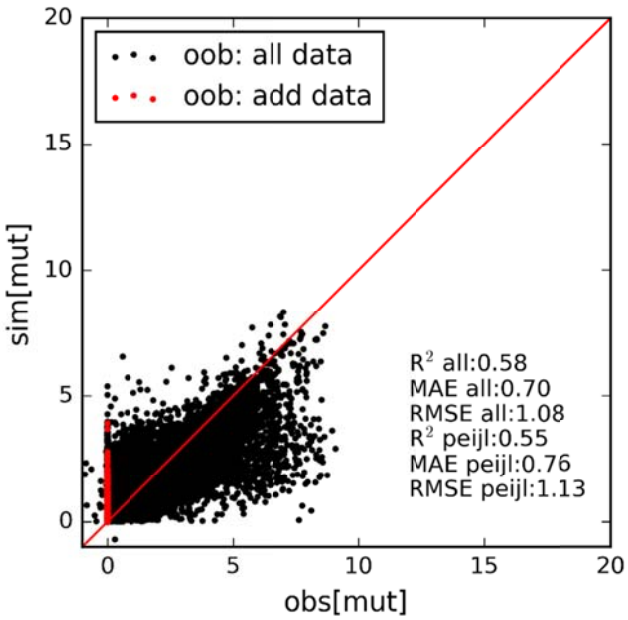
Figur 5-1. Beslutningstræ, hvor træningsdata opdeles på baggrund af en række forklarende variabler.

I RF metoden indgår en indbygget usikkerhedsmetodik. Hver gang der genereres et nyt træ benyttes kun 2/3 af træningsdata, og det betyder at 1/3 af data kan anvendes til validering ud fra uafhængige data. Denne uafhængige valideringstest (out-of-bag: oob) viser, at RF modellen kan beskrive over halvdelen (55 %) af den varians, der findes i træningsdata, se nedenstående Figur 5-2. Den halvdel, som RF ikke kan forklare, skyldes bl.a. små-skala variabilitet, som ikke kan fanges i 50 m opløsning, usikkerheder på de forklarende variabler samt usikkerheder på fremskrivning af den høje vandstand med sinus modellen, samt målefejlen på pejlinger.

Der er ved metoden opnået en middel absolut fejl (MAE) på modellen svarende til 76 cm alene vurderet ud fra pejleobservationer, hvilket er meget tilfredsstillende. Den samlede fejl på det RF genererede potentialekortet over typisk høj

grundvandsstand vil være større, da der er en usikkerhed på sinuskorrektionen, der ikke indgår i usikkerhedsestimatet af MAE, ligesom at der kan være usikkerhed på repræsentativitet af støttepunkter, der heller ikke er indregnet.

I Figur 5-3 er vist resultater af RF med beregnet middel absolut fejl beregnet hhv. med og uden støttepunkter. Betydningen af om man tager støttepunkter med i beregningen har ikke nogen væsentlig betydning (MAE ændres fra 70 cm til 76 cm), hvorimod de øvrige usikkerheder ikke umiddelbart kan kvantificeres.



Figur 5-2. RF valideringstest. "peijl" referer til beregning af usikkerheder alene ud fra pejledata fra boringerne hvorimod "all" også inkluderer de i alt 1900 støttepunkter (vist i rødt som punkter der har en observeret dybde på nul).

5.3 Random Forest usikkerhed

I en datadrevet model som RF, vil modellen give en relation mellem et sæt af forklarende variabler og dybden til det terrænnært grundvand.

Der er imidlertid ikke en unik relation mellem en kombination af forklarende variabler og en observeret vandstand. Dvs. den samme kombination af forklarende variabler i 50 m beregningsgrid forskellige steder, kan være associeret til forskellige grundvandstande. Det betyder i praksis, at estimatet fra RF modellen ikke vil være lige nøjagtigt alle steder.

For at kvantificere den usikkerhed, der er på det interpolerede potentialebillede, er der anvendt en metodik, som benævnes Quantile Regression Forest (QRF). QRF er en meget anvendt metode til at bestemme usikkerheden på en RF model. QRF analyserer fordelingen af de mulige estimater på baggrund af de 1000 beslutningstræer, som modellen indeholder. Fordelingen af de 1000 estimater kan tages som udgangspunkt for estimering af usikkerheden. Der er anvendt 16% og 84% fraktilerne til bestemmelse af usikkerheden svarende til +/- 1

standardafvigelse. Usikkerheden vil være lavere i områder hvor RF modellen indeholder mere ensarte træer, hvorimod en stor afvigelse blandt de 1000 beslutningstræer, vil blive indikereret med en høj usikkerhed.

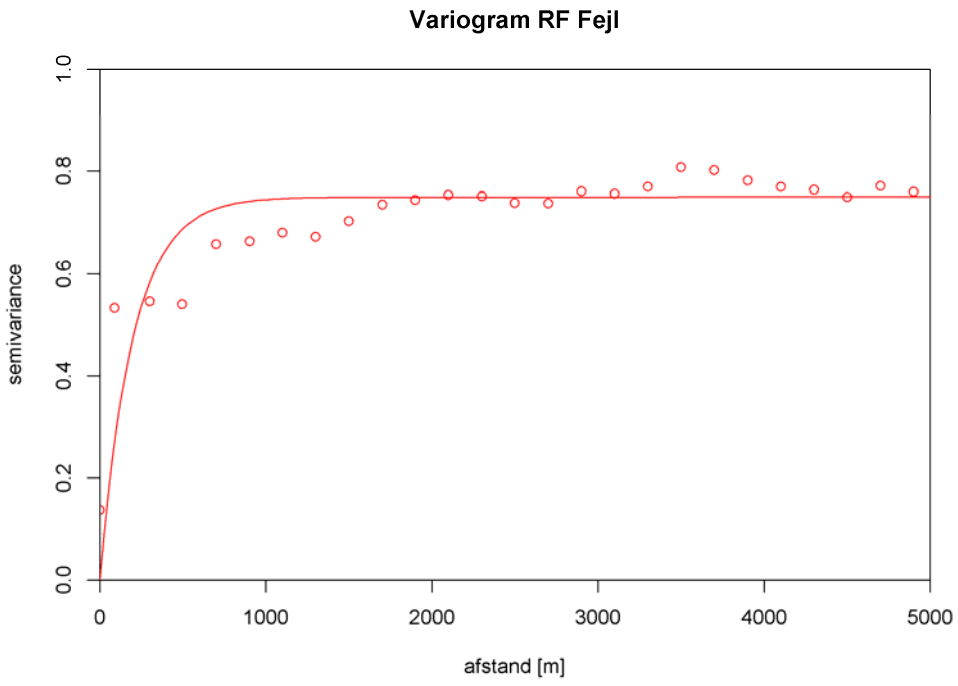
5.4 Fejlkriging

For at honorere de observerede data, kan RF modellen afslutningsvist kombineres med en rumlig model af afvigelsen (residualet) mellem RF estimatet og de observerede dybder til det terrænnære grundvand.

RF modellen benyttes således først til at give et estimat på dybden, herefter sammenlignes RF modellen med observationerne, og der laves en model over den rumlige fordeling af residualerne, der adderes til RF estimatet.

I denne sammenhæng er der anvendt en geostatistisk model som kaldes Kriging, og den samlede metode benævnes derfor Random Forest Residual Kriging (RFRK). Ved anvendelse af Kriging er det muligt at beskrive den rumlige korrelation i residualerne. Kriging af residualerne kan her benyttes til at korrigere RF modellen også for de 50 m beregningsceller, som ikke indeholder pejledata på basis af den anvendte geostatistisk interpolationsmetodik.

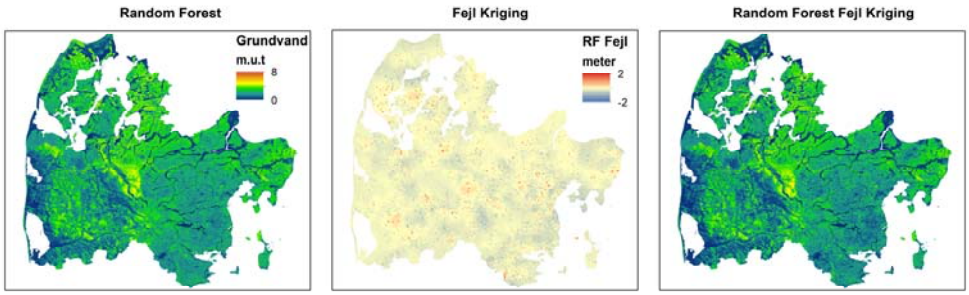
I forbindelse med Kriging etableres et variogram, der beskriver variansen samt den rumlig korrelationen af RF residualerne. Der er anvendt en variogram model, se Figur 5-3, med en korrelationslængde, som definerer den maksimale afstand indenfor hvilken, en observation indeholder relevant information, når det gælder en traditionel interpolation, bestemt til ca. en afstand på 500 m.



Figur 5-3. Variogram af RF fejlen brugt til fejlkriging

Figur 5-4 viser den anvendte RFRK metodik med RF estimatet som den overordnede trend og fejlkriging, til brug for interpolation af RF residualerne. Der kan ses en rumlig korrelation i RF fejlen, hvor ensartede positive eller negative fejl, ligger tættere på hinanden, når afstanden er mindre end 500 m, end ved større afstande.

At der er afvigelser mellem RF modellen og pejleobservationerne, kan både skyldes systematiske fejl i RF modellen, hvor afgørende forklarende variabler mangler, eller eksisterende variabler ikke er detaljeret nok repræsenteret, og kan derudover skyldes evt. usikkerheder på pejleobservationer, herunder usikkerhed på de anvendte sinusmodeller. Efter at fejlen adderes til RF laget, bliver RF estimatet tilpasset til observationerne, som dermed honoreres.



Figur 5-4. Skema over RFRK metodikken hvor RF fejlen adderes til RF estimatet.

5.5 Klimafremskrivning

For at kunne beregne dybden til det terrænnære grundvand i 2050 er anvendt resultaterne fra Klimagrundvandskort rapporten (http://www.klimatilpasning.dk/media/340310/klimagrundvandskort.pdf), som blevet udgivet i kombination med klimatilpasning (KFT).

Der er på Klimagrundvandskort beregnet ændringer med klima, der forudsiger en hhv. stor (våd klimamodel), middel (median klimamodel) og lille (tør klimamodel) ændring fra referenceperioden (1961-1990) til fremtiden (2021-2050) ud fra A1B scenariet og 9 forskellige regionale og globale klimamodeller.

Der er anvendt de tre udvalgte fremskrivninger, som anses for mest realistiske og repræsentative. På basis af den resulterende ændring i det terrænnære grundvandsspejl, blev træningsdatasættet justeret, således at hver observation repræsenterer 2050-forhold. Ændringen mellem referenceperioden og klimafremskrivningen er beregnet for hver pejling. Ændringerne er adderet, og værdierne som resulterer i vand over terræn, er sat til dybden 0 m.u.t. Efterfølgende er Random Forest modellen trænet igen, og kort over dybden til grundvandsspejlet i 2050 er estimeret.

5.6 Følsomhedsanalyse

Der er gennemført en følsomhedsanalyse til bestemmelse af de mest afgørende variabler i RF modellen, for at kunne vise de 3 vigtigste grupper af variabler for hver beregningscelle. Variablerne er grupperet i kategorierne: geologi, topografi, afstand til overfladevand, arealanvendelse, nedbør og koordinater, se Tabel 5-1.

For hver gruppe er RF modellen genberegnet 250 gange, og hver genberegning indeholder en tilfældig ændring i rækkefølgen af variabler, som tilhører den enkelte gruppe. Middelændringen mellem det oprindelig RF estimat og de 250 genberegninger er brugt til at estimere følsomheden. Middelændringer af de 6 grupper er sorteret og efterfølgende brugt til at identificere de 3 vigtigste variabler for hver celle.