NY FOSFORMODEL TIL ESTIMERING AF ÅRLIG VANDFØRINGSVÆGTET KONCENTRATION AF TOTAL FOSFOR FRA DIFFUSE KILDER I ID15-OPLANDE

Teknisk rapport fra DCE – Nationalt Center for Miljø og Energi nr. 246 2022

AARHUS
UNIVERSITET
DCE – NATIONALT CENTER FOR MILJØ OG ENERGI

NY FOSFORMODEL TIL ESTIMERING AF ÅRLIG VANDFØRINGSVÆGTET KONCENTRATION AF TOTAL FOSFOR FRA DIFFUSE KILDER I ID15-OPLANDE

Teknisk rapport fra DCE – Nationalt Center for Miljø og Energi nr. 246 2022

Søren E. Larsen
Ane Kjeldgaard
Jørgen Windolf
Henrik Tornbjerg
Brian Kronvang

Aarhus Universitet, Institut for Ecoscience

AARHUS
UNIVERSITET
DCE – NATIONALT CENTER FOR MILJØ OG ENERGI

Datablad

Serietitel og nummer:	Teknisk rapport fra DCE - Nationalt Center for Miljø og Energi nr. 246
Kategori:	Rådgivningsrapporter
Titel:	Ny fosformodel til estimering af årlig vandføringsvægtet koncentration af total fosfor fra diffuse kilder i ID15-oplande
Forfattere:	Søren E. Larsen, Ane Kjeldgaard, Jørgen Windolf, Henrik Tornbjerg og Brian Kronvang
Institutioner:	Aarhus Universitet, Institut for Ecoscience
Udgiver:	Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi ©
URL:	http://dce.au.dk
Udgivelsesår:	Juni 2022
Redaktion afsluttet:	Juni 2022
Faglig kommentering:	Jonas Rolighed, Ecoscience, Aarhus Universitet
Kvalitetssikring, DCE:	Signe Jung-Madsen og Lars Moeslund Svendsen, DCE, Aarhus Universitet
Sproglig kvalitetssikring:	Anne Mette Poulsen
Ekstern kommentering:	Miljøstyrelsen. Kommentarerne findes her: http://dce2.au.dk/pub/komm/TR246_komm.pdf
Finansiel støtte:	Miljøstyrelsen
Bedes citeret:	Larsen, S.E., Kjeldgaard, A., Windolf, J., Tornbjerg, H. & Kronvang, B. 2022. Ny fosformodel til estimering af årlig vandføringsvægtet koncentration af total fosfor fra diffuse kilder i ID15-oplande. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 80 s. - Teknisk rapport nr. 246 http://dce2.au.dk/pub/TR246.pdf
	Gengivelse tilladt med tydelig kildeangivelse

Sammenfatning: En ny model til estimering af den årlige vandføringsvægtede total fosfor (TP) koncentration i vandløb i ID15-oplande er udviklet på stikprøvedata ved brug af maskinlærings softwaren DataRobot på baggrund af data fra 207 typeoplande. Den nye model er af typen ' eXtreme Gradient Boosted Trees Regressor with early stopping' har en forklaringsgrad (R2) på 0,67-0,71 på træningsdatasættene i DataRobot. Resultater fra et uafhængigt valideringsdatasæt giver en forklaringsgrad (R^2) på 0,62 for modellen og på 0,41 efter tilbagetransformering, hvilket er en klar forbedring ift. den nuværende model, der kun bygger på data fra 24 vandløb med intensiv prøvetagning. Ved skift til den nye model vil den årlige fosfortilførsel til kystvandene i gennemsnit blive omkring 3 % lavere, end opgjort med den nuværende anvendte og biaskorrigerede TP-model.

Emneord:	vandløb, fosformodel, total fosfor, vandføringsvægtet koncentration, ID15-oplande, maskinlæring
Layout:	Karin Balle Madsen
Foto forside:	NOVANA-sensorstation. Fotograf: Brian Kronvang
ISBN:	978-87-7156-697-0
ISSN (elektronisk):	2244-999X
Sideantal:	8080
Internetversion:	Rapporten er tilgængelig i elektronisk format (pdf) som http://dce2.au.dk/pub/TR246.pdf

Indhold

Forord **5**

Sammenfatning **6**

Summary **7**

1 Introduktion **8**

2 Datagrundlaget **11**

3 Anvendte metoder **13**

Trin 1 13
Trin 2 14
Trin 3 14
Trin 4-6 15

4 Resultater **20**

4.1 Trin 1 – validering af åbent land NOVANA TP-modellen 20
4.2 Trin 2 – test af forskellige fosformodeltyper i Odense Å-oplandet 21
4.3 Trin 3 – udvikling af ny TP-model på intensive måledata af TP-koncentrationer 25
4.4 Trin 4 – udvikling af en prototype model i DataRobot 29
4.5 Trin 5 – sammenligning af en udviklet regressionsmodel i SAS contra DataRobot ML-01TP-model 37
4.6 Trin 6 – udvikling af endelig DataRobot ML-06 TP-model 43

5 Konklusion og perspektiver **57**

6 Referencer **60**

Bilag 1: Oversigt over temaer og referencer for inputdata anvendt i de forskellige modeller **62**

Dokumentation, inddata til P-model 62

Bilag 2: Oversigt over mulige modelparametre fra oplandene **64**

Bilag 3: Prototype ML TP-model – betydning af de mest forklarende variable vist i de tværgående blå søjler fra screen dump fra DataRobot **66**

Bilag 4: Oversigt over respons for variable som indgår i ML-01 TP-modellen **68**

Bilag 5: Oversigt over plots fra den endelige ML-06 TP-model i DataRobot **74**

Bilag 6: Sammenligning mellem arealanvendelse og jordtyper for målte og umålte arealer defineret ud fra NOVANA-programmets havbelastnings-stationsnet **78**

Forord

Miljøstyrelsen indgik i 2018 en aftale om et projekt med DCE, Aarhus Universitet, vedrørende en opdatering og modernisering af fosformodellen, som anvendes i NOVANA-programmet (Det Nationale Overvågningsprogram for Vandmiljø og Natur) til beregning af transporten af total fosfor i vandløb fra umålte oplande. I projektet blev der ligeledes gennemført analyser af eksisterende tidsserier af data fra intensivstationer, hvor total fosfor (TP) koncentrationen måles meget hyppigt, samt udvikling af nye sensormetoder til måling af næringsstofkoncentrationer. Dele af projektet er tidligere blevet rapporteret i selvstændige notater fra DCE (Tornbjerg et al., 2019; Van't Veen et al., 2020). I denne rapport beskrives derfor alene arbejdet med moderniseringen af fosformodellen. Det oprindelige formål med projektet var at opdatere den nuværende model med nye kortgrundlag og data. Men efterhånden som projektet skred frem, blev det klart, at man ikke ad den vej kunne nå et tilstrækkeligt godt resultat. Det blev derfor besluttet at forlænge projektet og i stedet udvikle en helt ny model for koncentrationen af TP i vandløb. Der har i projektperioden i fem trin været testet forskellige modeller kørt på forskellige datasæt og med forskellige inputparametre. Nærværende rapport beskriver hele processen frem mod udviklingen af den endelige nye TP-model. I tilknytning til projektet har der været en projektgruppe med deltagelse fra Miljøstyrelsen, hvor resultater fra projektet er blevet diskuteret, herunder parametervalg og modeltyper mv. De endelige faglige vurderinger og beslutning omkring parametre og model er blevet taget af medarbejdere fra DCE, AU.

Sammenfatning

I denne rapport er der i flere trin gennemført statistiske analyser med det formål at udvikle en ny model til estimering af den vandføringsvægtede total fosfor koncentration i vandløb, der afvander typeoplande med et oplandsareal på i gennemsnit ca. 20 km^2. Slutproduktet er en model for årskoncentrationen af total fosfor (TP) udviklet i maskinlæringssoftwaren 'DataRobot'.

Der er i de statistiske analyser arbejdet med at udvikle en model på logaritmetransformerede data. Der indgår i datasættet til udvikling af modellen et datasæt fra i alt 207 vandløbsoplande med 2389 observationer af den vandføringsvægtede årskoncentration af TP som input til krydsvalideringen i DataRobot. I et helt uafhængigt datasæt indgår der data fra i alt 142 vandløbsoplande med i alt 1261 observationer af den vandføringsvægtede årskoncentration af TP.

Den til slut udviklede maskinlæringsmodel er af typen 'eXtreme Gradient Boosted Trees Regressor with early stopping' og indeholder i alt 13 forklarende variable, herunder, som eksempler på de mest betydende variable i modellen, viden om omfanget af dræning i oplandet, befæstet areal i oplandet, dyrkningsgrad i oplandet, omfanget af brinkerosion i oplandet og den årlige nedbørsafvigelse fra et langtidsgennemsnit.

De tre lag i udviklingen af maskinlæringsmodellen i DataRobot har følgende forklaringsgrader: Træningsdatasættet, som består af 64 % af data (R^2 = 0,69), valideringsdatasættet, som består af 16 % af data (R^2 = 0,71), og hold out datasættet med 20 % af data (R^2 = 0,67). Hertil kommer en validering af modellen på det uafhængige datasæt, som har en god forklaringsgrad (R^2) både før (0,62) og efter (0,41) tilbagetransformering. Den nye udviklede TP-model har derfor en langt større forklaringsgrad end den tidligere anvendte TP-model i NOVANA-beregningerne.

Ved skifte fra den nuværende anvendte bias-korrigerede TP-model til den nyudviklede maskinlærings TP-model i denne rapport falder den hidtidige beberegnede TP-tilførsel til kystvande i Danmark med ca. 3 % i perioden 1990-2019. Skiftet til den nye TP-model vil i enkelte år maksimalt ændre tilførslen til 2.-ordens kystafsnit med fald på fra 0,9-6,9 %.

Usikkerheden på den udviklede nye maskinlærings-TP-model er beregnet på baggrund af valideringen af modellen på det helt uafhængige valideringsdatasæt af typeoplande (N=1261) samt kalibreringsdatasættet (N=2389). Root Mean Square Error (RMSE) er beregnet til at være lille og modellen derfor god (<0,2) for langt de fleste georegioner baseret på det uafhængige valideringsdatasæt. Tilsvarende er Mean Absolute Error (MAE) relativ lille for de fleste georegioner (0,003-0,055 mg P / l).

Summary

In this report, statistical analyses were carried out in several steps with the purpose of developing a new model for the simulation of flow-weighted total phosphorus concentrations in streams draining catchments with an average area of approx. 20 km^2. The final product is the presented model predicting the annual concentration of total phosphorus (TP), developed in the machine learning software 'DataRobot'.

The purpose of the statistical analyses was to develop a model based on logarithm-transformed data. The data set used for the development of the model included data on a total of 207 catchments with 2389 observations of the flow-weighted annual TP concentration as input for the DataRobot cross-validation. A completely independent dataset contained data from 142 catchments with 1261 observations of the flow-weighted annual concentration of TP.

The developed machine learning model is of the type 'Xtreme Gradient Boosted Trees Regressor with early stopping'. The model encompasses a total of 13 explanatory variables, including, as the most significant, the extent of drainage in the catchment, the built-up area in the catchment, the level of cultivation in the catchment, the extent of the ground erosion in the catchment and the annual deviation in precipitation from a long-term average.

The three layers in the development of the machine learning model in DataRobot had the following degrees of explanation: The training data set included 64% of the data (R^2 = 0.69), the validation data set included 16% of the data (R^2 = 0.71), and the holdout data set included 20% of the data (R^2 = 0.67). In addition, the model was validated against the independent dataset, which had a good degree of explanation (R^2) both before (0.62) and after (0.41) retransformation. This is a much higher degree of explanation than in the formerly used TP-model in NOVANA.

When switching from the current bias-corrected TP model to the newly developed machine-learning TP model presented in this report, there was an average decrease in the TP supply to coastal waters in Denmark of approx. 3% in the period 1990-2019. The switch to the new TP model will, in some years, change the input to second order coast sections with a maximum decrease from 0.9 to 6.9%.

The uncertainty of the developed new machine learning TP model was calculated on the basis of the validation of the model against the completely independent validation data set on type catchments (N = 1261) as well as the calibration data set (N = 2389). Root Mean Square Error (RMSE) is estimated to be very small and hence a good model (< 0,2) for the vast majority of the georegions based on the independent validation data. Also the mean absolute error (MAE) is relatively small in most georegions (0.003-0.055 mg P / l).

1 Introduktion

I NOVANA-overvågningsprogrammet er der siden 2007 anvendt en fosfor (P) model til beregning af månedlige vandføringsvægtede koncentrationer af total fosfor (TP) i umålte oplande til søer og kystvande (Ovesen et al., 2009; Tornbjerg et al., 2019). Den nuværende TP-model til beregning af månedlige vandføringsvægtede TP-koncentrationer blev udviklet i forbindelse med etableringen af den samlede DK-QNP belastningsmodel i perioden 2005-2007 (Ovesen et al., 2009). Modellen estimerer den vandføringsvægtede TP-koncentration i vandløb som et mål for bruttoudledningen fra diffuse kilder inklusive spredt bebyggelse. Modellen blev opstillet på NOVA-programmets (Nationale overvågningsprogram for Vandmiljøet) intensive måledata fra 24 vandløb fra perioden 1994-2002. I modellen indgår der data fra dyrkede typeoplande, dvs. oplande med et areal på mindre end 50 km^2.

Modellen er en regressionsmodel, og der indgår i modellen et antal forklarende parametre, som er dyrkningsgrad i oplandet, jordtypen (andel af sandede jorder), andel af vådområder i oplandet, baseflow-indeks i vandløbet (mål for andel af grundvand (årligt) og nedbøren (månedligt). Den nuværende anvendte NOVANA DK-P model er kun udviklet på data fra NOVA-perioden 1993-2002 og under anvendelse af ældre korttemaer over jordtyper og arealanvendelse. Derfor var formålet med projektet i udgangspunktet at opdatere NOVANA DK-P modellen med nye tilgængelige data og korttemaer, herunder en årlig opdatering af arealanvendelsesparametre som f.eks. dyrket areal. Derudover var det i projektet hensigten at modernisere den eksisterende DK-P-model ved anvendelse af nye modeludviklingsmetoder såsom anvendelse af tolags-krydsvalidering i forbindelse med almindelige regressionsmetoder. Det viste sig imidlertid, at moderniseringen af den nuværende NOVANA DK-P model ikke gav et tilstrækkeligt godt resultat, og det blev derfor besluttet at udvikle en helt ny fosformodeltype baseret på maskinlæringsmetoder.

Med henblik på forbedring af DK-P-modellen og gennemførelse af fremtidige nye måleindsatser for P i vandløb blev projektet gennemført med følgende fem overordnede elementer:

- Analyse af de intensive fosformåledata fra NOVANA.
- Review af anvendelsen af sensorstationer i udlandet (publiceret i DCE, AU-notat: van't Veen et al., 2020).
- Modernisering af den nuværende NOVANA DK-P model, som anvendes til fosforbelastningsberegninger til kystvande og søer (denne rapport).
- Udvikling af ny TP-model baseret på maskinlæringsmetoder (denne rapport).
- Ny beregning af tidsserie af fosforbelastninger til kystvande 1990-2020 med ny fosformodel (denne rapport).

I projektet er der under processen med udvikling af den nye TP-model til anvendelse i NOVANA gennemført flere trin, som vist i tabel 1. Rapporten er derfor også opdelt i kapitler, der afspejler disse trin.

I de enkelte trin er der udviklet på og testet forskellige modeltyper både ift. anvendelse af traditionelle regressionsmetoder, anvendelse af tolags-krydsvalidering og nye maskinlæringsmetoder. Desuden er der i projektet blevet

oparbejdet et meget stort datamateriale med input til modellen, der som et slutprodukt er delt med Miljøstyrelsen via et fællesdrev. Derfor er der også i de enkelte trin i modeludviklingen anvendt forskellige inputdata til udvikling af de forskellige TP-modeller. Resultaterne af hvert af de udførte trin i processen hen mod en ny moderniseret TP-model er beskrevet i de enkelte afsnit i rapporten, og deres indhold er opsummeret i tabel 1.

Tabel 1. De enkelte trin i udviklingen af den nye NOVANA total fosfor-model. LP: Landbrugspakke; ML: Maskinlæring; REG: Regression; Typeoplande: Oplande fra NOVANA, hvor der er en dominerende kilde til næringsstoftab, som f.eks. landbrug, natur etc.; LOOP: Landovervågningsoplande i NOVANA.

De enkelte trin i projektet	Formål	Metode	Datagrundlag	Navn	Resultatafsnit
Trin 1 (2019)	Test af eksisterende NOVANA TP-model (DK-P)	Regression	Intensive måledata fra NOVANA LOOP stationer (2003-2018)	NOVANA DK-P	4.1
Trin 2 (2019)	Test af eksisterende TP-modeller	Test på intensive måledata (daglige)	Odense Å, Kratholm (1990-2007)	NOVANA DK-P samt en model for TP, partikulært P og opløst orthofosfat fra Andersen et al. (2005) Med test af forskellige versioner af beregnet baseflow-indeks (BFI) i TP-model	4.2
Trin 3 (2020)	Modernisering af eksisterende NOVANA DK-P regressionsmodel for årlig vandføringsvægtet TP-koncentration	Tolags-krydsvalidering i SAS-software med proc reg	Data fra intensivstationer i perioden 1993-2018	En ny REG-P model på intensive måledata fra alle stationer	4.3
Trin 4 (2020)	Udvikling af en prototype model for årlig vandføringsvægtet koncentration af TP	Maskinlærings (ML) software DataRobot	Stikprøvedata fra alle typeoplande uden større punktilder og søer	ML-prototype Ny TP-model	4.4
Trin 5 (2021)	Udvikling af en maskinlæringsmodel i DataRobot og sammenligning med en tilsvarende regressionsmodel udviklet i SAS software	Maskinlæring (ML) software (DataRobot) og regression	Stikprøvedata fra alle typeoplande inkl. LP typeoplande uden større punktilder og søer	ML-01 Ny TP-model Ny REG TP-model	4.5
Trin 6 (2021)	Udvikling og test af 5 versioner af ny TP-model for årlig vandføringsvægtet koncentration på baggrund af alle typeoplande inkl. de nye Landbrugspakke-stationer fra 2016	Maskinlærings (ML) software DataRobot	Stikprøvedata fra alle typeoplande inkl. LP oplande uden større punktilder og søer	ML-version 2-6 Ny TP-model	4.6

Processen omkring udvikling af den nye TP-model i trin 6 har inkluderet mange forskellige forsøg med anvendelse af forskellige statistiske metoder samt datasæt. Efter hvert trin i processen blev der vurderet på modellens performance og derudfra besluttet, hvordan der skulle arbejdes videre. I trin 1-3

anvendes normale regressionsmetoder og data fra 24 såkaldte intensivstationer i NOAVA i et forsøg på at modernisere den nuværende NOVANA DK-P model. Da dette ikke gav et tilfredsstillende resultat, blev det besluttet ikke at arbejde videre med modernisering af den nuværende NOVANA DK-P model, men i stedet at udvikle en helt ny type model. I trin 4-6 anvendes maskinlæringsmetoder og data fra stikprøver fra alle 349 typeoplande i NOVANA-programmet. Den endeligt udviklede nye TP-model med dens resultater for TP koncentrationen i ID15-oplande er derfor beskrevet under trin 6 i rapporten.

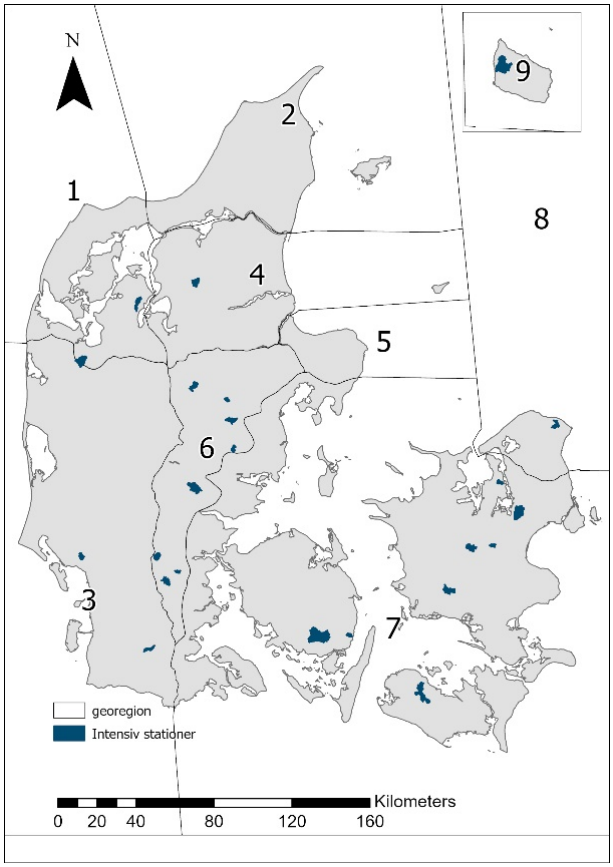
2 Datagrundlaget

Data til validering af den eksisterende fosformodel og udvikling af en ny TP-model i projektet har været de tilgængelige NOVANA-data i ODA-databasen (overfladevandsdatabasen med NOVANA-data) for beregnet måneds- og årstransport af TP og vandafstrømning. Der er tale om data fra intensive målinger af TP-koncentrationen ved i alt 24 vandløb i perioden 1993-2003 samt de efterfølgende intensive måledata fra Landovervågningsoplandene (2004-2017). Desuden er der anvendt stikprøvedata fra alle typeoplande for perioden 1990-2019, også de mange nye typeoplande, der er oprettet efter Landbrugspakken i 2016.

For alle målestationer er den årlige / månedlige vandføringsvægtede koncentration af TP fra diffuse kilder (inkl. spredt bebyggelse) blevet beregnet ved at fratrække udledninger fra punktkilder i oplandet og herefter dividere den diffuse transport af TP med vandafstrømningen. I resten af rapporten angiver TP-koncentration derfor altid den vandføringsvægtede TP-koncentration. For alle målestationer er punktkildernes andel af den målte TP-transport blevet opgjort, og hvis den udgør mere end 25 % af årets transport ved målestationer, er dette års data blevet udeladt fra kalibrering og validering af modeller.

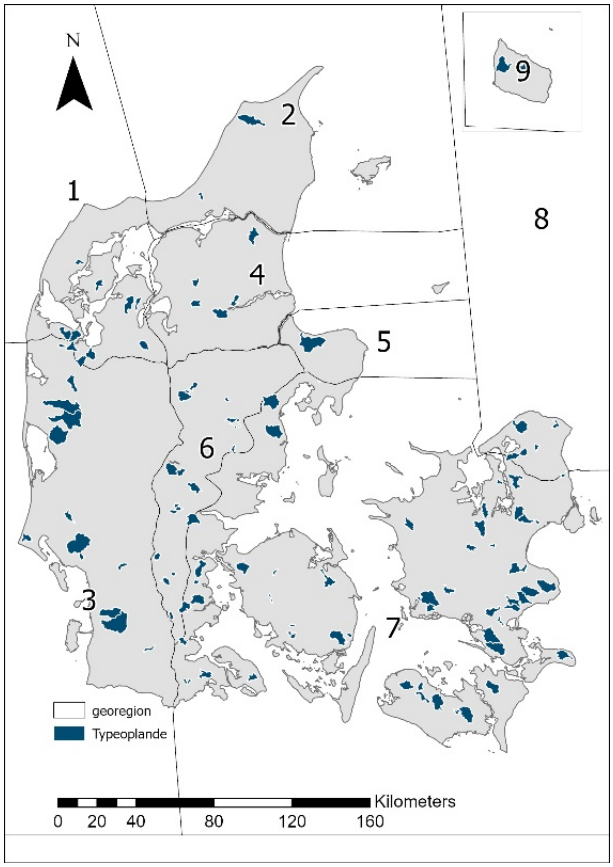
De i trin 1-3 anvendte intensivt målte oplandes geografiske placering er vist i figur 1. Placeringen af de anvendte typeoplande i modeludviklingens trin 4 er vist i figur 2, og typeoplan-de til trin 5 og 6 er vist i figur 3.

Inputdata til de forskellige modeller er generet i ArcGis eller i SAS-software ud fra en række forskellige kilder (bilag 1).

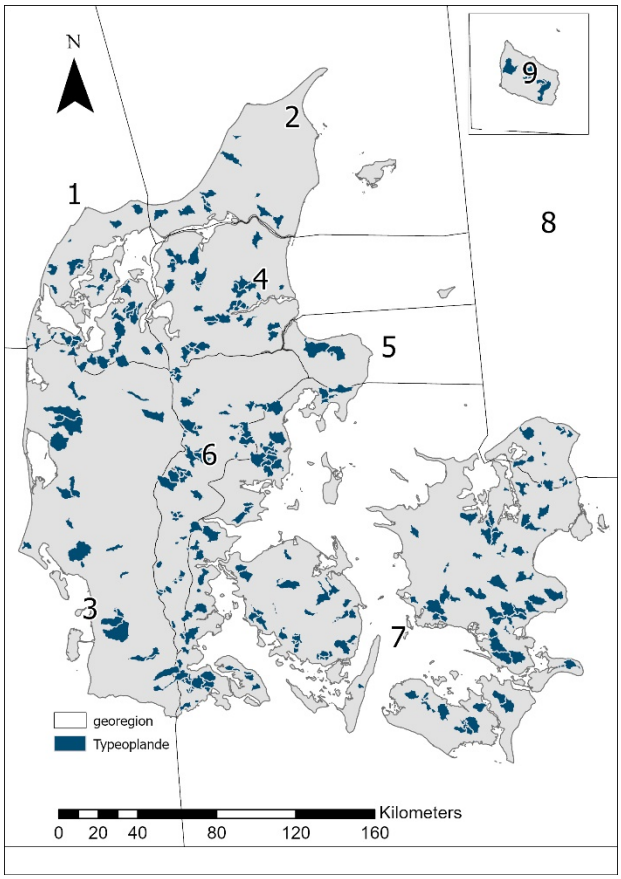


Figur 1. Beliggenhed af oplande til de 24 intensivstationer anvendt i trin 1-3. Numre er georegionerne.

Figur 2. Beliggenhed af de 107 typeoplande anvendt i trin 4 i modeludviklingen. Numre er georegionerne



Figur 3. Beliggenhed af alle 349 typeoplande anvendt i trin 5-6 af modeludviklingen. Numre er georegionerne



3 Anvendte metoder

Trin 1

Den nuværende anvendte model til simulering af måneds-vandføringsvægtede TP-koncentrationer (NOVANA DK-P modellen) blev udviklet i 2005-2007 på baggrund af data fra 24 målestationer, etableret i NOVA-overvågningsprogrammet i perioden 1993-2002 (figur 1). De intensive målestationer bestod af en automatisk vandprøvetager (ISCO), som udtog en puljet vandprøve til analyse af fosfor. I perioden 1993-2002 anvendtes både en tidsproportionalt prøvetagningsstrategi, en flomproportional eller en flowproportional strategi. Typisk er der analyseret på en ugentlig puljet vandprøve, men ved flomproportional prøvetagningsstrategi er der analyseret på puljede døgnprøver hver for sig fra dage med stigende vandføring. NOVANA DK-P modellen til simulering af månedlige TP-koncentrationer i vandløb er vist i ligning (1) og er nærmere beskrevet i Ovesen et al. (2009) og Tornbjerg et al. (2019).

$$TP(i,j) = 1{,}2436 \cdot \exp(\alpha_j + \beta_1 \cdot \mathrm{D} + \beta_2 \cdot \mathrm{J} + \beta_3 \cdot \mathrm{V} + \beta_4 \cdot \mathrm{BFI_i} + \beta_5 \cdot \mathrm{N_{i,j}}) \quad \textit{(ligning 1)}$$

i: år; j: måned;

$\alpha_{j:}$ angiver et koncentrationsniveau for hver kalendermåned.

D: Dyrkningsgrad er andelen af dyrket land i oplandet (angivet som procentpoint).

J: Jordtype er andelen af sandjorde (j1+j2+j3) i oplandet (angivet som procentpoint).

V: Vådområder er andelen af vådområder i oplandet (angivet som procentpoint).

BFI: Årlig baseflow-indeks (BFI).

N: Nedbør står for den ukorrigerede månedlige nedbør i mm i et givet år.

$\beta_{1\text{-}5}$ er parameter estimater i modellen.

.

BFI i år (i) beregnes ud fra en NOVANA BFI-model, der som forklarende variable indeholder georegion, nedbør, dyrkningsgrad, jordtype og andel befæstet area i oplandet.

BFI–modellen afhænger af, hvilken georegion der skal beregnes for (her vist for georegion 1):

BFI = 1.3152-0.8620-0.0001957*N-0.005439*J-0.007967*H+0.01091*J; (ligning 2),

hvor

N: ukorrigeret nedbør pr. år (mm fra DMI's gridnet).

J: procent sandjorde i oplandet (jordtypekort DJF).

H; procent befæstet areal i oplandet (jordtypekort DJF).

Modellerne er yderligere beskrevet i Ovesen et al. (2009) og Larsen et al. (2016).

I perioden 2003-2018 er den tidsproportionale intensive prøvetagning fortsat i Landovervågningsoplandene (LOOP) under NOVANA-programmet (Blicher-Mathiesen et al., 2021). Data fra disse stationer blev i trin 1 anvendt til en validering af den eksisterende TP-model (NOVANA DK-P model).

Trin 2

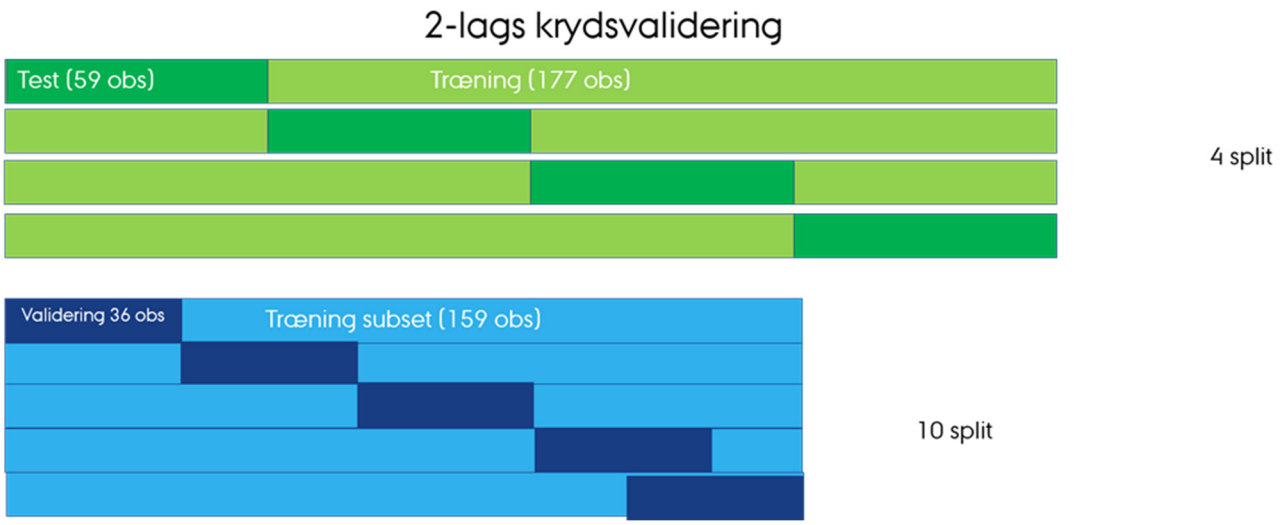
På dette trin i processen blev det af følgegruppen for projektet besluttet at teste forskellige versioner af den nuværende NOVANA DK-P model l samt en model for opløst orthofosfat. Testen skete på en lang tidsserie af intensive (daglige) måledata fra en målestation i Odense Å ved Kratholm i perioden 1990-2007. Modelversioner, der blev testet, er den nuværende NOVANA måneds TP-model. I denne model indgår baseflow-indeks (BFI) som forklarende variabel. Det blev besluttet at teste NOVANA DK-P modellen i en version med et BFI beregnet ud fra forskellige metoder:

1) beregnet ud fra delmodellen angivet ovenfor under trin 1 (ligning 2); 2) beregnet ud fra udtræk af data om daglig vandføring fra den nationale vandressourcemodel (DK-modellen); 3) beregnet ud fra faktiske målinger af vandføring ved målestationen.

Trin 3

På dette trin blev det besluttet at forsøge at modernisere den eksisterende DK-TP-model til simulering af årlig vandføringsvægtet TP-koncentration med anvendelse af tolags-krydsvalidering (statistisk metode, der anvendes under modeludviklingen) i SAS-software med anvendelse af rutinen 'proc reg' i SAS (SAS Inc.). Data til modeludviklingen er alle data fra intensivstationerne i perioden 1993-2017.

Krydsvalidering anvendes i bl.a. regressionsanalyser for at undgå overtilpasninger af den opstillede sammenhæng. Det sker ved at gentage træning af model og teste denne i flere omgange (splits) ved opdeling af data i et træningsdatasæt og et testdatasæt og gentage dette i flere omgange (splits) (figur 4). Der er i arbejdet med udvikling af modellen anvendt en tolags-krydsvalideringsmetode, hvor data til opstilling af regressionsmodeller først er opdelt i et øvre lag af data, der køres igennem i fire split, og et nedre lag med uafhængige data fra øvre lag, der køres igennem i ti split. Efter hvert split vurderes den opstillede models egnethed ud fra et statistisk mål for modelgodhed (RMSE). Metoden er vist i figur 4 med det antal data, som indgår i de forskellige lag. Der er arbejdet med fire split i det øvre lag og 10 split i det nedre lag i modeludviklingen.



Figur 4. Eksempel på anvendelse af tolags-krydsvalidering i udviklingen af fosformodellen med udgangspunkt i datasættet fra intensivstationerne i NOVA- og NOVANA-programmet 1993-2017. I træningsdatasættet for modelen (de lysegrønne) er der 177 observationer i alle split og 59 observationer i testdatasættet (mørkegrønne). Tilsvarende er der 159 observationer (lyseblå) i trænings-subsættet i alle lag og 36 observationer (mørkeblå) i alle valideringsdatasæt.

Trin 4-6

I efteråret 2020 blev det klart, at der var et homogenitetsbrud i den tidsserie af korrigeret gridnedbør, der bl.a. indgår i beregningen af stoftransporter i NOVANA-regi og indgår som forklarende parameter i DK-P TP-modellen (Svendsen & Jung-Madsen (red)., 2020).

Det blev derfor nødvendigt at opdele udviklingen af TP-modellen i flere etaper for at kunne arbejde videre, indtil et nyt datasæt for nedbør var udarbejdet, som kunne indgå i modellen. Et nyt datasæt for korrigeret gridnedbør, der dog kun retter op på en del af den observerede inhomogenitet, blev tilgængeligt foråret 2021 (Vejen et al., 2021).

I trin 4-6 arbejdes med udviklingen af en helt ny TP-model i maskinlæringssoftwaren 'DataRobot'. DataRobot indeholder 70 forskellige versioner af maskinlæringsrutiner, som alle testes i et trin af modeludviklingen. Vi har anvendt autopilotfunktionen i DataRobot til alle kalibreringer af modellen.

Responsvariablen er de årlige vandføringsvægtede TP-koncentrationer, som transformeres med den naturlige logaritmefunktion, før data læses ind i DataRobot maskinlærings- (ML) rutinerne for at sikre, at data er normalfordelte. Tilsvarende transformeres den årlige nedbør (i mm), som er en forklarende variabel, med den naturlige logaritme. I referencen er der en udtømmende beskrivelse af alle de forskellige ML-rutiner, som er medtaget (DataRobot©, 2021).

Den endelige model for TP vælges som den ML-rutine, der giver den bedste estimering af TP-koncentrationerne i kalibreringsdatasættet. Hvordan DataRobot© evaluerer modelfit forklares senere i dette afsnit.

DataRobot© deler kalibreringsdatasættet op i tre fraktioner:

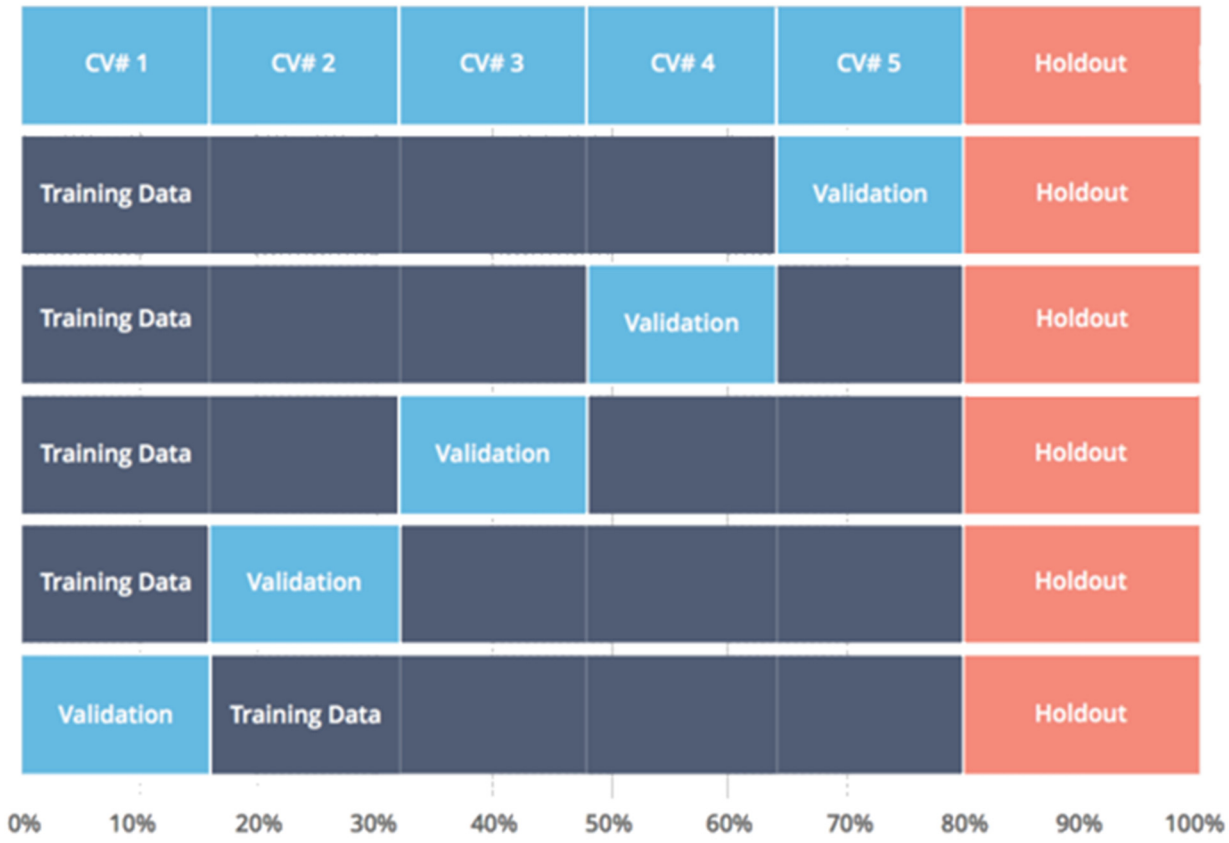
1. Træningsdatasættet (omtales også tit som 'validation'): Udgør 64 % af kalibreringsdata.

2. Valideringsdatasættet (omtales også tit som 'cross-validation': Udgør 16 % af de anvendelige kalibreringsdata.

3. 'Holdout'-datasæt: Udgør 20 % af de anvendelige kalibreringsdata.

Træningsdatasættet anvendes til at estimere / bygge modeller. Modelkoefficienter eller splits i beslutningstræet (splits er opdelinger af beslutningstræet) er bestemt ved anvendelse af dette datasæt. Valideringsdatasættet anvendes til at få et unbiased estimat for modellens nøjagtighed og præstation på et datasæt, som ML-rutinen ikke har set før. Resultaterne af denne uafhængige validering er vigtig, når de forskellige ML-rutiner skal sammenlignes ved endeligt valg af model. Endelig anvendes 'holdout'-datasættet til at give en yderligere mulighed for at evaluere den model, som vælges baseret på 5-fold krydsvalideringen. Dette er fordelagtigt, da valideringen med valideringsdatasættet typisk er lidt for optimistisk, da disse data indgår i den endelige evaluering af godheden af forklaringsgraden i de enkelte rutiner.

I standardopsætningen af DataRobot© anvendes en 5-fold krydsvalidering, som illustreres i figur 5, hvor man også kan se 'holdout'-datasættet.



Figur 5. Princippet i 5-fold krydsvalidering (fra DataRobot© manualen).

Selve modelkalibreringen og 5-fold-krydsvalidering foregår på følgende måde:

Først allokeres tilfældigt og på forhånd 20 % af kalibreringsdatasættet til 'holdout'. Dernæst opdeles de resterende 80 % af observationerne i kalibreringsdatasættet efter følgende procedure:

Data deles tilfældigt op i fem lige store grupper (folds) – 16 % hver.

For hver af de fem grupper udføres følgende 4 punkter:

1. Brug gruppen som valideringsdata.

2. Anvend de resterende fire grupper (64 %) som træningsdata.

3. Kalibrer en model på dette træningsdataset, og evaluer den på trænings- og valideringsdata (RMSE- Root Mean Square Error)

4. Behold de to RMSE-værdier.

Dernæst beregnes gennemsnit af de fem RMSE-værdier (et gennemsnit for træningstrinene og et for valideringstrinene). Derved beregnes en gennemsnitlig performance for de fem kørsler.

Modellen med den bedste performance (laveste RMSE) anvendes til prædiktion.

På denne måde er hver observation i kalibreringsdatasættet brugt en gang til validering og fire gange til træning af modellen.

I DataRobot udvælges den metode, der performer bedst på de indlæste forklarende variable. Når datasættet med TP-koncentration og alle de valgte forklarende variable er indlæst i DataRobot, så afvikler DataRobot alle de 70 rutiner i autopiloten. Standardopsætningen vedrørende ”eXtreme Gradient Boosted Trees Regressor” i DataRobot©, som viste sig at være den bedste algoritme til TP-modeludvikling, er i store træk anvendt (se tabel 2). Kun en parameter vedrørende opsætningen er ændret i forhold til standard, nemlig ’early stopping’, som er sat til 10 (se forklaring på DataRobot-optimering i tabel 2).

Gradient Boosting ML (Breiman, 1997; Breiman et al., 2001; Friedman, 2001; Hastie et al., 2009; Chen and He, 2015) (GBM) er en algoritme, som kan anvendes til at opnå kalibrering af præcise simuleringsmodeller, og den anses generelt for at være en af de bedste, hvis ikke den bedste, ML-rutine, der findes til mange formål (Kaggle, 2015). GBM er en generalisering af adaboost-algoritmen (Freund and Schapire, 1997) og minder om random forest-algoritmer. Extreme Gradient Boosting ML (XGBoost) er en optimeret version, som afvikles hurtigere og giver større prædiktionsnøjagtighed. Boosting er forklaret i Hastie et al. (2009) sammen med en generel beskrivelse af statistisk ML. I Géron (2019) findes en mere generel beskrivelse af ML.

De enkelte ML-rutiners præstation i forhold til at estimere data evalueres via ”least squares”, det vil sige med størrelsen RMSE. Modelkalibreringen evalueres kun ved brug af RMSE for training- og valideringsfraktionerne, RMSE for ’holdout’-datasættet anvendes kun til en sidste evaluering af modellens præstation, efter at den bedste model er valgt.

Tabel 2. Parameteropsætning (kun de vigtigste er omtalt her).

Parameter	Værdi	Forklaring
Learning rate	0,3	Kontrollerer, hvor hurtigt modellen tilpasser sig data, altså hvor hurtigt træningen foregår.
Antal boostingtrin	100	Til reduktion af bias og varians.
Maksimumsdybde	3	Det maksimale dybde af træerne i modellen, det vil sige det maksimale antal af splits i den vertikale retning af træmodellen.
Antal iterationer	200	Rutinen vil køre mindst 200 gange – efter 200 iterationer vil rutinen tjekke for early stopping via moving average.
Early stopping	10	Denne parameter kontrollerer early stopping. Hvis RMSE ikke er blevet større i de 10 sidste iterationer, stopper rutinen. Dette starter først efter 200 iterationer,

Softwaren DataRobot© indeholder muligheden for at anvende den valgte ML-rutine til at prædiktere TP-koncentrationer ud fra et givet datasæt, som indeholder værdier for alle de nødvendige forklarende variable. Dette er foretaget på de tre datasæt nævnt ovenfor: Først og fremmest på et datasæt, som indeholder værdier for de forklarende variable for alle ID15-oplande for Danmark, men også på kalibreringsdatasættet, som jo blev anvendt til at opsætte modellen med, jf. beskrivelsen af kalibreringsproceduren i DataRobot©. Desuden er det foretaget på det, vi kalder det 'uafhængige' valideringsdatasæt af typeoplande, som slet ikke har været anvendt i DataRobot©.

De prædikterede værdier skal tilbagetransformeres med eksponentialfunktionen samt ganges med en korrektionsfaktor (Ferguson, 1986). Ferguson-korrektionen anvendes til at løfte de simulerede værdier op til middelværdien i en log-normal fordeling. Uden denne korrektion er de tilbagetransformerede værdier en medianværdi. Korrektionen beregnes med formlen: exp(0,5·MSE). SE er "mean squared error" beregnet på træningsdatasættet, dvs. på 64 % af kalibreringsdatasættet. De prædikterede årlige TP-koncentrationer beregnes altså ved:

$$TP_{pred} = \exp(\ln(TP)_{pred}) \cdot \exp(0{,}5 \cdot MSE) = 1{,}028 \cdot \exp(\ln(TP)_{pred}) \qquad (3)$$

Med andre ord løftes TP-koncentrationerne med 2,8 % ved anvendelse af Fergusonfaktoren. Til analyse af usikkerhed i modellen er der anvendt følgende statistiske beregninger:

RMSE (Root Mean Squared Error) beregnes som standardafvigelsen af model residualerne, altså modelprædiktionsfejlene. RMSE beregnes som

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{i_{målt}} - x_{i_{model}}\right)^2}.$$

Residualer er et mål for hvor langt væk de målte værdier ligger fra modellen. Så RMSE er et mål for hvor spredte disse residualer er, dvs. hvor koncentrerede de målte data ligger rundt om det bedste modelfit. MAE (Mean Absolute Error) er et mål for den gennemsnitlige modelfejl uden at tage hensyn til retningen af fejlen. MAE beregnes som

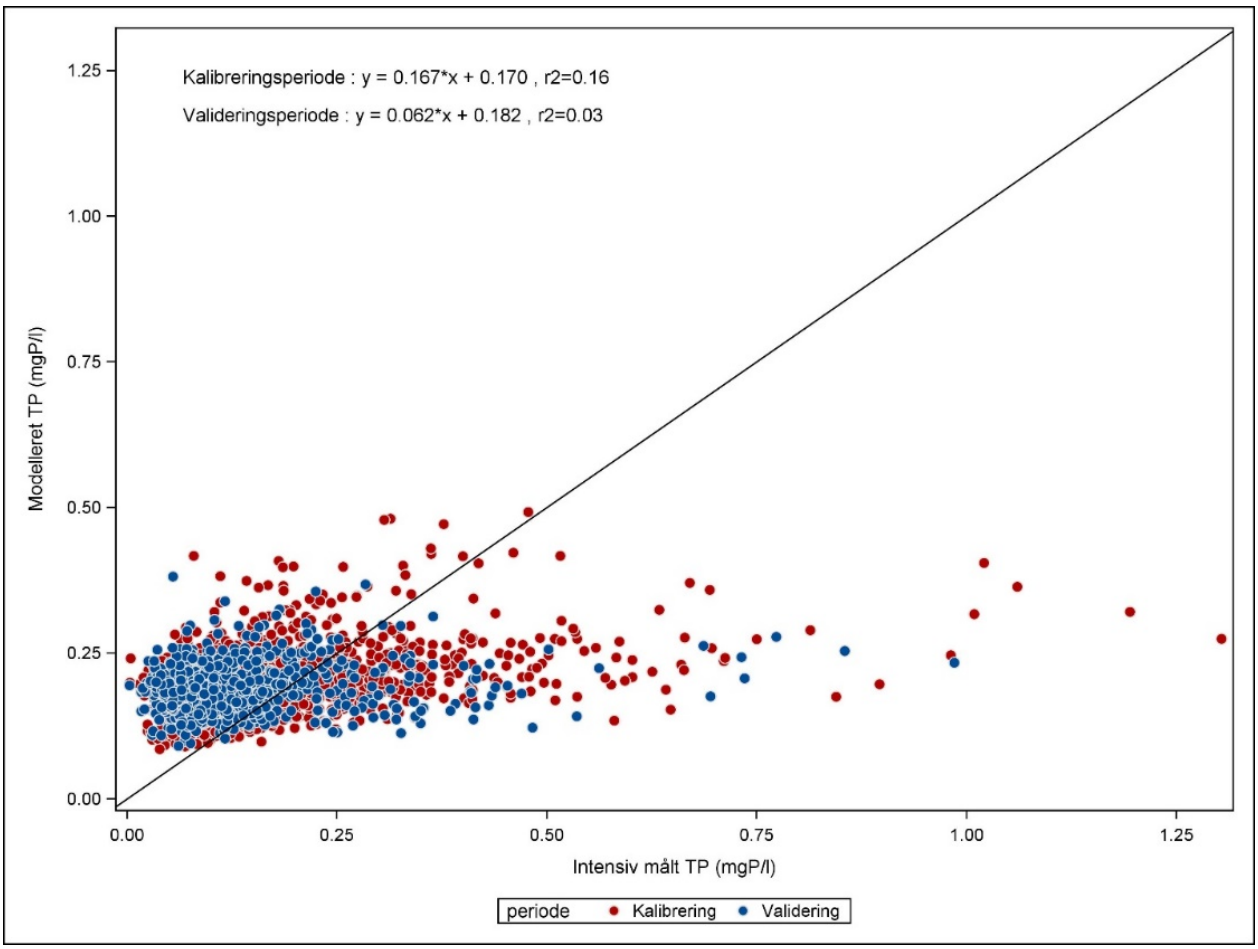
$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|x_{i_{målt}} - x_{i_{model}}\right|,$$

hvor $|x|$betegner den absolutte værdi af x. Både RMSE Og MAE har samme enhed som x værdierne og udtrykker begge middel modelprædiktionsfejl. Ligesom MAE ligger RMSE heller ikke vægt på retningen af modelfejlen. RMSE ligger mere vægt på store fejl, hvorimod MAE indregner alle modelfejl med samme vægt. Både RMSE og MAE kan antage værdier fra 0 til ∞ og jo mindre værdien er af dem jo bedre er modellen.

4 Resultater

4.1 Trin 1 – validering af åbent land NOVANA TP-modellen

Det første trin i projektet indeholdt en validering af den nuværende NOVANA DK-P model (Ovesen et al., 2009). Modellen anvendes til at simulere den vandføringsvægtede månedlige TP-koncentration fra diffuse kilder inklusive spredt bebyggelse. Valideringen af DK-P modellen i nærværende rapport blev gennemført på intensive målinger af TP-data fra NOVANA-vandløb gennemført i perioden 2003-2017. Valideringen viser, at TP-modellen har en meget dårlig forklaringsgrad (R^2=0,03) af de målte månedlige TP-koncentrationer, også sammenlignet med den beregnede R^2 på 0,16 fra den periode, som modellen er kalibreret på (nyt udtræk af data fra 1993-2002) (figur 6). Det fremgår også tydeligt af figur 6, at DK-P modellen generelt estimerer TP-koncentrationer inden for et meget mere snævert koncentrationsinterval, end målingerne fra vandløbene viser. Desuden overestimerer modellen i både kalibrerings- og valideringsperioden generelt koncentrationen ved de hyppigst forekommende TP-koncentrationer målt ved intensivprøvetagning i vandløb (< 0,25 mg P / l).



Figur 6. Målte og simulerede måneds-vandføringsvægtede koncentrationer af total fosfor med anvendelse af åbent lands fosformodellen fra NOVANA (DK-P) i henholdsvis kalibreringsperioden for modellen (1993-2002; nyt dataudtræk for de 24 intensivt målte vandløb) (angivet med røde prikker) og valideringsperioden (efter 2002) (angivet med blå prikker), hvori indgår ét års data fra de 24 intensivt målte vandløb og LOOP-data fra fem stationer i perioden 2004-2017. Linjen er 1:1.

Delkonklusioner fra trin 1 er:

- At der i projektet skal arbejdes videres med test af forskellige eksisterende modeltyper i et havbelastningsopland, hvor der findes intensive data af målte fosforkoncentrationer – Odense Å, Kratholm-stationen. Formålet er at vurdere anvendeligheden af forskellige regressions-modeltyper (år, måned) ved sammenligning af nuværende NOVANA DK-P model med en tidligere udviklet TP-model (Andersen et al., 2005).

- At der skal testes for betydningen af at anvende forskellige inputdata for beregning af baseflow-indeks (BFI), som indgår i NOVANA DK-P modellen. Der er tale om anvendelse af model til beregning af BFI (ligning 2), målinger af vandføring til beregning af BFI eller udtræk af data for daglig vandføring fra DK-modellen til beregning af BFI.

4.2 Trin 2 – test af forskellige fosformodeltyper i Odense Å-oplandet

Som et andet trin i analyserne gennemførtes en pilottest af forskellige TP-modeller udviklet ud fra de intensive måledata af fosforkoncentrationer fra målestation Odense Å, Kratholm. Her er der i en længere årrække blevet målt daglige koncentrationer af TP (1990-2009). I testen indgår den månedlige TP-model, der hidtil er blevet anvendt i NOVANA-opgørelser af havbelastningen i to versioner samt en årsmodel for simulering af diffus vandføringsvægtet TP fra Andersen et al. (2005) – den såkaldte VMPX-model – i tre versioner. Alle modeller simulerer ved hjælp af en række hydrologiske parameter og oplandsparametre den vandføringsvægtede koncentration af TP fra diffuse kilder i oplandet inklusive udledninger fra spredt bebyggelse, men med en forskellig beregnet BFI-parameter. BFI-parameteren beregnes enten med samme delmodel som i NOVANA DK-P modellen (formel 2), fra et udtræk af simuleret døgnmiddelvandføring fra DK-modellen for alle ID15-oplande i Odense Å oplandet eller fra målinger af BFI ved Kratholm målestation.

NOVANA DK-P modellen fremgår af formel (1), mens formlen for VMPX-modellen udviklet af Andersen et al. (2005) er:

$$C_i = (-2{,}215 + 0{,}00115 \times D + 0{,}00519 \times S + 0{,}0184 \times H - 1{,}621 \times BFI_i - 0{,}0668 \times B) \times 1{,}06 \quad (4),$$

hvor

C_i: Årlig vandføringsvægtet TP-koncentration i vandløbet

D: Andel dyrket areal i opland (%).

S: Andel sandjord i oplandet (%).

H: Hældning af oplandet langs vandløbet .

BFI: Årligt baseflow-indeks (ligning 2).

B: Andel riparisk buffer areal langs vandløb (%).

De forskellige TP-modeltyper, som blev testet i Odense Å, er følgende:

- NOVANA DK-P_1: Den oprindelige NOVANA DK-P model til simulering af månedlige vandføringsvægtede TP-koncentrationer med et baseflow-indeks (BFI) beregnet med BFI-delmodel (ligning 2) for alle ID15-oplande i Odense Å-oplandet.
- NOVANA DK-P_2: Den oprindelige NOVANA DK-P model til simulering af månedlige vandføringsvægtede TP-koncentrationer med et baseflow-indeks (BFI) beregnet på baggrund af data fra den nationale vandressourcemodel (DK-modellen; Højberg et al., 2015) med simuleret døgnmiddel-vandføring for alle ID15-oplande i Odense Å-oplandet som inputdata til beregningen.
- VMP1: Tidligere udviklet årsmodel for simulering af vandføringsvægtede TP-koncentrationer (Andersen et al., 2005). I modellen er BFI-parameteren beregnet på baggrund af vandressourcemodel- (DK-model) simulerede døgnmiddel-vandføringer i alle ID15-oplande i Odense Å-oplandet.
- VMP2: Tidligere udviklet årsmodel for simulering af vandføringsvægtede TP-koncentrationer (Andersen et al., 2005). I modellen er BFI-parameteren beregnet ud fra BFI ved Kratholm målestation, som derefter er anvendt i alle ID15-oplande i Odense Å-oplandet.
- VMP3: Tidligere udviklet årsmodel for simulering af vandføringsvægtede TP-koncentrationer (Andersen et al., 2005). I modellen er BFI-parameteren beregnet ud fra en BFI-delmodel i alle ID15-oplande i Odense Å-oplandet, som anvendes i NOVANA DK-P modellen (ligning 2).

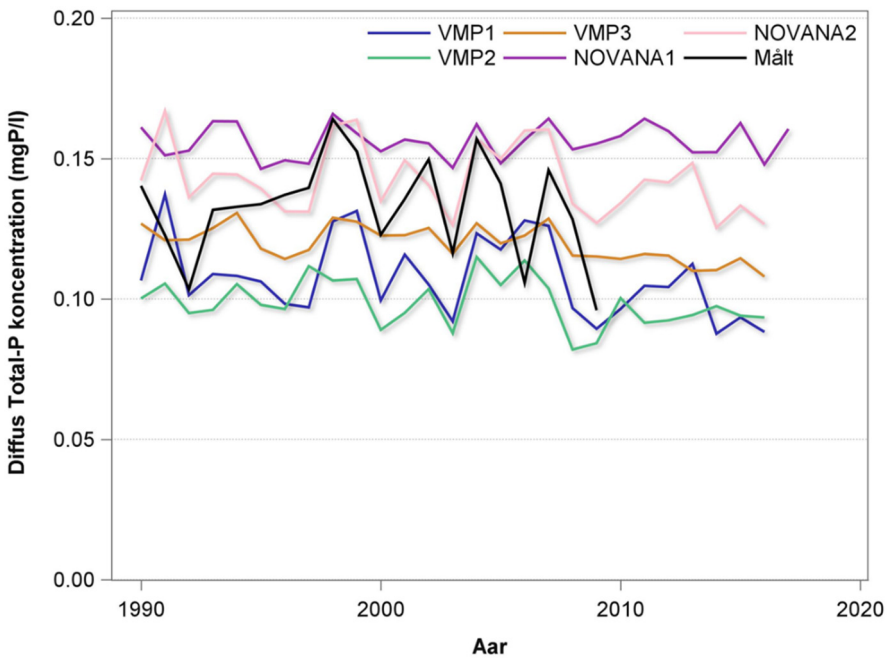
Modellerne blev testet ved opsætning på i alt 39 ID15-oplande inden for Odense Å-oplandet med beregning / udtræk af data for BFI for hver ID15. For hvert ID15-opland blev der herefter beregnet en transport af TP fra diffuse kilder, hvorefter der blev summeret til hele oplandet. Vandføringsvægtet TP-koncentration er så beregnet ved at dividere TP-transport med vandafstrømning for den pågældende periode.

Resultaterne af simuleringen med alle fem modeller mod de målte årlige vandføringsvægtede TP-koncentrationer i Odense Å Kratholm er vist i figur 7 og 8. De to NOVANA-modelversioner af den vandføringsvægtede årlige TP-koncentration overestimerer generelt den målte årlige vandføringsvægtede TP-koncentration det meste af tiden (1990-2008) (figur 7 og 8). Dette kan forklares ved, at den månedlige NOVANA TP-model formentlig overestimerer TP-koncentrationen i sommerhalvåret, da modellen ikke inddrager betydningen af TP-retention i vandløbssystemet fra fx sedimentation og biologisk optag.

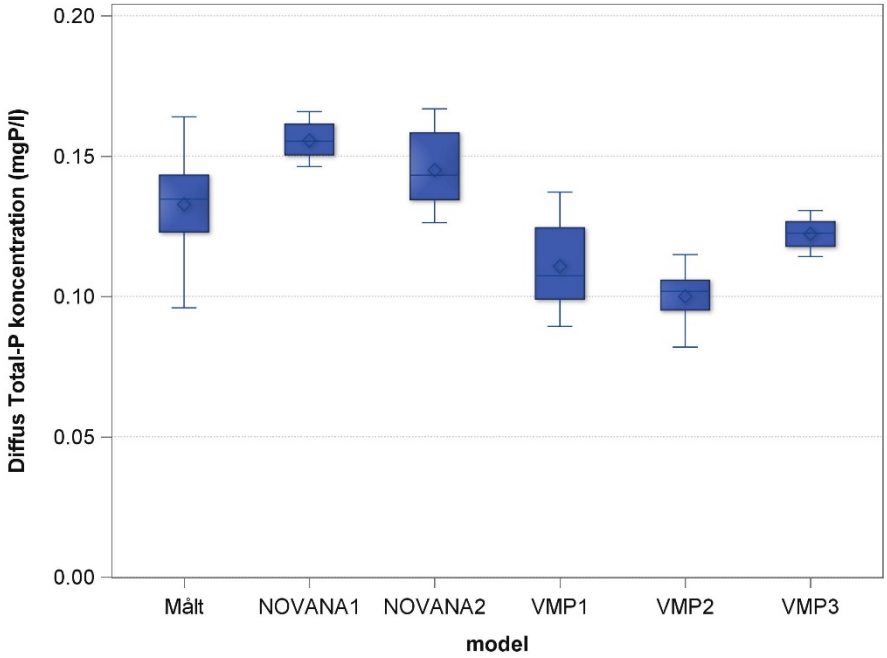
Derimod underestimerer de tre modelversioner, af den af Andersen et al. (2005) udviklede årlige TP-model, de målte TP-koncentrationer (figur 7 og 8). Dette kan muligvis forklares ved, at årsmodellerne ikke er gode nok til at inddrage betydningen af kortvarige høje TP-koncentrationer under regnbyger mv., hvor der både kan ske udskyl af fosfor fra vandløbsbund og -brinker samt tilførsel af især partikelbundet fosfor fra marker med overfladisk afstrømning og drænvand.

Modellen, som generelt simulerer tættest på de målte TP-koncentrationer i Odense Å, er årsmodellen med baseflow-indeks (BFI) simuleret med NOVANA BFI-modellen (VMP3) (figur 7 og 8).

Figur 7. Målte årlige vandføringsvægtede total fosfor-koncentrationer sammenholdt med simulerede koncentrationer med fem forskellige modeltyper i perioden 1990-2009.



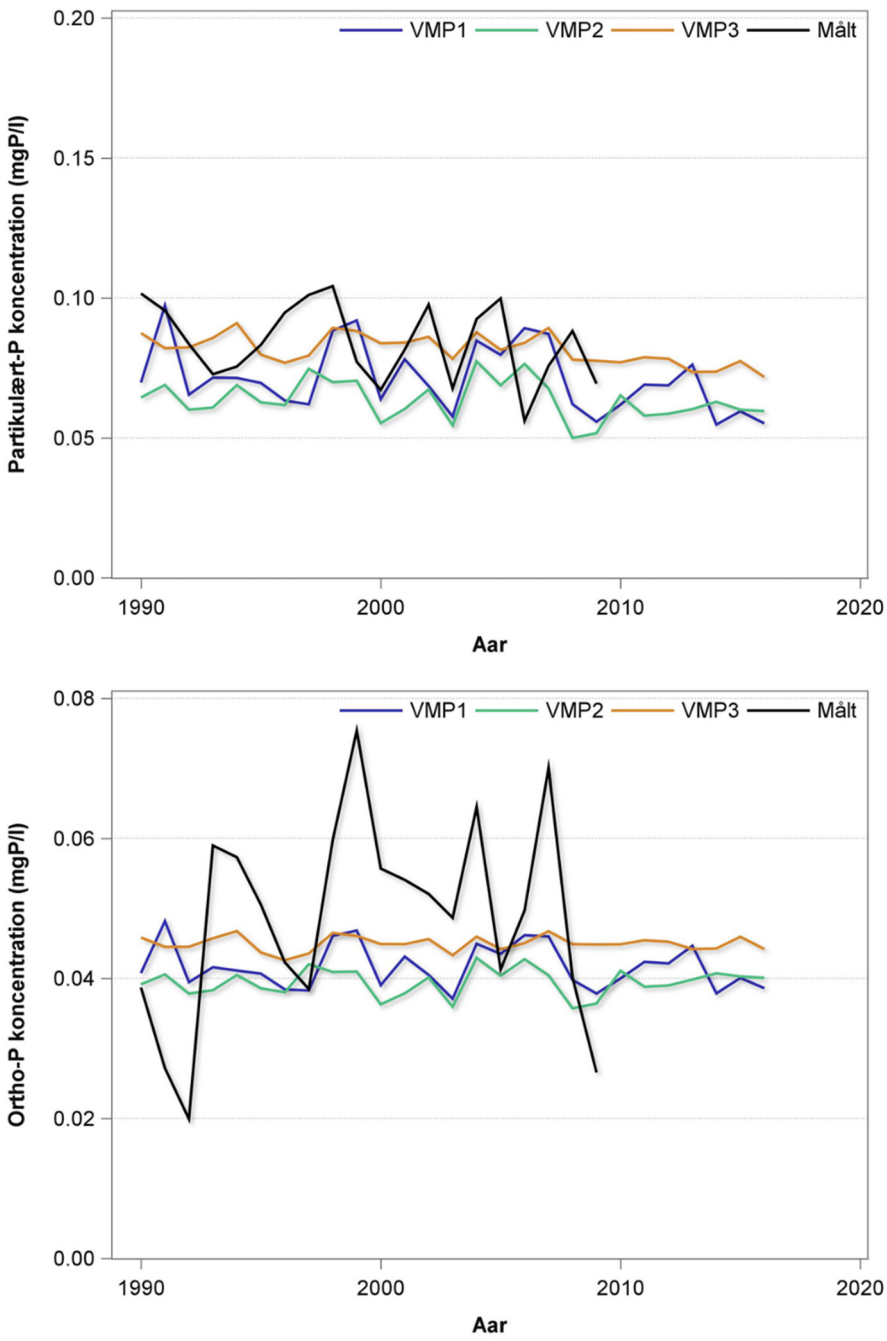
Figur 8. Målte årlige vandføringsvægtede total fosfor-koncentrationer sammenholdt med simulerede koncentrationer med de fem forskellige beregninger med to modeltyper i perioden 1990-2009. I figur er o = gennemsnit, median er stregen, 25 % og 75 % percentiler er bund og top af boksen, og linjeenderne er 10 og 90 % percentilerne.



I Andersen et al. (2005) indgår foruden årsmodellen for TP også to modeller, hvor hhv. koncentrationen af partikulært P (PP) og opløst orthofosfat P (OUP) estimeres. I figur 9 er de to modeller kørt med tre forskellige estimater af BFI tilsvarende årsmodellen.

Modelversionen (VMP3) er i gennemsnit bedst til simulering af hhv. de årlige koncentrationer af PP og OUP. Dog er ingen af modellerne særligt gode til simulering af dynamikken af hverken PP eller OUP (figur 9).

Figur 9. Analyse af årlig variation i den målte vandføringsvægtede koncentration af partikelbundet fosfor (TP minus opløst orthofosfat P) og opløst orthofosfat P i vandløb sammenholdt med modelberegnede koncentrationer for partikelbundet P og opløst orthofosfat P.



Analysen af de forskellige modellers simuleringsevne i Odense Å-oplandet, som er unik pga. de mange år med intensive målinger af koncentrationen af TP, viser, at de nuværende modeller udviklet på baggrund af intensive data fra mindre dyrkede typeoplande har problemer med at simulere de observerede P-koncentrationer med tilstrækkelig stor nøjagtighed. Da modellerne er udviklet til kun at simulere de diffuse kilder til P i oplandet (inkl. spredt bebyggelseTP) inkluderer modeller opstillet på baggrund af fratræk af udledninger fra større punktkilder et problem, især ved forsøg med udvikling af modeller på månedsniveau. Det skyldes en pt. manglende sikker opgørelse af udledningen fra større punktkilder hen over året. På baggrund af ovenstående overvejelser blev følgende besluttet:

- At der skal arbejdes videre med udvikling af en årsmodel for TP med anvendelse af tolags-krydsvalidering på baggrund af alle intensive data fra perioden 1994-2017.
- At der skal foretages en opdatering af forklarende parametre til den nye model, herunder inddragelse af ny gridnedbør fra DMI og input-data fra de nye fosforrisikokort (Andersen og Heckrath, 2020)
- At det fremadrettet i NOVANA-rapporteringen, ved anvendelse af en udviklet årsmodel for TP, er nødvendigt at anvende et reference-datasæt af målestationer i forskellige regioner af landet, når der er behov for at beregne månedstransport af TP.

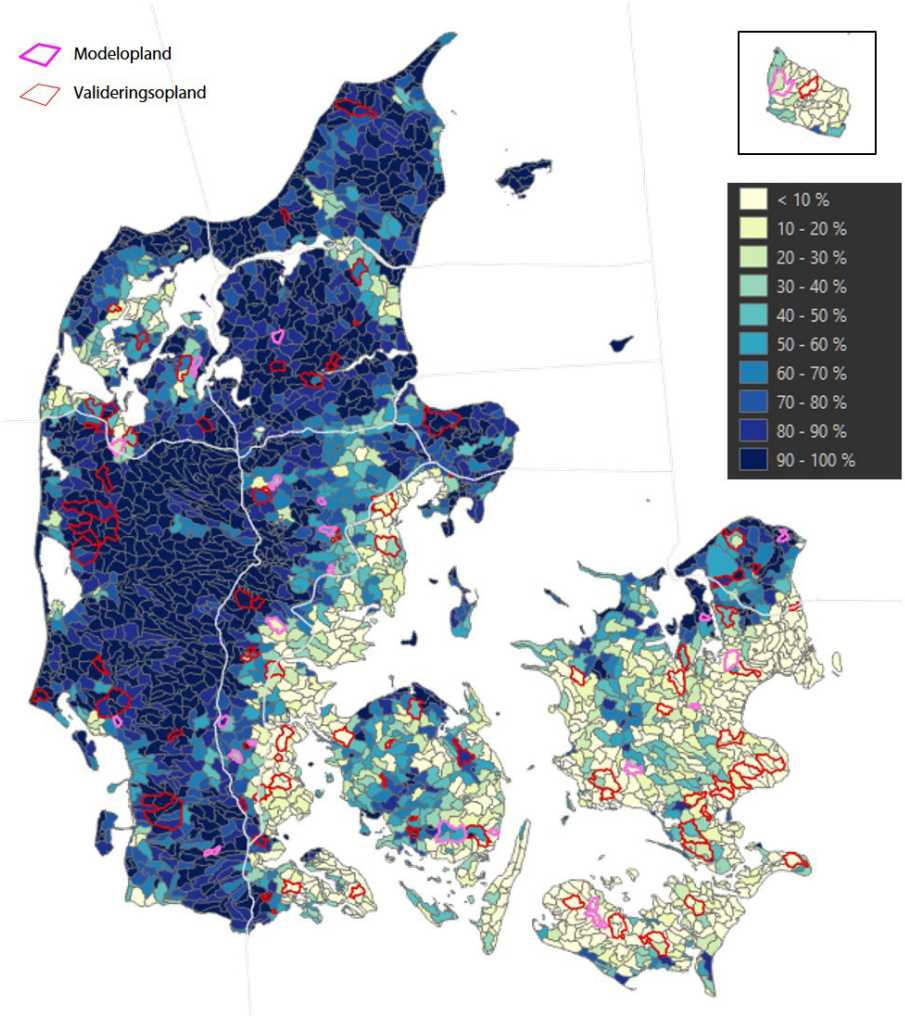
4.3 Trin 3 – udvikling af ny TP-model på intensive måledata af TP-koncentrationer

På baggrund af resultaterne fra de gennemførte valideringer af forskellige TP-modeller (trin 1 og 2) blev det besluttet at arbejde videre med at udvikle en årlig TP-model til simulering af vandføringsvægtet TP-koncentration på baggrund af alle intensive data fra NOVANA-målingerne i perioden 1993-2017. De nye etablerede forklarende parametre fra oplandene med intensive måledata i vandløb, som indgår i modeludviklingen, er vist i tabel 3. Variationen i mulige modelparametre i både kalibreringsdatasættet (intensivstationerne i NOVA) og valideringsdatasættet (landbrugs-typeoplande fra NOVANA) er vist på kort i bilag 2.

Tabel 3. Oversigt over mulige modelparametre i fosformodellen, som er beregnet for alle ca. 3200 ID15-oplande (yderligere informationer om temaer med referencer findes i bilag 1).

Parameter	Navn	Enhed	Kilde
Dyrket areal	DYRK	%	Markblok- og markkort, Corine-data
Befæstet areal	BEF	%	Basemap fra 2016
Potentiel drænet	DRÆN	%	PropDrain fra 2018
Sandjord	SAND	%	Jordtype2014 jb1-jb3
Lerjord	LER	%	Jordtype jb4-jb6
Humus	HUMUS	%	Jordtype2014 jb7
Organogen jord	ORG	%	Organogen2010
Hældning i oplandet beregnet langs vandløbstraceen	SLO	‰	DTMrain 1,6m grid og GeoDanmark vandløb
Erosionspotentiale	ERO	Tons/år	Fra Ohnen et al. (2018)
Spredt bebyggelse	SPRB	kg	FDC-punktkilder
BFI	BFI	0-1	Modelberegnet årlig BFI (ligning 2)
Nedbør årlig	NEDB	mm	Ukorrigeret nedbør 10x10 km grid fra DMI
Nedbørsafvigelse mm	NEDBAFV	mm	Afvigelse fra middelnedbøren 1991-2010
Nedbørsafvigelse	NEDBAFVREL	%	Afvigelse fra middelnedbøren 1991-2010
Georegion	GEO1-GEO9	1-9	Ni georegioner i DK (jf. figur 1)

Et eksempel er variationen i sandjordsprocent i oplandene (figur 10), hvor deres beliggenhed i Danmark er vist for både kalibrerings- og valideringsdatasættet.



Figur 10. Variationen i sandjordsprocenten i ID15-oplande med markering af kalibreringsoplande (modeloplande intensivstationer) og valideringsoplande (typeoplande, landbrug i NOVANA).

Den første TP-model, som blev udviklet med tolags-krydsvalidering i SAS-software, indeholder 17 parametre, hvoraf de ni dog er klassevariable for georegionerne:

Sgrt(TP konc) =
f(GEO1+GEO3+GEO4+GEO6+GEO7+GEO8+GEO9+BFI+ORG+SAND+HUMUS+ERO+SLO+NEDB+NEDBAFV)

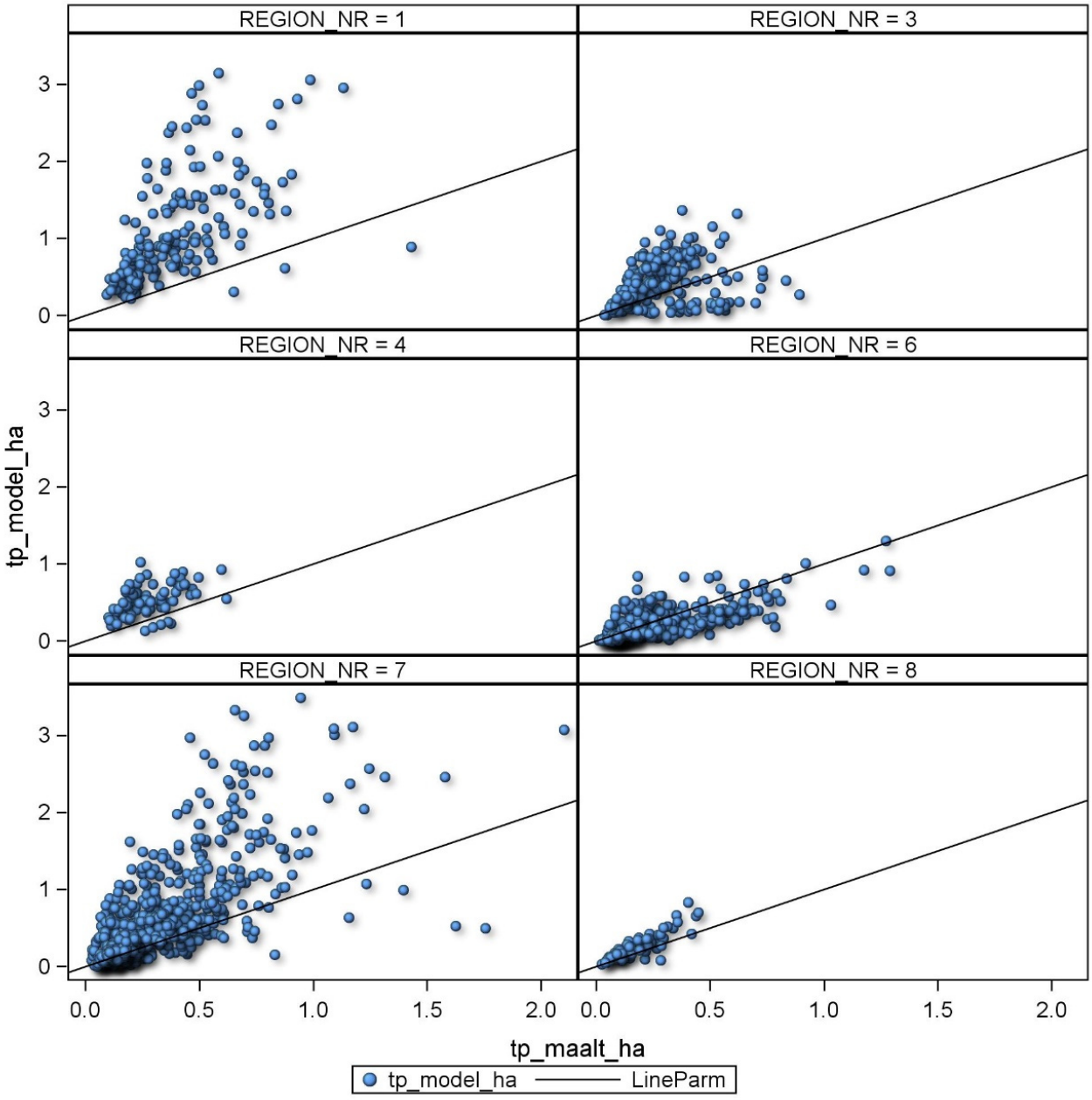
Den lineære models parameterestimater er vist i tabel 4, og modellen har en forklaringsgrad (R^2) på 0,49 ($P<0,0001$).

Tabel 4. Parameterestimater for den udviklede årsmodel for vandføringsvægtet koncentration af total fosfor i vandløb (der er ingen intensivstationer i georegion 2 og 5).

Parameter	Parameterestimat	Standard error	Signifikansniveau
GEO1	-0,625	0,336	0,065
GEO3	-0,819	0,405	0,045
GEO4	-0,592	0,384	0,125
GEO6	-0,894	0,411	0,031
GEO7	-0,565	0,306	0,067
GEO8	-0,429	0,310	0,169
GEO9	0		
BFI	0,451	0,312	0,151
ORG	0,048	0,024	0,041
SAND	-0,00230	0,000771	0,0034
HUMUS	-0,0128	0,00626	0,0433
ERO	3,259	0,937	0,0007
SLO	0,00870	0,00251	0,0007
NEDB	0,00105	0,000325	0,0015
NEDBAFV	-0,000654	0,000301	0,0315

Den udviklede TP-models forskellige residual- og kalibreringsplot er vist i figur 11.

TP-modellen er blevet testet mod uafhængige observationer fra 109 typeoplande fra NOVANA, hvor der er gennemført stikprøvetagning for TP-koncentrationen i perioden 1990-2017. I valideringsplottet i figur 11 er vist det årlige tab af TP, og valideringen er opdelt på georegioner (dog mangler georegion 2 og 5 pga. manglende oplande med intensive målinger; se figur 1). Det ses, at modellen i nogle georegioner simulerer det årlige tab af TP med en meget væsentlig bias (georegion 1 og 7), og for mange georegioner ses en dårlig simulering (stor spredning). Der kan også konstateres en bias i andre georegioner, men her er forskellen mindre udtalt mellem simuleret og målt årlig TP-koncentration (figur 11 og tabel 5).



Figur 11. Valideringsplot af simuleringen af det årlige tab af total fosfor (kg P/ha) med TP-modellen (y-aksen) mod det målte årlige tab i typeoplandene i NOVANA-programmet i perioden 1990-2017. Indsat 1:1 linjen.

Tabel 5. Oversigt over median valideringsresultater mellem målte vandføringsvægtede TP-koncentrationer og modelberegnede vandføringsvægtede TP-koncentrationer i vandløb inden for seks georegioner i Danmark. Kun vandløb/år, hvor udledningen af TP fra punktkilder udgør mindre end 10 % af årstransporten, er medtaget.

Geo-region	Antal år (N)	Antal stationer	P-konc. model (mg P/l)	P-konc. målt (mg P/l)	Difference (mg P/l)	Forskel (%)
1	214	9	0,33	0,12	0,21	175
3	365	14	0,14	0,08	0,06	75
4	144	6	0,22	0,10	0,12	120
6	504	20	0,06	0,08	-0,02	-25
7	1237	54	0,19	0,12	0,07	58
8	164	6	0,20	0,14	0,06	43

Delkonklusioner fra trin 3 er:

- At det er muligt at udvikle en regressionsmodel på baggrund af de intensive måledata, som giver en fornuftig forklaringsgrad (R^2) (0,49).
- At den udviklede regressionsmodel simulerer tab af total fosfor i dyrkede typeoplande med en meget stor bias i flere georegioner (f.eks. 1 og 7)
- At der for at få et bedre datagrundlag for hele landet i stedet skal forsøges udviklet en model for den årlige vandføringsvægtede koncentration af TP på baggrund af data fra et større antal typeoplande, hvor koncentrationen er målt ud fra stikprøvetagning.

4.4 Trin 4 – udvikling af en prototype model i DataRobot

Efter forsøget med udvikling af TP-modellen på intensive data fortsatte arbejdet i projektet med at forsøge at udvikle en ny TP-model i maskinlærings- (ML) softwaren DataRobot. Udviklingen af prototype ML TP-modellen sker på baggrund af stikprøver af total fosforkoncentrationen i 107 typeoplande uden større punktkilder i oplandet. Udviklingen af den nye model til simulering af årlig vandføringsvægtet TP-koncentration i vandløb foretages i flere etaper og under anvendelse af fem-fold krydsvalidering til endelig udvælgelse og test af TP-modellen.

Eftersom der i 2020 og primo 2021 pågik et arbejde i DMI med at udarbejde et nyt korrigeret nedbørsdatasæt (se afsnit 2), udvikles en prototypemodel, der senere kan opdateres med et revideret nedbørsdatasæt. Prototype ML TP-modellen er blevet udviklet i DataRobot med anvendelse af de udvalgte forklarende variable vist i tabel 6 (referencer til data findes i bilag 1).

Tabel 6. Oversigt over parametre i Prototype ML TP-modellen, som er beregnet for alle ca. 3200 ID15-oplande. Nummeret i kolonne 2 angiver rækkefølgen, som DataRobot giver i betydningen af de forklarende variable, der inddrages i den endelige prototype MS TP-model.

Variabelnavn	**Nr. i prototype model**	**Parameternavn**	**Enhed**	
Responsvariabel		Ln(tpdif_mgl)	Ln(mg P/l)	Naturlig logaritme til den diffuse vandføringsvægtede total P-koncentration inklusive spredt bebyggelse målt i typevandløb (<25 % udledning af P fra større punktkilder for typeoplandet)
Forklarende variable				
Dræningsgrad	1	DRÆN	%	Mål for detailafvandingen med drænrør i oplandet
Dyrkningsgrad	2	DYRK	%	Mål for dyrkningens indflydelse i oplandet
Sandjordsandel	5	SAND	%	Mål for forekomst af sandjord i oplandet
Høj risiko områder	9	ERO	%	Mål for forekomst af overfladisk afstrømning til vandløb
Lavbundsarealer med humusjord	4	ORG	%	Mål for forekomst af organiske lavbundsjorder i oplandet
Årlig temperatur	6	TEMP	°C	Mål for klimaets betydning – lufttemperatur
Årlig nedbørsafvigelse	8	NEDBAFV	mm	Årlig afvigelse fra middelnedbør over perioden (1990-2018) – ekstremer
Georegioner	11-19	GEO1-GEO9		Regionale forskelle – geologi
Årlig nedbør	10	Ln(NEDB)	ln(mm)	Nedbør som drivparameter (ukorrigeret)
Årlig afvigelse afstrømning	7	Q-indeks		($Q_{år}$-Q_{gns}/Q_{gns}) Data leveres fra DK-modellen
Udledning af TP fra spredt bebyggelse	3	SPRB_ind	kg x ln(mm)	Udtrykt ved styrken af belastning multipliceret med log(årlig nedbør_mm)

I modeludviklingen indgik i alt 2.166 årlige data på vandføringsvægtet TP-koncentration fra typeoplandsvandløb for perioden 1990-2018. Modellen skal alene kunne simulere den diffuse andel af den årlige total fosforkoncentration inkl. udledningen fra spredt bebyggelse. Derfor er den årlige fosforkoncentration i modeldatasættet renset for en eventuelt P-udledning fra andre punktkilder i oplandene.

For at data fra vandløb kan komme med i kalibrering af ML TP-modellen er der sat et krav om, at der minimum findes en tidsserie på seks år. Udledningen af TP fra punktkilder må for ethvert år højst udgøre 25 % af årets TP-transport, for at data fra det pågældende år og vandløb kan indgå i modeludviklingen. Dette kriterie er indført for at undgå usikkerheder ved fratræk af punktkildeudledninger fra den beregnede årstransport. Endelig er et atypisk vandløb med meget høje total fosforkoncentrationer (50000048: Kighanerenden) udelukket som en outlier. I alt er der afviklet 117 forskellige algoritmer i DataRobot med femfold-krydsvalidering hvor 80 % af datasættet indgår i modeltræningen, og 20 % er afsat til en uafhængig validering (holdout-datasæt).

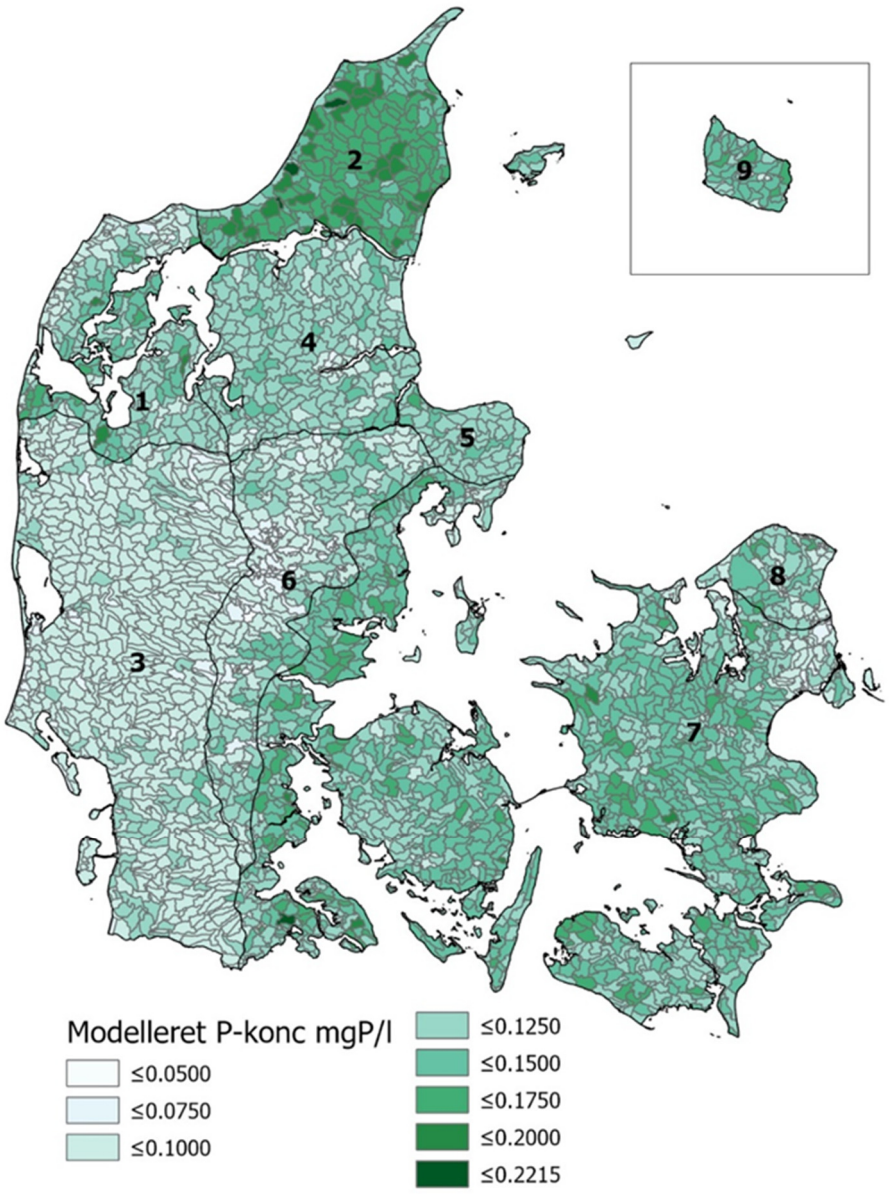
I DataRobot sker den bedste modeltilgang af de 70 metoder i softwaret 'Extreme Boosted Trees Regressor with Early Stopping'. I valideringen af den bedste model er der opnået en signifikant forklaringsgrad med $R^2 = 0{,}625$. I krydsvalideringen er der opnået en signifikant forklaringsgrad med $R^2 = 0{,}645$, mens der ved test på holdout-datasættet også er opnået en signifikant forklaringsgrad med $R^2 = 0{,}655$. Modellen består af følgende forklarende variable:

Ln(TP konc) = f(DRÆN; DYRK; SPRB_ind; ORG; SAND; TEMP; Q_indeks; NEDBAFV; ERO; ln(NEDB); GEO1-9) (se tabel 6 for nærmere beskrivelse af variable).

Rækkefølgen i betydning af de forklarende variable, som indgår i den udledte TP-model, fremgår af tabel 6 og er bestemt i DataRobot. Eksempler på, hvordan de enkelte modelvariable indgår i prototypemodellen, findes i bilag 3.

I DataRobot er der herefter indlagt alle de forklarende modelvariable for alle 3.215 ID15 i Danmark. Herefter er den trænede prototype ML TP-model anvendt i DataRobot til at simulere de årlige koncentrationer af TP for hvert år i perioden 1990-2018 (figur 12).

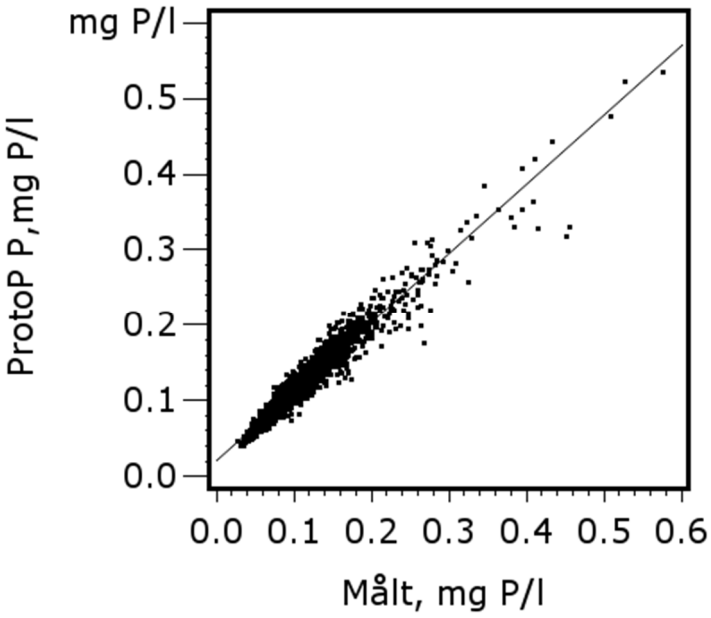
Figur 12. Den modelsimulerede (prototype) gennemsnitlige årlige vandføringsvægtede total fosforkoncentration i alle ID15-oplande i Danmark for perioden 1990-2018.



En delopgave i projektet var, at der gennemføres en validering af den udviklede TP-model samt en sammenligning af de nye modelresultater med tidligere versioner af TP-modellen (den oprindelige model fra 2009 og en bias-korrigeret version fra 2019).

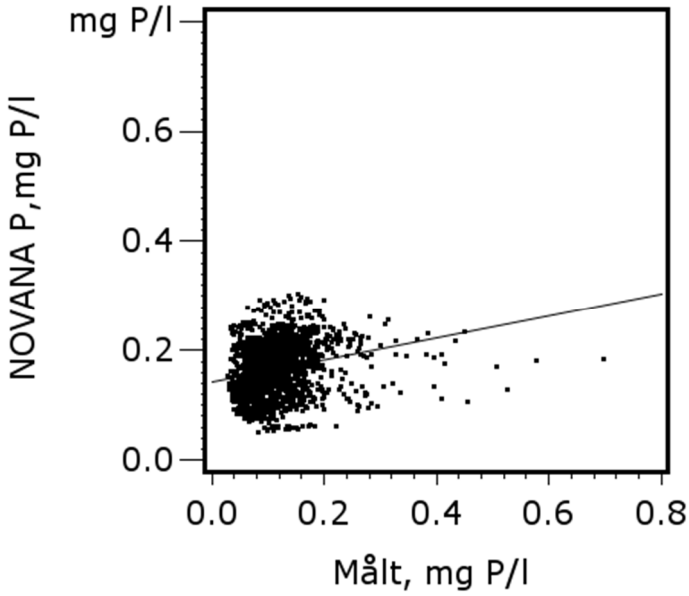
Således er der med resultaterne fra den udviklede prototype ML-model for årlig TP-koncentration gennemført en sammenligning i tre trin: 1) En sammenligning af modelsimuleringer mod målingerne i vandløb i de typeoplande, som indgår i modelopstillingen. 2) Validering mod et helt uafhængigt datasæt på 80 vandløb, der afvander lidt større oplande end typeoplande. 3) En sammenligning af modelsimuleringer på 2.-3.-ordens farvandsniveau af vandføringsvægtet TP-koncentration på årsniveau fra umålte oplande (ID15). Sammenligningen sker med simuleringerne med den nye prototype TP-model, som sammenholdes med de tidligere beregnede koncentrationer med NOVANA DK-QNP-modelkomplekset (med og uden bias-korrektion) (Tornbjerg et al., 2019; Thodsen et al., 2019).

Sammenligning af resultater fra prototype ML TP-modellen med målingerne i vandløb i typeoplande er vist i figur 14. Median oplandsstørrelsen i datasættet er på 15 km^2. Sammenhængen mellem observerede og estimerede værdier er signifikant og har en meget god forklaringsgrad og med en hældning tæt på en (Y= 0.91X + 0.02; R^2=0.94 n=2166 p<0.0001) (figur 13).



Figur 13. Simulering af vandføringsvægtet årlig total fosfor-koncentration med den nye prototype ML TP-model i vandløb i typeoplande sammenlignet medmålinger i samme.

Til sammenligning er i figur 14 vist den tidligere modelversion opstillet ud fra intensive fosformålinger i 24 oplande (se afsnit 4.2). I denne var forklaringsgraden meget ringere (Y= 0.20X + 0.14; R^2=0.05 n=2166 p<0.0001).

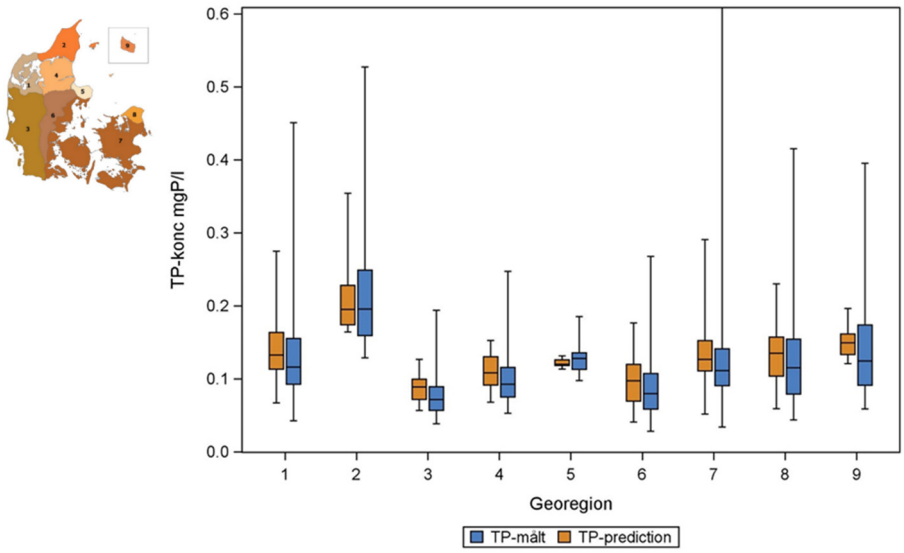


Figur 14. Simulering af vandføringsvægtet årlig total fosfor-koncentration med NOVANA-fosformodel uden bias-korrektion i vandløb i typeoplande mod målinger i samme.

En test af, hvordan prototype MS TP-modellen simulerer den vandføringsvægtede årlige koncentration af TP i de ni georegioner, som landet er opdelt i, er vist i figur 15. Det ses, at modellen simulerer TP-koncentrationen i de

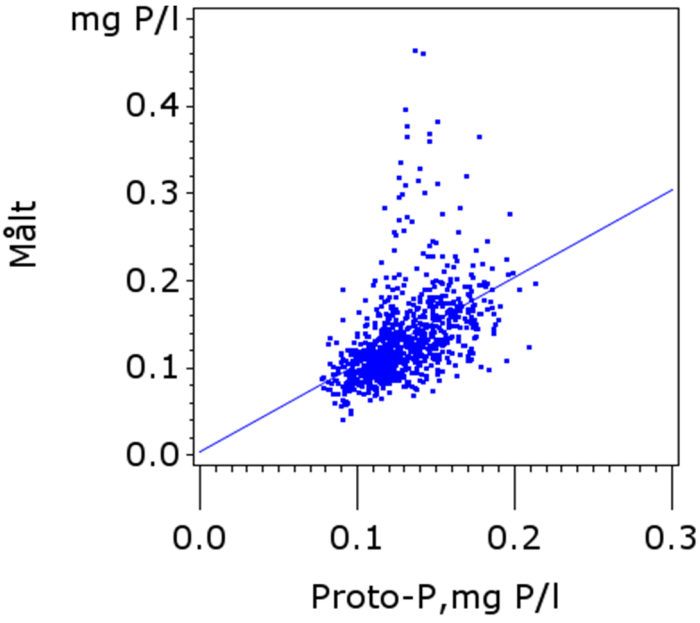
enkelte georegioner rimeligt godt, dog med de største afvigelser mellem model og målt i georegion 1 og georegion 6-9 (figur 15).

Der er gennemført en validering af den udviklede prototype ML TP-models simuleringer mod et uafhængigt datasæt, som modellen ikke er trænet på. Datasættet består af målinger i vandløb fra i alt 83 oplande med et gennemsnitligt oplandsareal på 192 km^2).



Figur 15. Georegioner i Danmark og prototype ML TP-model simulering af årlig vandføringsvægtet TP-koncentration mod de målte koncentrationer i vandløb, der afvander typeoplande inden for ni georegioner i Danmark.

Sammenligningen mellem prototype-modelsimuleringer og de målte vandføringsvægtede årlige TP-koncentrationer i de 83 vandløb er vist i figur 16.



Figur 16. Modelsimulering af årlig vandføringsvægtet TP-koncentration med prototypemodellen sammenlignet med de målte koncentrationer i 83 vandløb.

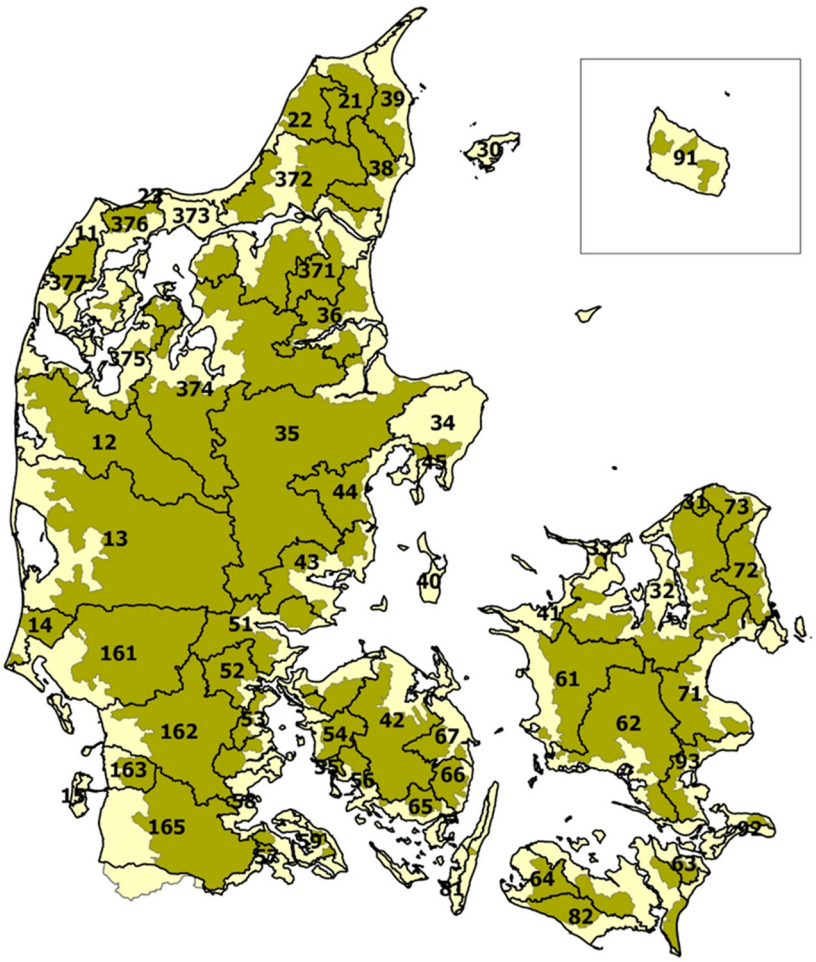
Valideringen af prototype ML TP-model mod målingerne i de 83 vandløb er signifikant og har en hældning på næsten en, men har en forholdsvis ringe forklaringsgrad (Y = 0,98X + 0.008; R^2=0.23, n=863, p<0.0001).

Der er i valideringsplottet flere høje målte koncentrationer, som måske kan stamme fra fosforlækage fra søer mv., som modellen ikke har forudsætninger for at simulere. En nærmere gennemgang af datasættet med udlugning af disse data vil derfor formentlig kunne være med til at øge forklaringsgraden.

Den udviklede prototype ML TP-model er blevet testet på umålte kystoplande i Danmark inden for de i alt 55 2.-3.-ordens kystoplande (figur 17).

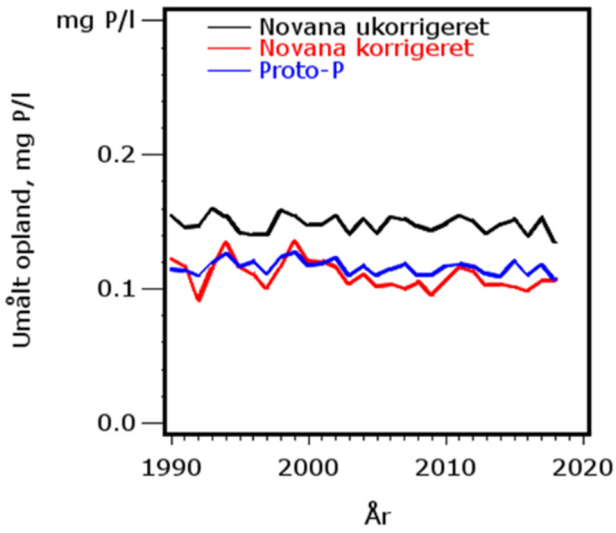
Resultatet af simuleringerne med prototypefosformodellen er for både vandføringsvægtet koncentration og tabet af TP fra diffuse kilder (inkl. spredt bebyggelse) blevet sammenlignet med den nuværende NOVANA DK-P model (bias-korrigeret) (Thodsen et al., 2019) og den tidligere DK-P NOVANA-model (Ovesen et al., 2009). Prototype ML TP-modellens resultater afviger ikke væsentligt fra den nuværende DK-P model på landsresultatet (figur 18 og 19).

Dog giver prototype ML TP-modellen lidt højere beregnede TP-koncentrationer og TP-belastninger end den nuværende NOVANA-model efter 2005 (figur 18 og 19). Prototypemodellens resultater er for beregninger for landet som helhed langt under resultaterne fra den oprindelige (ukorrigerede) DK-QNP model (figur 18 og 19).

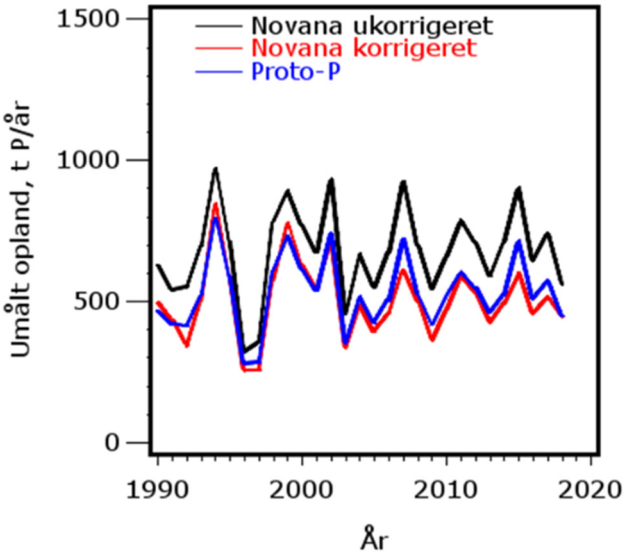


Figur 17. Oversigt over de målte (grønne) og umålte (hvide) arealer i Danmark i forhold til opgørelser af vand- og næringsstofbelastninger i NOVANA (Thodsen et al., 2021). Tallene angiver nummeret på oplandet til 2.-ordens kystafsnit.

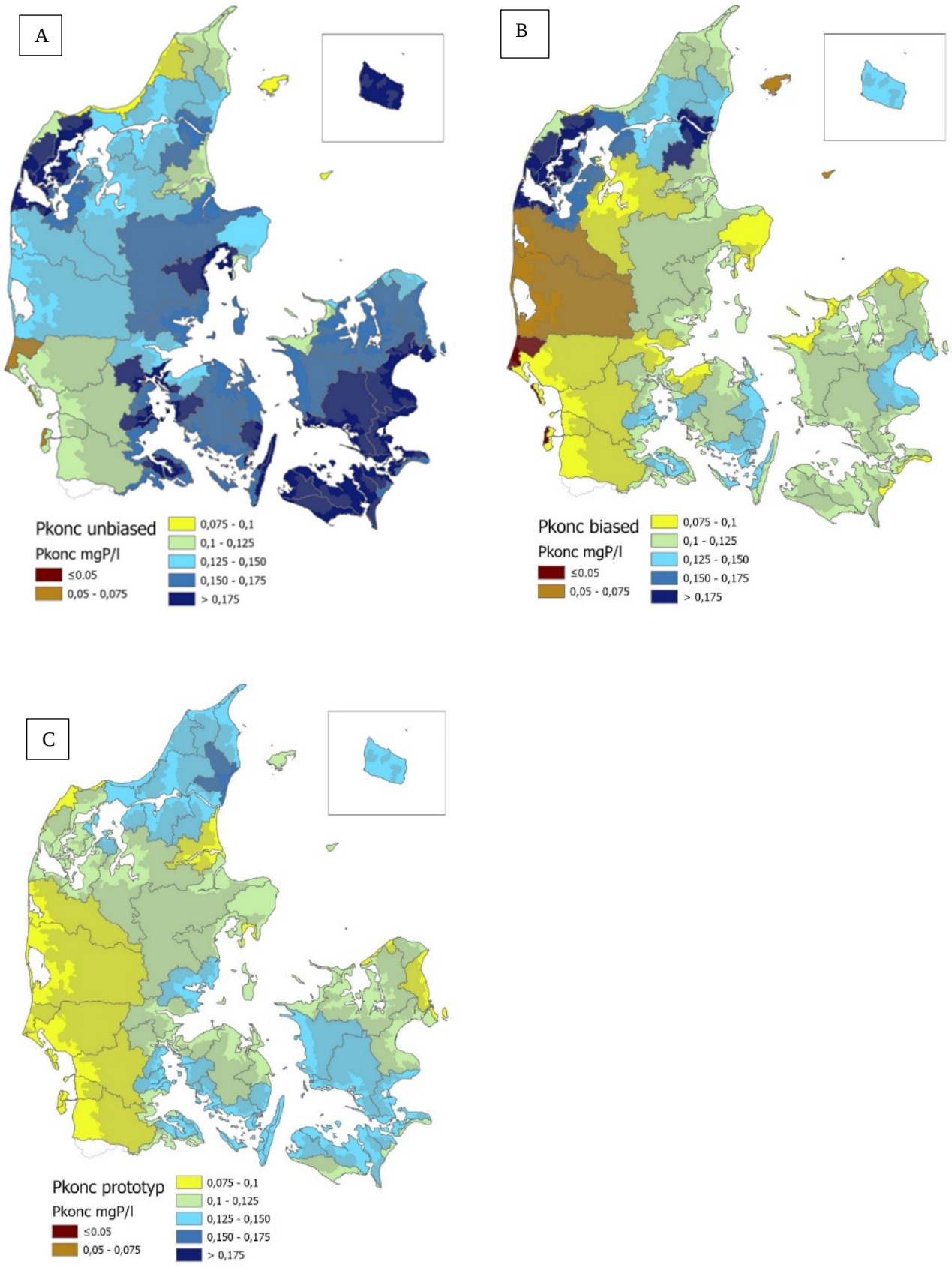
Figur 18. Årlig vandføringsvægtet total fosforkoncentration beregnet som gennemsnit af ID15 for hele landets umålte areal med henholdsvis den rå NOVANA TP-model (DK-QNP ukorrigeret), den nuværende anvendte NOVANA fosformodel (korrigeret) og prototype ML TP-model.



Figur 19. Årlig transport af total fosfor beregnet som sum af ID15 for hele landets umålte areal med henholdsvis den rå NOVANA fosformodel (DK-P ukorrigeret), den nuværende anvendte NOVANA fosformodel (DK-P korrigeret) og prototype ML TP-model.



De regionale forskelle i de modelberegnede gennemsnitlige vandføringsvægtede TP-koncentrationer med de tre modeller kan ses i figur 20.



Figur 20. Gennemsnitlig vandføringsvægtet koncentration af TP i vandløb i perioden 1990-2018 i de forskellige 2.-3.-ordens kystafsnit i Danmark beregnet med den oprindelige ukorrigerede NOVANA-model (A), den nuværende anvendte bias-korrigerede fosformodel (B) og den nye ML prototypefosformodel (C).

Der er tydelige og relativt store regionale skift mellem de simulerede TP-koncentrationer med den oprindelige og den bias-korrigerede NOVANA-model (figur 20). Derimod er der flere ligheder mellem resultaterne fra den anvendte bias-korrigerede NOVANA TP-model og den nye ML prototype TP-model (figur 20). Der sker dog regionale skift i de beregnede vandføringsvægtede TP-koncentrationer, mest udtalt i Thy- og Himmerlandsregionen (figur 20). Generelt modelleres højere koncentrationen af TP i vandløb i de forskellige regioner med den oprindelige NOVANA TP-model end med den bias-korrigerede og de laveste koncentrationer med den nye prototype MS TP-model. Det skyldes, at den oprindelige NOVANA TP-model er opstillet på intensive måledata for TP-koncentrationer (Tornbjerg et al., 2019), mens den bias-korrigede TP-model er tilpasset stikprøvemålinger ved havstationerne i NOVANA (Thodsen et al., 2019). ML prototype TP-modellen er derimod kalibreret på stikprøvedata fra typeoplande.

Den udviklede prototype ML TP-model har i Datarobot en meget god forklaringsgrad (R^2 = 0,65), men i den helt uafhængige validering på 83 typeoplande falder forklaringsgraden markant (R^2 = 0,23). Modellens evne til at simulere TP-koncentrationer i de forskellige georegioner viser sig i flere tilfælde heller ikke at være optimal. Derfor er delkonklusioner fra trin 4:

- At den udviklede prototype ML TP-model i DataRobot har en høj forklaringsgrad på kalibreringsdatasættet, hvilket kunne tyde på, at modellen er blevet overfittet i DataRobot. Det sker, når ML-rutinen ikke stoppes efter nogle indsatte kriterier, men får lov til at køre for længe og derfor opnår maksimal præcision.
- At der i den efterfølgende modeludvikling skal opdateres med flere nye forklarende variable i modeludviklingen og inddrage nye nedbørsdata (grid) fra DMI.
- At der skal gennemføres en sammenligning mellem udvikling af en almindelig regressionsmodel og DataRobot ML TP-modellen på stikprøver fra alle typeoplande.

4.5 Trin 5 – sammenligning af en udviklet regressionsmodel i SAS contra DataRobot ML-01TP-model

I foråret 2021 modtog DCE, AU, nye reviderede grid-nedbørsdata fra DMI. Derefter blev der også oparbejdet en del nye forklarende variable til brug for udvikling af en TP-model, som det fremgår af oversigten i tabel 7. De nye variable er blevet udregnet for alle ca. 3200 ID15-oplande og derefter for deloplande, der indgår i modellen til kalibrering og validering. Data er hentet fra forskellige kilder (se tabel 7 og bilag 1).

Til modelkalibrering blev der anvendt data fra i alt 114 typeoplande uden større søer og udledning fra punktkilder (N=2040) (figur 21). Desuden indgik et uafhængigt datasæt fra typeoplande i valideringen af de udviklede modeller, som består af op til 300 oplande med langt de fleste i de seneste år pga. de nye Landbrugspakke-målestationer, som blev oprettet fra 2017 (figur 21).

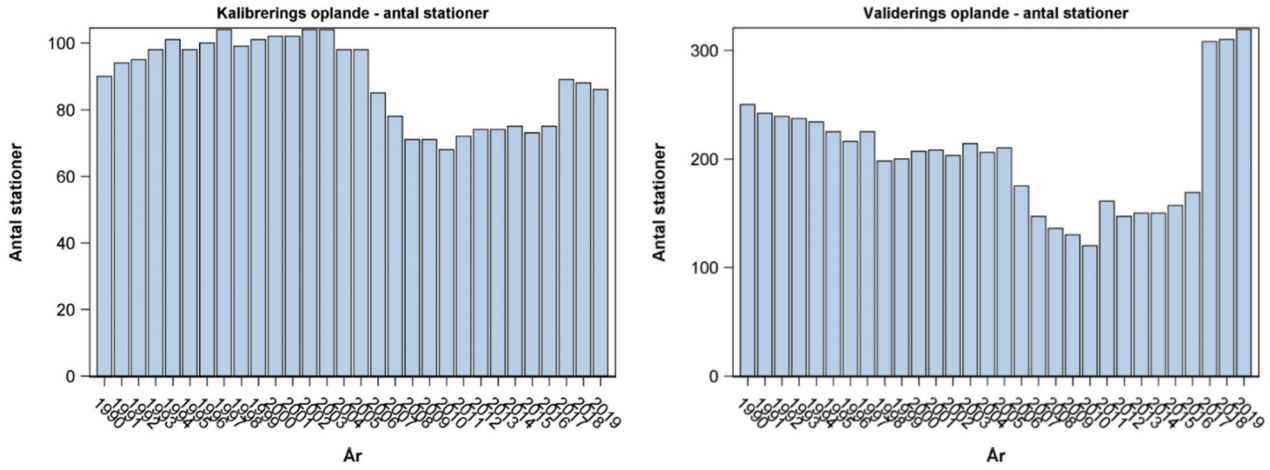
TP-koncentrationen i modelkalibrering og -validering blev afskåret, så kun TP-koncentrationsdata <0,20 mg P/l indgik for at undgå få høje 'outliers' i modeludviklingen. Disse outliers kan skyldes flere forhold som: 1) for usikre oplysninger om udledninger fra punktkilder i oplandet; 2) at der tilfældigt er målt en meget høj TP-koncentration under en ekstrem vandføring i vandløbet,

som vil øge TP-transporten betydeligt. Data i kalibreringen blev ligeledes afskåret, så kun årlige data, hvor betydningen af punktkilder (uden spredt bebyggelse) var mindre end < 25 % i forhold til den målte transport, blev medtaget i modelkalibreringen.

Efter den indledende dataudtynding indgik der et antal oplande med måledata i de enkelte år til modelkalibrering og -validering, som det fremgår af figur 21.

Tabel 7: Oversigt over de forskellige forklarende variable, som brutto har indgået i modelkalibreringer i 2021 (se bilag 1 for referencer). Alle parametre til modellen er etableret for de ca. 3200 ID15-oplande, som dækker landet.

Parameter i opland	**Enhed**	**Type**	**Betydning**	**Datagrundlag/reference**
Dræningsgrad (DRÆN)	%	Stationær	Hvor stor en andel er drænet (potentielt)	Drained_v3 potentialekort Agroøkologi, AU
Dyrkningsgrad (DYRK)	%	Årlig	Hvor stor en andel er dyrket	Markblokkort og beregnet fra 1990-2019
Befæstningsgrad (BEFÆST)	%	Årlig	Hvor stor en andel er befæstet	Basemap
Sandjordsandel (SAND)	%	Stationær	Andel af sandjord fra 3 øvre horisonter	JB-jordtypekort i 3 horisonter – gennemsnit
Lerjord i underjord (LER_U)	%	Stationær	Andel af lerjord i underjord (D-horisont)	JB-jordtypekort – dybeste horisont 1-2 m
Nedbørsafvigelse (NEDBVAFVREL)	%	Årlig	Afvigelse fra middelnedbør i opland for hele perioden	DMI
Frostdøgn (FROSTD)	Antal	Årlig	Antal frostdøgn (lufttemperatur mindre end 0 ^{o}C)	DMI
Brinkerosion (BE)	kg P/ha	Stationær	Tab fra brinkerosion i oplandet	Andersen og Heckrath (2020)
Makroporer (MAKRO)	kg P/ha	Stationær	Tab fra makroporer i marken i oplandet	Andersen og Heckrath (2020)
Matriceudvaskning (MATRIX)	kg P/ha	Stationær	Tab fra matrice udvaskning i oplandet	Andersen og Heckrath (2020)
Vanderosion (VERO)	kg P/ha	Stationær	Tab fra jorderosion i oplandet	Andersen og Heckrath (2020)
Overfladisk afstrømning (OVERF)	tons sediment	Stationær	Tab af sediment til vandløb fra overfladisk afstrømning i oplandet	Heckrath (2020)
Diffust tab (DIFF)	Kg P/ha	Stationær	Samlet diffust tab af fosfor fra oplandet	Andersen og Heckrath (2020)
Organiske jorder (ORG)	%	Stationær	Andel af opland med arealer med > 3 % kulstodindhold	Kulstof_3-6_12 – Agroøkologi, AU
BFI	Tal (0-1)	Årlig	Baseflow-indeks beregnet med en delmodel (ligning 2)	Ecoscience, AU
Korrigeret nedbør (NEDB)	mm	Årlig	Korrigeret nedbør (DMI)	DMI
Georegioner (GEO1-9)	Tal (1-9)	Klasse	Georegioner i Danmark	Agroøkologi, AU
Udledning fra spredt bebyggelse (SPRB)	kg P/ha	Årlig	Sum udledning i oplandet	PULS databasen
Årsmiddel lufttemperatur (TEMP)	^{o}C	Årlig	Årlig gennemsnit temperatur i opland	DMI



Figur 21. Antallet af vandløbsstationer, som indgår i henholdsvis kalibrerings- og valideringsdatasættet for TP-modellerne udviklet i DataRobot (ML-01) og SAS.

Resultatet af den udviklede regressionsmodel i SAS blev en model med en forklaringsgrad (R^2) på 0,30 og med i alt 11 forklarende variable (tabel 8).

Tabel 8. Den opstillede regressionsmodel for årlig vandføringsvægtet TP-koncentration i SAS.

Parameterestimater

Variabel	DF	Parameter Estimat	Standard Error	t Value	Pr > |t|
Intercept	1	-2,49389	0,11805	-21,13	<,0001
TEMP	1	-0,05375	0,01042	-5,16	<,0001
DYRK	1	0,00266	0,00048512	5,49	<,0001
DRÆN	1	0,00635	0,00069125	9,18	<,0001
BE	1	0,15584	0,08565	1,82	0,0690
MAKRO	1	0,78680	0,21257	3,70	0,0002
VERO	1	-2,94818	0,52894	-5,57	<,0001
BEFÆST	1	0,03152	0,00270	11,65	<,0001
ORG	1	0,00690	0,00092274	7,48	<,0001
BFI	1	-0,50947	0,07160	-7,12	<,0001
NEDBAFVREL	1	0,00264	0,00052609	5,02	<,0001

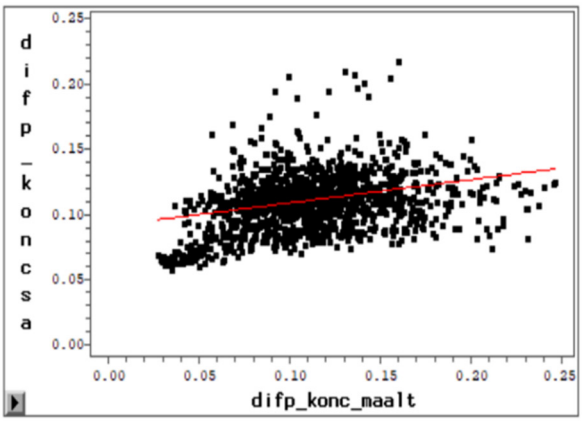
Et samplot mellem modelestimater af TP-koncentration med den nye regressionsmodel og de målte koncentrationer viser dog, at der er en ringe forklaringsgrad (R^2 = 0,099; P<0,001) med en stor afgivelse mellem målte og modellerede TP-koncentrationer (figur 22).

type = Model opland

difp_koncsas_ferg = difp_konc_maalt	
Response Distribution:	Normal
Link Function:	Identity

Model Equation

difp_koncsas_ferg = 0.0911 + 0.1781 difp_konc_maalt



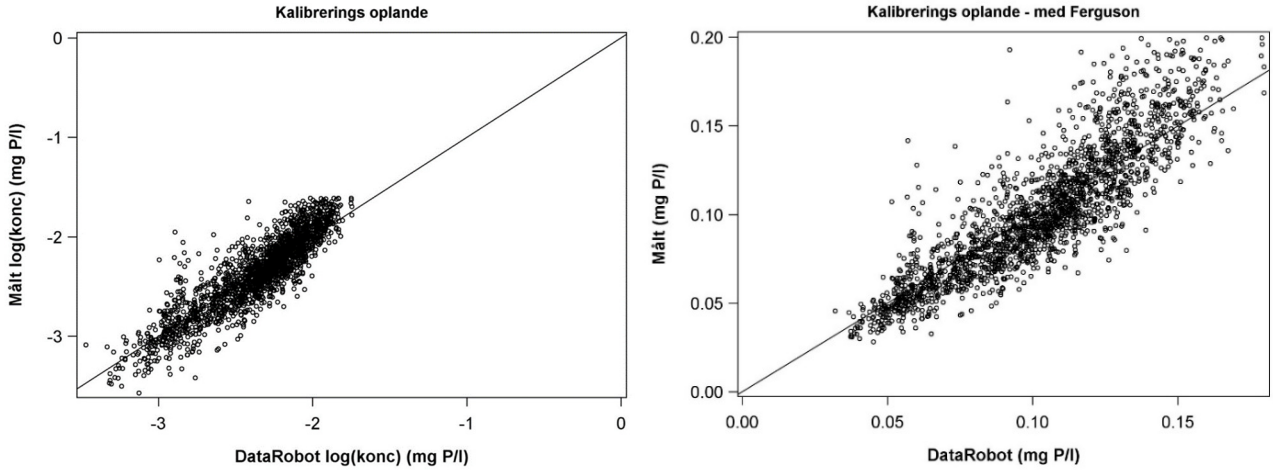
Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Model Mean Square	Error DF	Error Mean Square	R-Square	F Stat	Pr > F
	1	1	0.0597	1288	0.0004	0.0997	142.63	<.0001

Figur 22. Resultat af regressionsmodellen udviklet i SAS-software, der har en forklaringsgrad på R^2 = 0,099 på valideringsdatasættet.

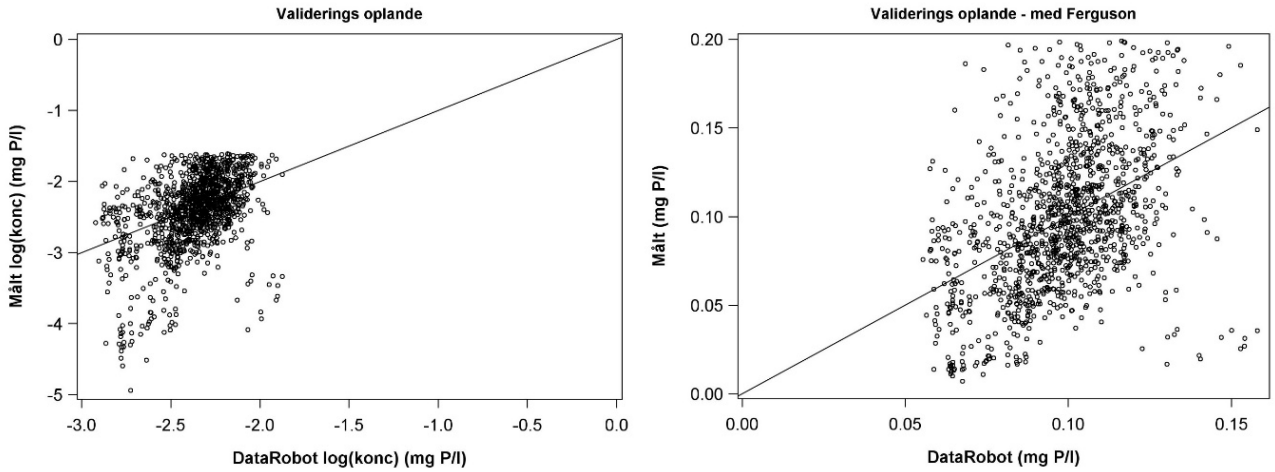
Den udviklede ML-01 TP-model i DataRobot på samme datamateriale fra typeoplande som regressionsmodellen viser en meget større forklaringsgrad (R^2 = 0,80 på log-transformere data og R^2 = 0,77 på tilbagetransformerede data med en Ferguson-korrektion (figur 23). I DataRobot-modellen indgår i alt 10 forklarende variable (se tabel 7 og bilag 4):

Ln(TP konc) = f(DRÆN; BEFÆST; BFI; DYRK; MAKRO; ORG; TEMP; BE; NEDBAFVREL; VERO) (se tabel 7 for nærmere beskrivelse af variable).



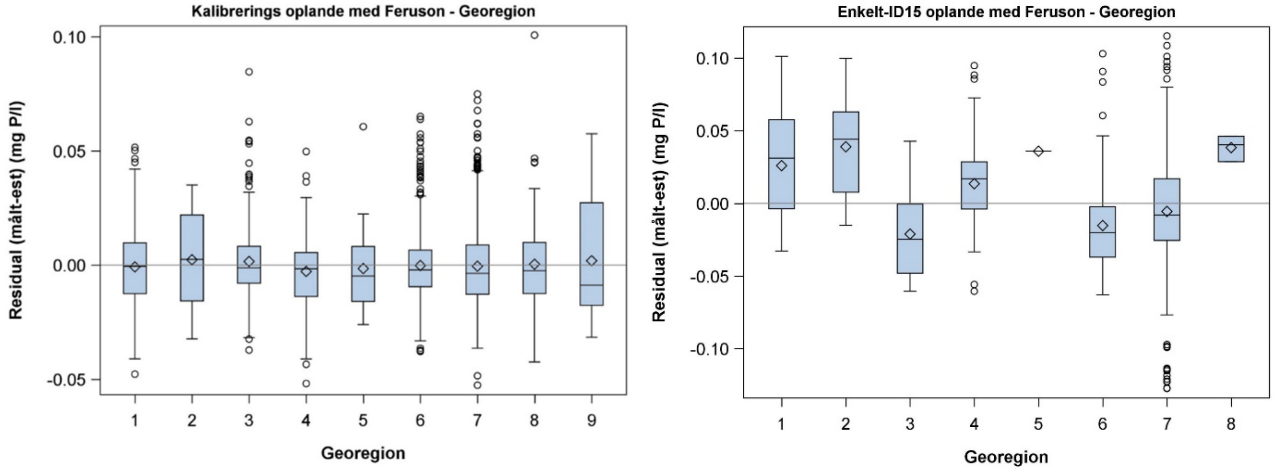
Figur 23. Sammenhæng mellem DataRobot ML-01 TP-model simuleret TP-koncentration (log og tilbagetransformeret med efterfølgende Ferguson korrektion) for alle kalibreringsoplande, for alle tre datasæt i DataRobot i modeludviklingen (N= 2040).

En validering af den udviklede DataRobot ML-01 TP-model på et 'uafhængigt' datasæt af måledata fra typeoplande, som er udeladt fra kalibreringen i Datarobot, viste dog et meget dårligere resultat end i modelkalibreringen. Forklaringsgrad på logaritmetransformerede data er dårlig (R^2 = 0,21, og det samme gælder for de tilbagetransformerede data med Ferguson-korrektion (R^2 = 0,19) (figur 24).



Figur 24. Sammenhæng mellem DataRobot-model ML-01 simuleret TP-koncentration (log- og tilbagetransformeret med efterfølgende Ferguson korrektion) for alle kalibreringsoplande for det uafhængige valideringsdatasæt af oplande (N=1261).

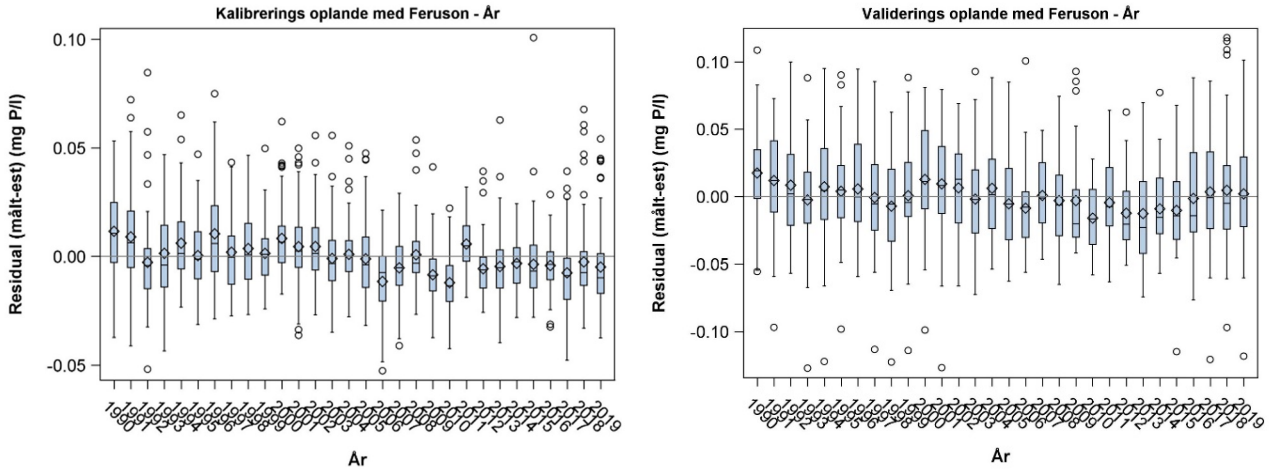
Residualplot for kalibreringsdata og valideringsdata med den udviklede ML-01 TP-model i DataRobot er vist i figur 25. Der er en god overensstemmelse mellem modelsimulerede og målte TP-koncentrationer i de ni georegioner for kalibreringsdatasættet, med de største afvigelser i georegionerne 5-9 (figur 25). Men på valideringsdatasættet viser TP-modellen en meget dårlige performance (figur 25).



Figur 25. Residualplot af forskel målt mod simuleret TP-koncentration med den nye ML-01 TP-koncentrationsmodel inden for ni georegioner i Danmark vist for logaritme-transformerede data for kalibreringsdatasættet (venstre) og valideringsdatasættet (højre), begge med tilbagetransformerede data med Ferguson-korrektion. (Diamant er gennemsnit, streg er median).

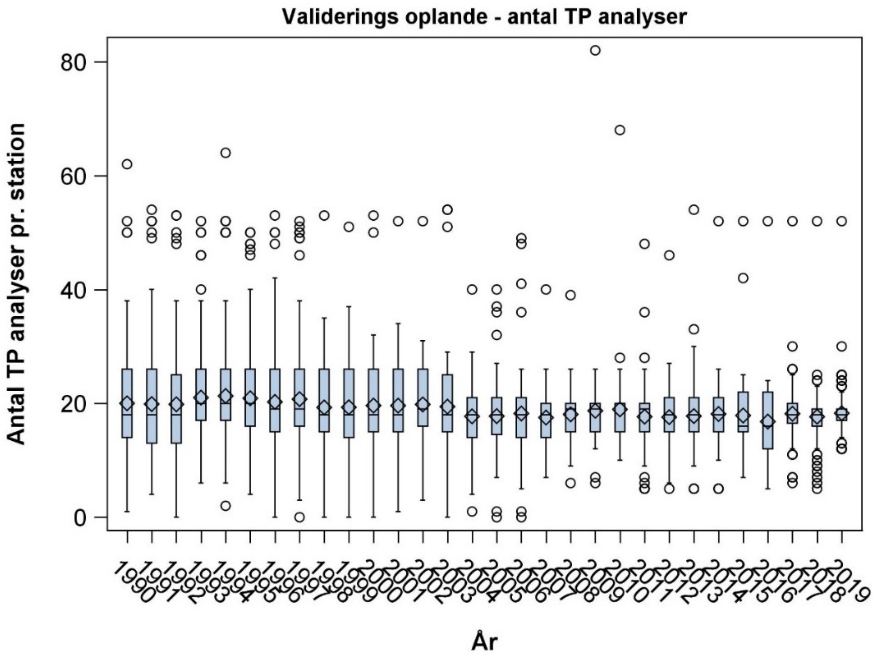
Dog udviser residualplottet for de årligt simulerede TP-koncentrationer et mønster, hvor modellen overestimerer de målte koncentrationer efter 2004 både for kalibrerings- og valideringsdata (figur 26). Dette kan forklares ved, at der skete en revision af NOVA-programmet i 2003, hvor antallet af årlige stikprøver pr. vandløbsstation faldt (figur 27). Det vides fra tidligere analyser,

at lavere målefrekvens i gennemsnit medfører en stigende underestimering af den 'sande' TP-koncentration (Kronvang og Bruhn, 1996).



Figur 26. Residualplot af forskel målt mod simuleret TP-koncentration med ML-01 TP-model for simulering af TP-koncentrationen i perioden, hvor modellen er testet (1990-2019), vist for henholdsvis kalibreringsdatasættet og valideringsdatasættet, begge med tilbagetransformerede data med Ferguson-korrektion.

Figur 27. Statistisk oversigt over antallet af stikprøvetagninger ved målestationerne i typeoplande i årene 1990-2019. Bemærk, at antallet falder efter 2003, hvor der skiftes fra NOVA til det nye NOVANA-overvågningsprogram (diamant er gennemsnit, og streg er median. Bokse er henholdsvis 25 % og 75 % kvartil).



Delkonklusioner fra trin 5 er:

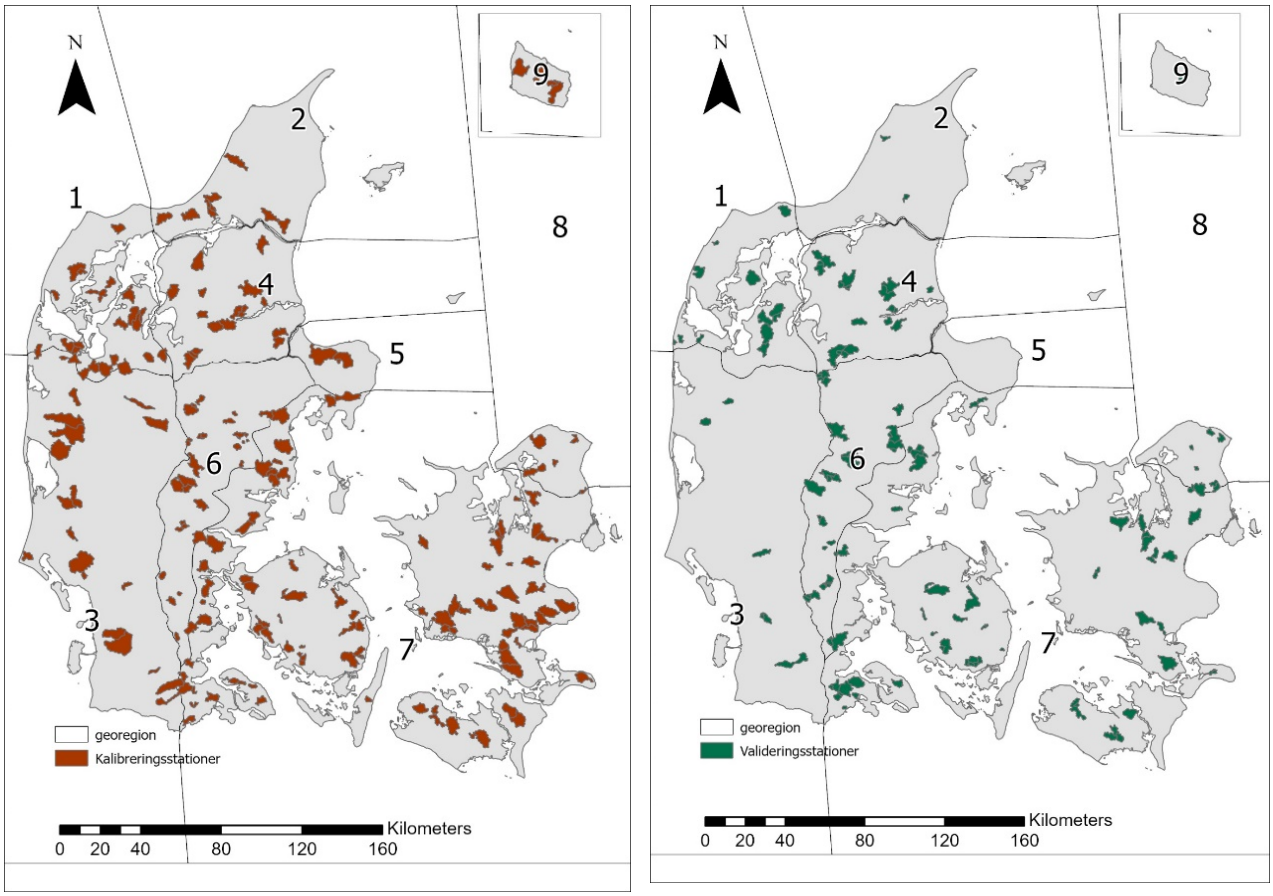
- At der på baggrund af ovenstående resultater skal ske en rekalibrering af DataRobot ML-01 TP-modellen med inddragelse af en periodeparameter for ændringen i stikprøveantal (to perioder: 1990-2003 & 2004-2019).
- At det skal undersøges, om der er behov for at inddrage en parameter for perioden 2009-2015, hvor der ved analyse af TP-koncentration i vandløbsprøver er benyttet UV-metode til oplukning i stedet for den anbefalede autoklavemetode, hvilket sandsynligvis medføre en underestimering af den målte TP-koncentration (Larsen et al., 2020).

- At det skal undersøges, om der skal inddrages en bedre rumlig parameter end georegion for beliggenhed af de enkelte oplande (X, Y-UTM centrumskoordinater).
- At det skal forsøges at substituere BFI-variablen som forklarende variabel i modellen med de inputdata, som anvendes i denne delmodel (se ligning 2).

4.6 Trin 6 – udvikling af endelig DataRobot ML-06 TP-model

I de endelige modelkalibreringer i DataRobot er der arbejdet med i alt fem udviklede versioner af ML TP-modeller. Undervejs blev der i de forskellige modelversioner flyttet flere oplande fra valideringsdatasættet uden for DataRobot til kalibreringsdatasættet for at opnå bedre modelsimuleringer inden for de enkelte ni georegioner. Desuden har der været arbejdet med modeller i DataRobot af forskellig type og med forskellige nye forklarende variable såsom en periode før og efter ændring i stikprøvetagning, UV-metode periode, X,Y koordinat for oplandes beliggenhed, sandjordsprocent mv.

Data blev, som i trin 5, afskåret, så kun data med TP-koncentrationer <0,20 mg P/l blev inddraget (reducerer datasæt med ca. 7 %, og TP-udledning fra punktkilder (uden spredt bebyggelse) skulle være < 25 % af TP-tabet hvert år i perioden (reducerer kalibreringsdatasæt med yderligere 17 %). Der arbejdes igen med en logaritmetransformeret model. Oplandene i henholdsvis kalibreringsdatasættet og valideringsdatasættet kan ses i figur 28. Der indgår efter beskæring af datasæt fra de i alt 207 vandløbsoplande 2389 observationer i kalibreringsdatasættet og data fra i alt 142 vandløbsoplande i det uafhængige valideringsdatasæt (N=1261).



Figur 28. Beliggenhed af oplande i det endeligt anvendte kalibreringsdatasæt (N=207) (venstre figur) og valideringsdatasæt (N=142) (højre figur).

De to grupper data til kalibrering og validering uden for DataRobot-modellen er meget ens i forhold til de enkelte forklarende variable, som er anvendt i modelkalibreringen (tabel 10). Således er det gennemsnitlige oplandsareal i henholdsvis kalibrerings- og valideringsdatasættet på 22,3 km^2 og 17,7 km^2 og dyrket areal på 59,3 % og 53,0 % (tabel 9).

Tabel 9: Oversigt over gennemsnit og variation i de forklarende variable i den endelige TP-model for henholdsvis kalibreringdatasættet i DataRobot og valideringsdatasættet uden for DataRobot (i parentes er standardafvigelsen angivet).

	Kalibreringsoplande (N = 207)	**Valideringsoplande (N = 142)**
Oplandsareal (km^2)	22,3 (±17,1)	17,7 (±13,4)
Dyrkningsgrad (%)	59,3 (±18,3)	53,0 (±21,3)
Befæstet areal (%)	10,4 (±5,0)	10,0 (±6,2)
Andel af leret underjord (%)	12,6 (±4,9)	11,8 (±4,9)
Drænet areal (%)	55,4 (±16,1)	52,9 (±15,9)
Andel med organisk jord (>3 %) (%)	11,9 (±8,4)	12,9 (±10,1)
P-tab ved Brinkerosion (kg P/ha)	0,12 (±0,12)	0,081 (±0,11)
P-tab ved makropore flow (kg P/ha)	0,055 (±0,010)	0,015 (±0,012)
P-tab ved matrice udvaskning (kg P/ha)	0,015 (±0,010)	0,015 (±0,012)
P-tab ved vanderosion (kg P/ha)	0,017 (±0,013)	0,015 (±0,014)
Korrigeret årlig nedbør (mm)	853 (±101)	852 (±113)
Nedbørsafgivelse fra middel (%)	6 (±43)	-12 (±56)
Antal frostdøgn	27 (±8)	31 (±10)

Kalibrering af TP-model

Den endeligt udviklede model i DataRobot – ML-06 TP-modellen – er af typen 'eXtreme Gradient Boosted Trees Regressor with early stopping' og indeholder de forklarende variable:

Ln(TP konc) = f(DRÆN; BEFÆST; PER; DYRK; BE; NEDBAFVREL; SAND; ORG; XUTM; YUTM; LERU; FROST; NEDB) (se tabel 10 for nærmere beskrivelse af variable).

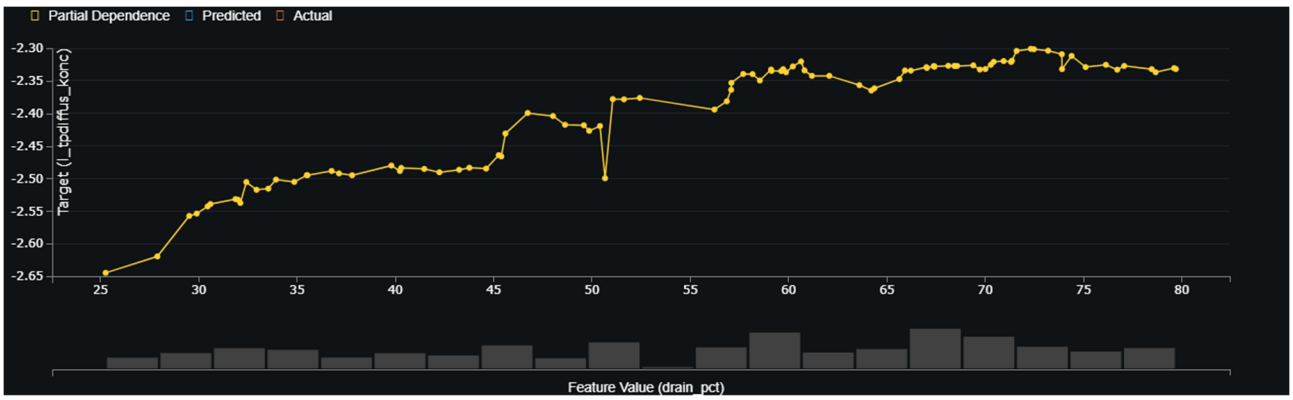
De tre lag i udviklingen af modellen har følgende forklaringsgrader: Træningsdatasættet, som består af 64 % af data (R^2 = 0,69), valideringsdatasættet, som består af 16 % af data (R^2 = 0,71), og holdout-datasættet med 20 % af data (R^2 = 0,67).

I figur 26-32 er betydningen af de forklarende variable, som har mest indflydelse på simulering af den årlige vandføringsvægtede TP-koncentration, vist. Desuden er der i bilag 5 indsat figurer, der viser betydningen af de enkelte variable i ML-06 TP-modellen for simulering af logaritmetransformerede TP-koncentrationer.

Tabel 10: Oversigt over de forklarende variable, som indgår i den endelige ML-06 TP-modelversion sorteret efter deres forklaringsgrad i modellen. Alle variable er beregnet for alle ca. 3200 ID15-oplande i landet og, hvis nødvendigt, for hvert år i perioden 1990-2019.

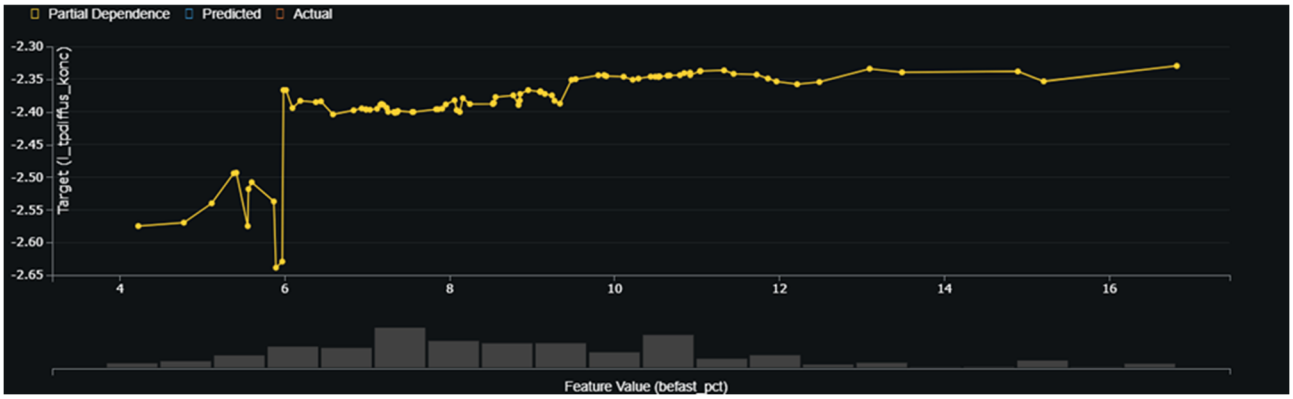
Parameter i opland	Enhed	Type variabel	Vigtighed af variabel i model
Dræningsgrad (DRÆN)	%	Stationær	1
Befæstningsgrad (BEFÆST)	%	Årlig	2
Periode (PER)	1: 1990-2003; 2: 2004-2019	Klasse	3
Dyrkningsgrad (DYRK)	%	Årlig	4
Brinkerosion (BE)	kg P/ha	Stationær	5
Nedbørsafvigelse (NEDBAFVREL)	%	Årlig	6
Sandjords andel (SAND)	%	Stationær	7
Organiske jorder (ORG)	%	Stationær	8
Centrum X UTM (XUTM)	-	Stationær	9
Centrum Y UTM (YUTM)	-	Stationær	10
Lerjord i underjord (LERU)	%	Stationær	11
Frostdøgn (FROST)	Antal	Årlig	12
Korrigeret nedbør (NEDB)	mm	Årlig	13

Betydningen af dræningsgrad i oplandene i simulering af TP-koncentrationen med DataRobot ML-06 modellen, når alle andre variable holdes fast, fremgår af figur 30. Figuren viser, at der med stigende dræningsgrad simuleres en generelt stigende vandføringsvægtet TP-koncentration. Da dræning med drænrør i marken øger risikoen for transport af både opløst og især partikelbundet fosfor til vandløb, er det i god tråd med vores viden om fosfortabsprocesser (Andersen og Heckrath, 2020).



Figur 30. Betydning af dræningsgrad i oplandet for simulering af TP-koncentrationen i modellen. (Y-aksen er logaritmisk (ln), og Y-aksens range er 71-100 µg P/l). Grå søjler under figuren er antal observationer.

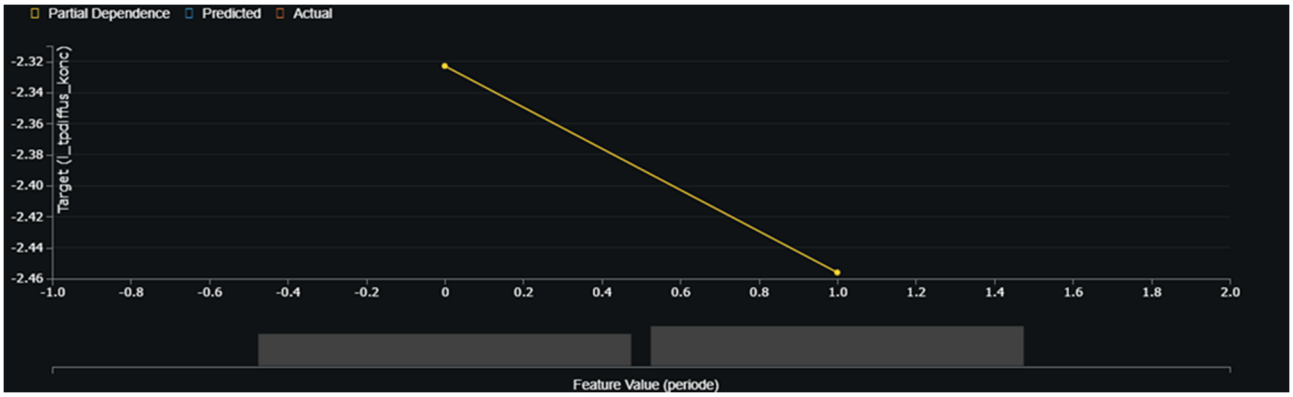
Betydningen af befæstede arealer i oplandene for simulering af TP-koncentrationen med DataRobot ML-06 TP-modellen, når alle andre variable holdes fast, fremgår af figur 31. Figuren viser, at der overordnet er en stigende TP-koncentration med stigende befæstningsgrad, men at der ved en befæstningsgrad på 6 % er et spring i den simulerede vandføringsvægtede TP-koncentration til et højere niveau. Derefter er der kun en svagt stigende betydning af befæstningsgrad i modellen. Springet ved omkring de 6 % i befæstningsgrad er formentligt forårsaget af meget få observationer under 6 % befæstningsgrad, og af at der omkring dette niveau måske sker et skift fra naturvandløb til vandløb, der generelt afvander dyrkede arealer. Om befæstningsgraden alene kan medføre det viste spring i TP-koncentrationen, er uvist. En mulighed er, at der er en stigende risiko for forbindelser til vandløb fra befæstede arealer ved et vist omfang befæstet areal i et opland. Stigende befæstningsgrad kan derfor være en påvirkning fra regnvandsbetingede overløb i oplandet.



Figur 31. Betydning af andel befæstet areal i oplandet for simulering af TP-koncentrationen i modellen. (Y-aksen er logaritmisk (ln) og spænder over et range på 71-100 µg P/l).

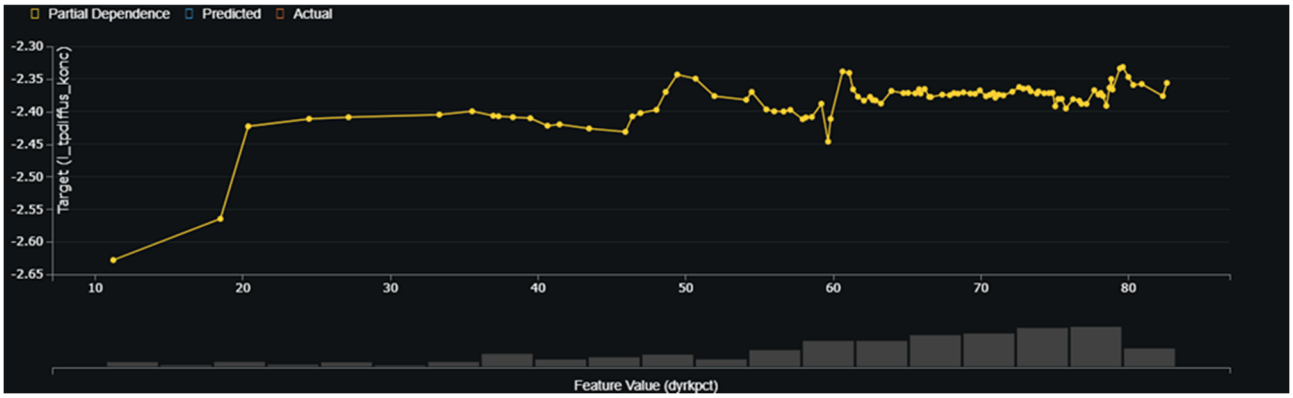
Periodeparameteren forklarer skiftet i frekvens af stikprøvemålinger mellem perioden 1990-2003 (NOVA) og perioden 2004-2019 (NOVANA). I den sidste periode er der generelt udtaget færre årlige stikprøver til analyse af TP-koncentration end i den første periode (se figur 32). Forskellen mellem de to perioder er ca. 15 µg P/l med laveste koncentration i den seneste periode (2004-2019). Andre faktorer såsom bedre rensning af spildevand kan dog ikke helt afvises også at have betydning, som f.eks. rensning af spildevand fra spredt bebyggelse. Men det abrupte skift taler for, at det observerede spring hænger sammen med skiftet i antal udtagne vandprøver pr. år mellem de to perioder.

Der blev også forsøgt med en parameter for anvendelse af UV-metoden til analyse af TP (2009-2015), men denne viste sig ikke at kunne indarbejdes i den endelige ML TP-model, da den ikke indgik signifikant. Det betyder, at eventuelle homogenitetsbrud ved anvendelse af UV-metoden i perioden 2009-2015 for analyse af TP ikke ser ud til at have medført et homogenitetsbrud, der kan erkendes af analyserne i DataRobot.



Figur 32. Betydning af periodeparameter for simulering af TP-koncentrationen i modellen. (Y-aksen er logaritmisk (ln) og spænder over et range på 85-98 µg P/l).

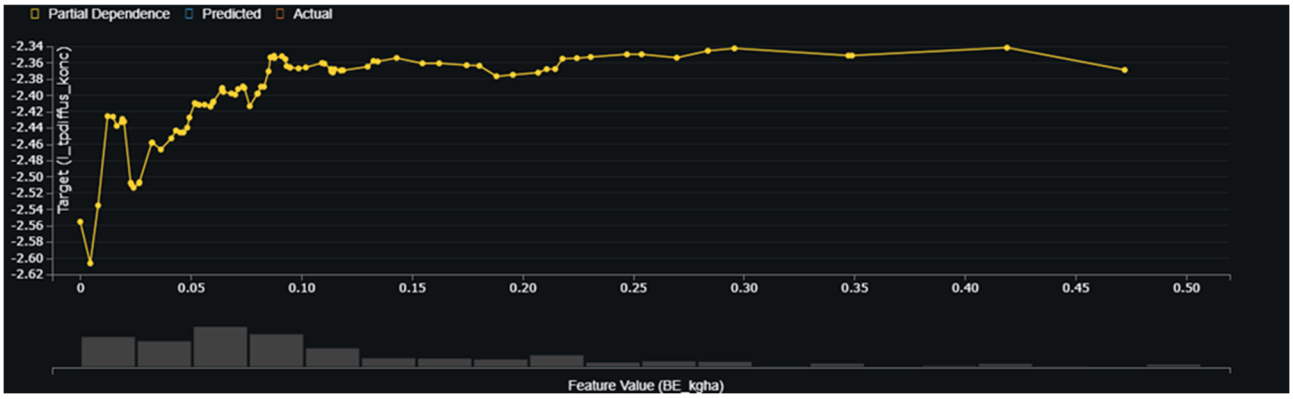
Betydningen af dyrkningsgrad i oplandene for simulering af TP-koncentrationen med DataRobot ML-06 TP-modellen, når alle andre variable holdes fast, fremgår af figur 33. Figuren viser, at der med stigende dyrkningsgrad – især ved lave dyrkningsgrader (0-20 %) – simuleres en stigning i vandføringsvægtet TP-koncentration, mens der herefter ved stigende dyrkningsgrader over 20 % kun er meget lille respons. Denne respons skal selvfølgelig samtolkes med effekten af dræningen, vist i figur 30, da dyrkede arealer med ler i underjord oftest er drænet. Den bratte øgning af fosforkoncentrationen omkring en dyrkningsgrad på 20 % indikerer, at selv ved lave dyrkningsgrader vil en overgang fra naturopland til, at en mindre del af oplandet er dyrket, give anledning til en øget fosforkoncentration. Den mindre koncentrationsforøgelse ved stigende dyrkningsgrad efter springet omkring en dykningsandel på 20 % kan muligvis skyldes, at dyrkningen har medført en nettotilførsel af fosfor over tid, hvilket har øget jordens P-indhold og dermed risikoen for fosfortab med de forskellige transportveje (Andersen og Heckrath, 2020).



Figur 33. Betydning af dyrkningsgrad i oplandene for simuleringen af TP-koncentrationen i modellen. (Y-aksen er logaritmisk (ln) og spænder over et range på 71-100 µg P/l).

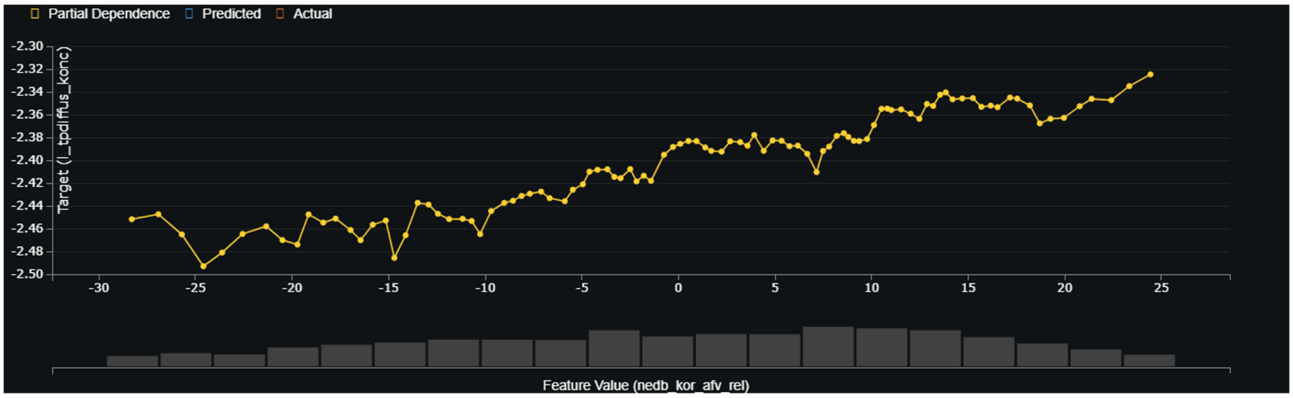
Betydningen af brinkerosion for tilførsel af TP til vandløb i oplandene for simulering af TP-koncentrationen med DataRobot ML-06 TP-modellen, når alle andre variable holdes fast, fremgår af figur 34. Der ses en ret stor, tilnærmelsesvis lineær, stigning i TP-koncentration ved stigende tab af fosfor fra brinkerosion indtil en grænse på omkring 0,1 kg P / ha tab fra brinkerosion. Herefter ændres TP-koncentrationen ikke ved mere brinkerosion. Den manglende respons ved stigende risiko for fosfortab ved brinkerosion over 0,1 kg P / ha, kan

muligvis skyldes, at TP-koncentrationen stammer fra målinger med stikprøver. Det er kendt at stikprøvetagning ikke fanger især transporten af partikelbundet fosfor ligeså godt som intensiv prøvetagning (Tornbjerg et al., 2019).



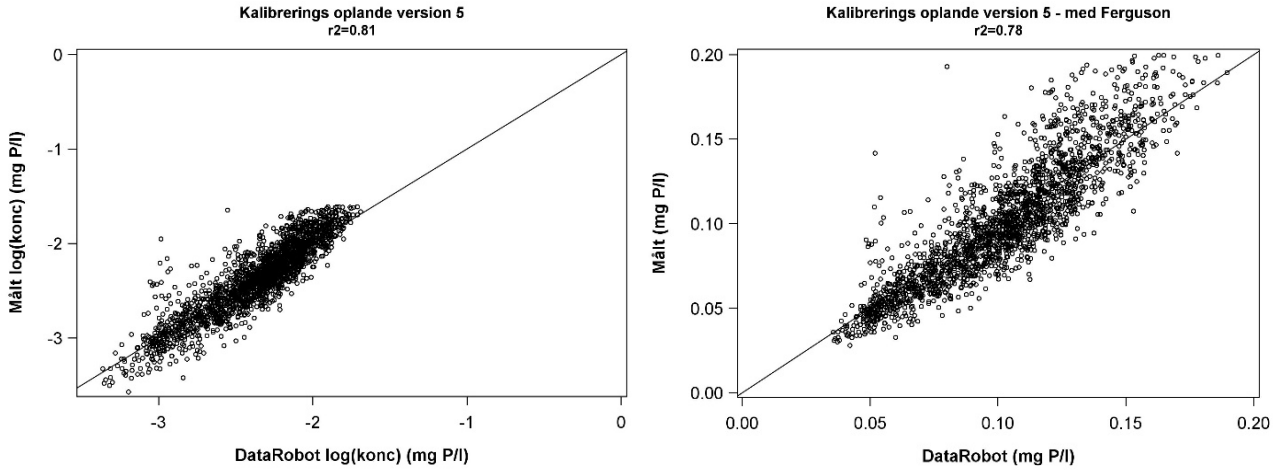
Figur 34. Betydning af brinkerosion (kg P/ha) for simulering af TP-koncentrationen i den endelige model. (Y-aksen er logaritmisk (ln) og spænder over et range på 73-96 µg P/l).

Betydningen af den relative afvigelse af årsnedbøren fra en langtidsmiddel (1990-2018) er vist i figur 35. Der er generelt en stigning i den simulerede TP-koncentration med stigende nedbørsafvigelse, især når årets nedbør er større end gennemsnitsnedbøren (positiv afvigelse). I tørre år ses der ikke en tilsvarende stigende TP-koncentration op mod gennemsnitsnedbøren, hvilket formentlig kan forklares ved, at der kun er få transportveje, som bidrager til TP-koncentrationen i vandløb i meget tørre år. Ved stigende nedbør over gennemsnittet åbnes muligheden for, at der kommer bidrag til vandløbet af TP fra flere transportveje såsom dræn, erosion og overfladisk afstrømning, mere brinkerosion mv.



Figur 35: Betydning af den relative afvigelse fra årets middelnedbør i perioden for simulering af TP-koncentrationen i modellen. (Y-aksen er logaritmisk (ln) og spænder over et range på 82-100 µg P/l).

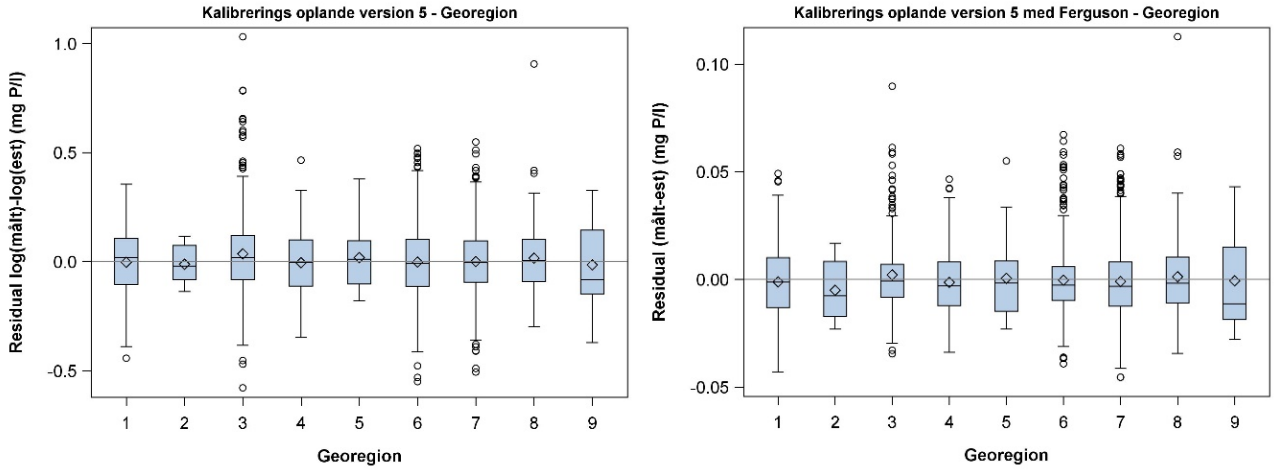
Kalibreringsresultatet af modellen er vist for simuleret TP-koncentration mod målt TP-koncentration både for logaritmetransformerede og tilbagetransformerede data i figur 33. Der er en meget god forklaringsgrad (R^2) på 0,81 for de logaritmetransformerede data og en forklaringsgrad på 0,78 for de tilbagetransformerede data med Fergusonkorrektion (figur 36).



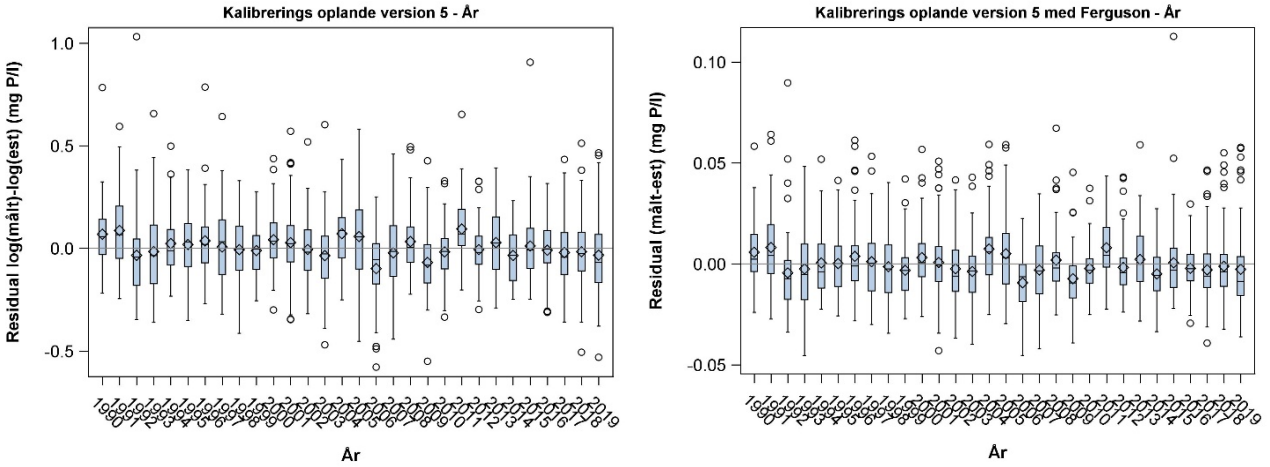
Figur 36: Sammenhæng mellem DataRobot-modelsimuleret TP-koncentration (log- og tilbagetransformeret) for alle kalibreringsoplande for alle tre datasæt inkluderet i DataRobot-modeludviklingen (N=2389).

Residualplot af den kalibrerede ML-06 TP-model i DataRobot for simulering mod målte TP-koncentrationer i de ni georegioner for både de simulerede logaritmetransformerede resultater og de tilbagetransformerede kan ses af figur 37). TP-koncentrationen er dårligst simuleret i georegion 9, som er Bornholm (figur 37). Her indgår også kun få oplande i modeludviklingen, og undergrunden er væsentlig anderledes end i resten af landet.

De tilsvarende residualplot for de enkelte år i kalibreringsperioden er vist i figur 38 for både logaritme- og tilbagetransformerede data for TP-koncentrationen.



Figur 37. Residualplot af forskel mellem målt og simuleret TP-koncentration med den endelige ML-06 TP-koncentrationsmodel inden for ni georegioner i Danmark vist for logaritme-og tilbagetransformerede data med Fergusonkorrektion.

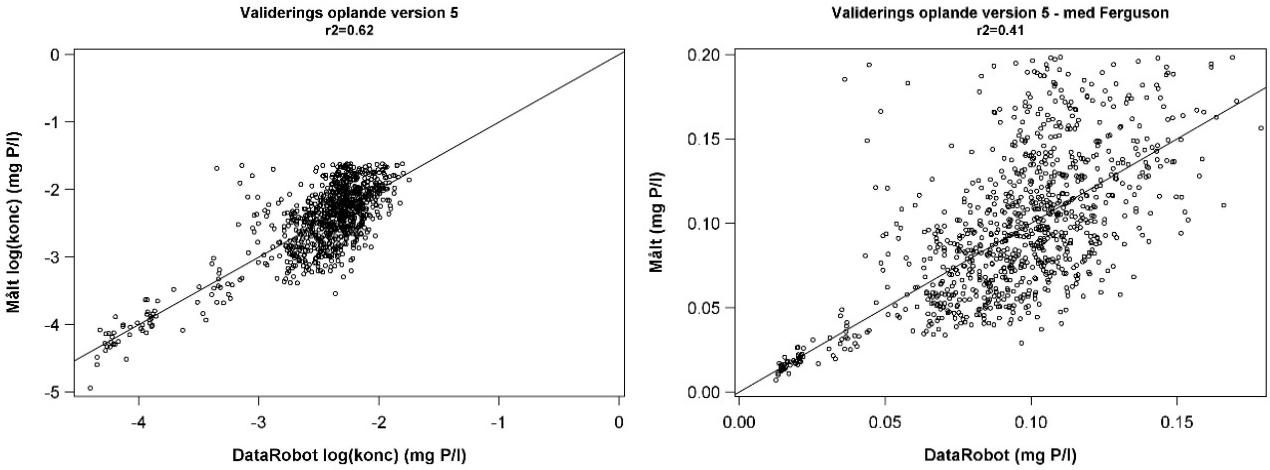


Figur 38. Residualplot af forskel mellem målt og simuleret TP-koncentration med den endelige ML-06 TP-koncentrationsmodel i perioden, hvor modellen er testet (1990-2019), vist for logaritme-og tilbagetransformerede data med Fergusonkorrektion.

Validering af TP-model

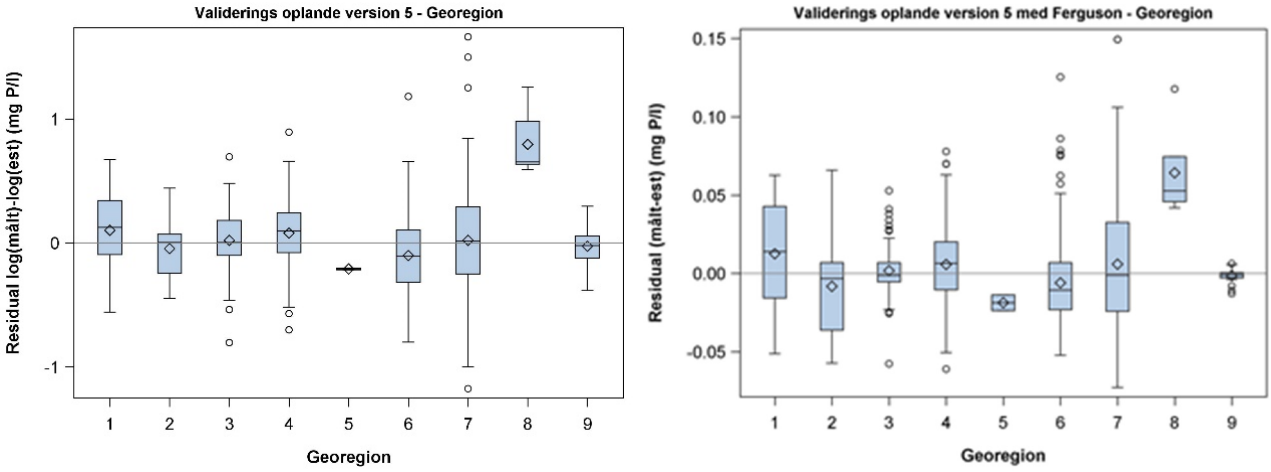
Der er gennemført en validering af den endelige ML-06 TP-koncentrationsmodel på de uafhængige valideringsstationer, som dækker landets regioner. Denne valideringsøvelse er bedre end valideringen i DataRobot i holdout-datasættet, da der indgår mange flere vandløb. Valideringen kan således bibringe os en vurdering af modellens simuleringsgodhed samt et udtryk for usikkerheden i modellens fremtidige anvendelse i umålte oplande.

Valideringen med simuleret vandføringsvægtet årskoncentration af TP mod de målte koncentrationer er vist i figur 39 for både logaritme- og tilbagetransformerede data med Fergusonkorrektion. Der er i begge tilfælde en god forklaringsgrad (R^2) på henholdsvis 0,62 og 0,41.



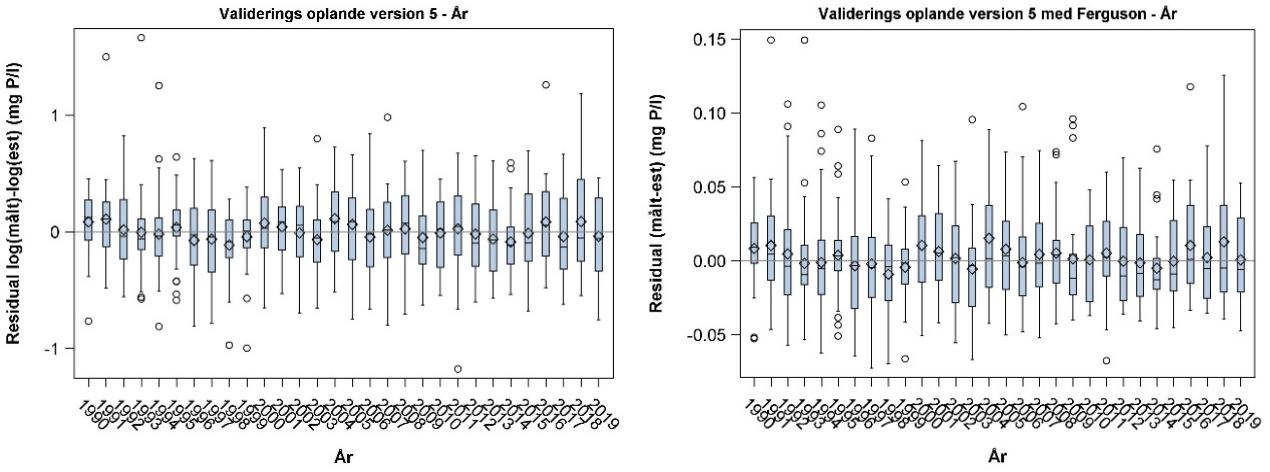
Figur 39. Sammenhæng mellem DataRobot ML-06 TP-model simuleret TP-koncentration (log og tilbagetransformeret) for alle kalibreringsoplande, for alle tre datasæt i modeludviklingen (N=1261).

Den tilsvarende analyse af residualresultater for målte minus simulerede TP-koncentrationer i de ni georegioner i Danmark er også generelt tilfredsstillende, dog med en for lav modelleret koncentration i georegion 8 (Nordsjælland) (figur 40). Dette forhold skyldes formentlig, at der er mange søer og vådområder i georegionen, som måske frigiver akkumuleret fosfor.



Figur 40. Residualplot af forskel mellem målt og simuleret TP-koncentration med den endelige ML-06 TP-koncentrationsmodel inden for ni georegioner i Danmark vist for logaritme-og tilbagetransformerede data med Fergusonkorrektion.

Residualplot med resultater fra valideringen for hvert år i perioden 1990-2019 viser, at de årlige TP-koncentrationer er simuleret godt med gennemsnittet liggende lige omkring 0-linien og en god fordeling på begge sider. Der er således ingen systematiske afgivelser fra 0-residualen (modelbias) (figur 41).



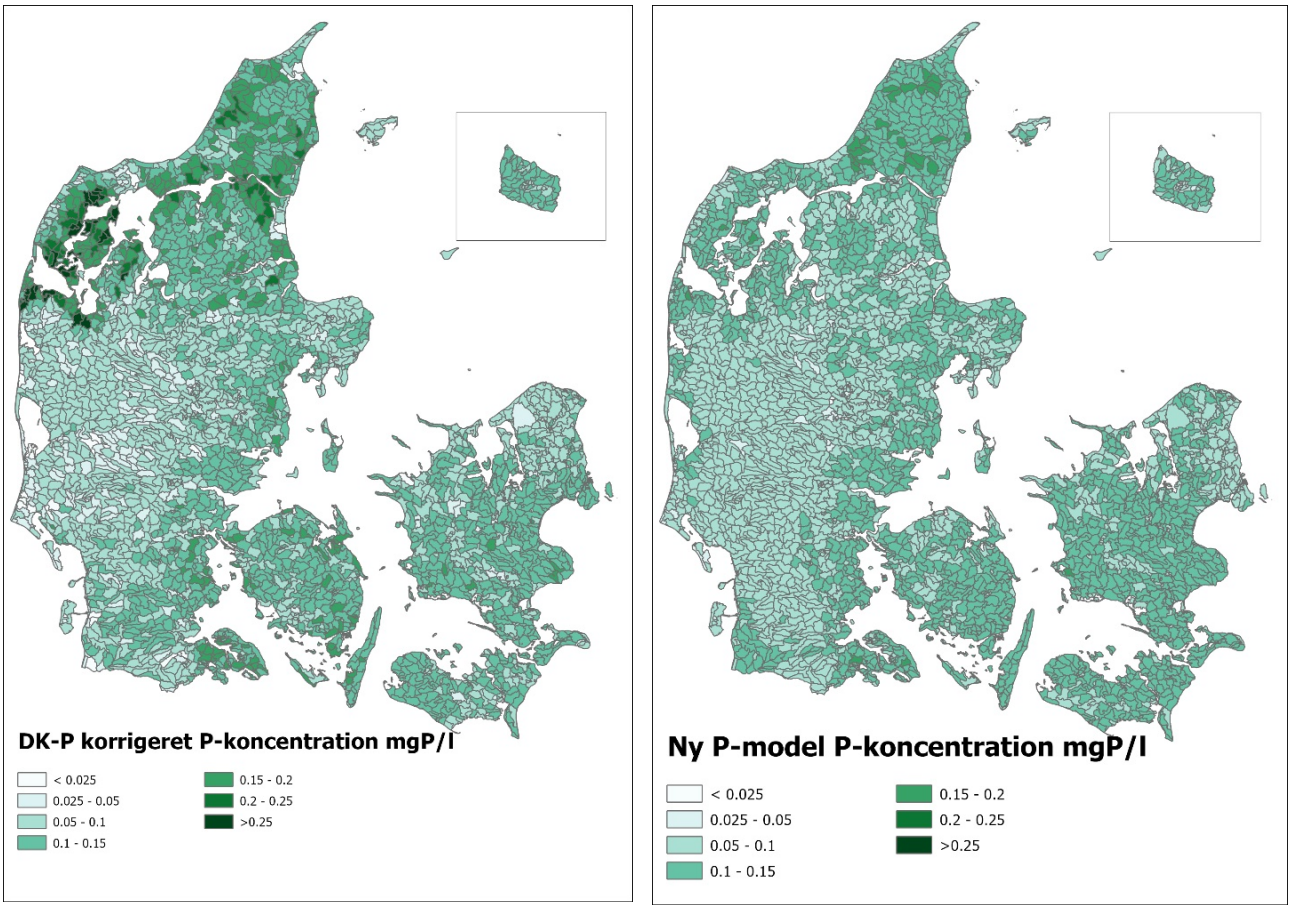
Figur 41. Residualplot af forskel mellem målt og simuleret TP-koncentration med den endelige ML-06 TP-koncentrations model i perioden, hvor modellen er testet (1990-2019), vist for logaritme- og tilbage transformerede data med Fergusonkorrektion.

Konsekvenser af anvendelse af den endelige ML-06 TP-model

En sammenligning mellem den nuværende anvendte bias-korrigerede TP-model (NOVANA DK-P) og den nye TP-model (ML-06) til simulering af den årlige vandføringsvægtede TP-koncentration i vandløb opdelt på ID15-polygoner er vist i figur 42. Der er generelt samme mønster i TP-koncentration i landets vandløb, men med lidt lavere koncentrationer i den nye model end i den gamle bias-korrigerede model, der er anvendt i NOVANA-rapportering af havbelastning siden 2018. Forskellen i resultater for den årlige beregnede tilførsel af TP til de marine kystvande i hele landet er vist for hvert år i perioden 1990-2019 i tabel 11. Tabellen viser både resultatet fra den oprindeligt udviklede TP-model fra 2007, den bias-korrigerede TP-model taget i brug i 2018 og den nye udviklede TP-model. Skiftet i 2018 til den bias-korrigerede model medførte et generelt fald i den beregnede TP-tilførsel til de marine kystvande fra Danmark (tabel 11). Ved et skifte fra den nuværende bias-korrigerede TP-

model til den TP-model, der er udviklet i dette projekt, vil der ske en gennemsnitlig ændring af TP-tilførslen til kystvande i Danmark på ca. -3 %, med variationer mellem de enkelte år i perioden 1990-2019 (-0,9 til -6,9 %) (tabel 11).

Betydningen af skiftet i TP-model for beregningen af TP-tilførsel til 2.-ordens kystafsnit kan ses i figur 43. Der sker generelt et fald i den beregnede TP-tilførsel til kystafsnittene, størst i områder omkring Limfjorden, det sydøstlige Jylland og områder af Sjælland. Det er kun mindre områder, hvor der vil ske en stigning, som f.eks. i Storåens opland, dele af Djursland, helt kystnære dele af Vestsjælland og sydhavsøerne (figur 43).

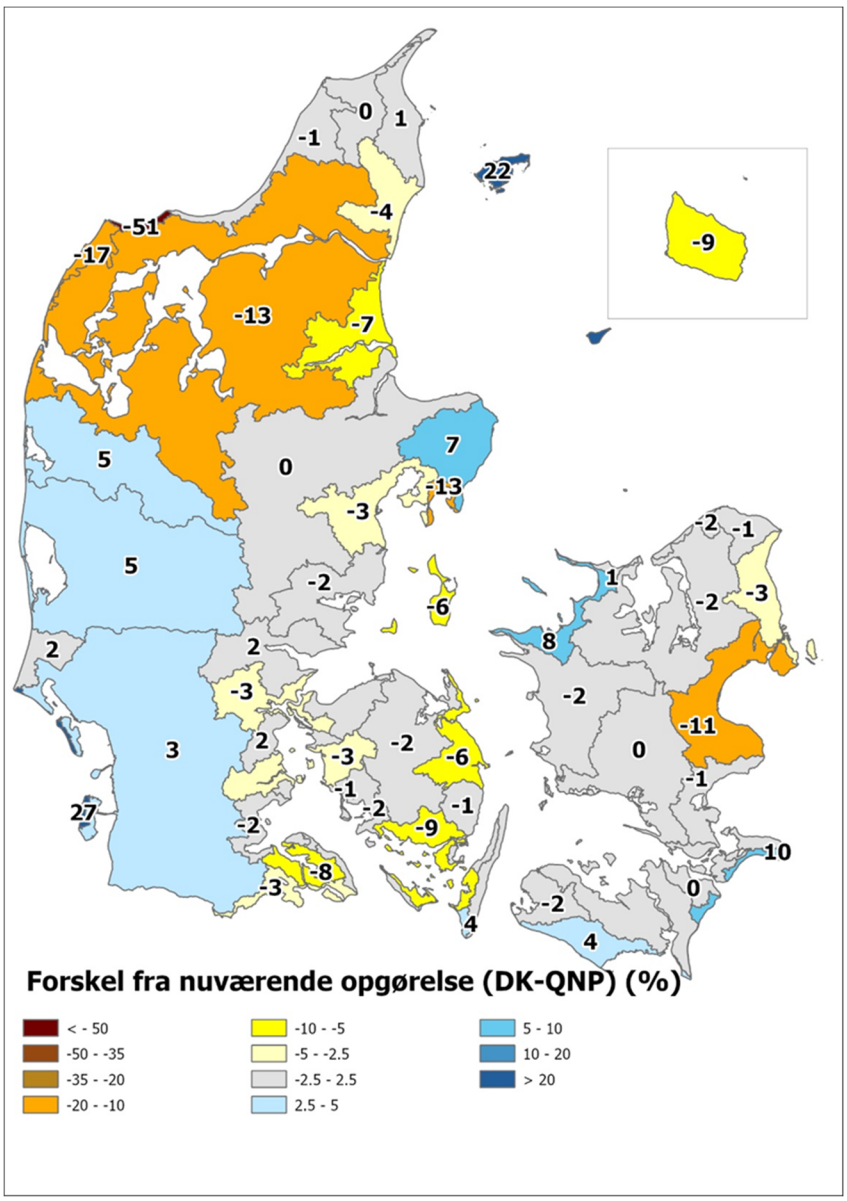


Figur 42. Modelberegnet gennemsnitlig årlig vandføringsvægtet koncentration af diffus total fosfor i vandløb inden for ca. 3200 ID15-oplande i Danmark. I kortet til venstre er anvendt den eksisterende DK-P bias-korrigerede NOVANA-model og i kortet til højre den endelige ML-06 TP-model fra DataRobot.

Tabel 11. Årlige forskelle i beregnet samlet tilførsel af TP til danske kystafsnit via vandløb (inkl. punktkilder til ferskvand og direkte spildevandsudledninger til kystvande) beregnet med tre modeller. Den oprindeligt udviklede NOVANA DK-P model fra 2009 (DK-P model), den nuværende anvendte bias-korrigerede NOVANA DK-P model (DK-P_{bias}) fra 2018 og den endeligt udviklede ML-06 TP-model fra DataRobot

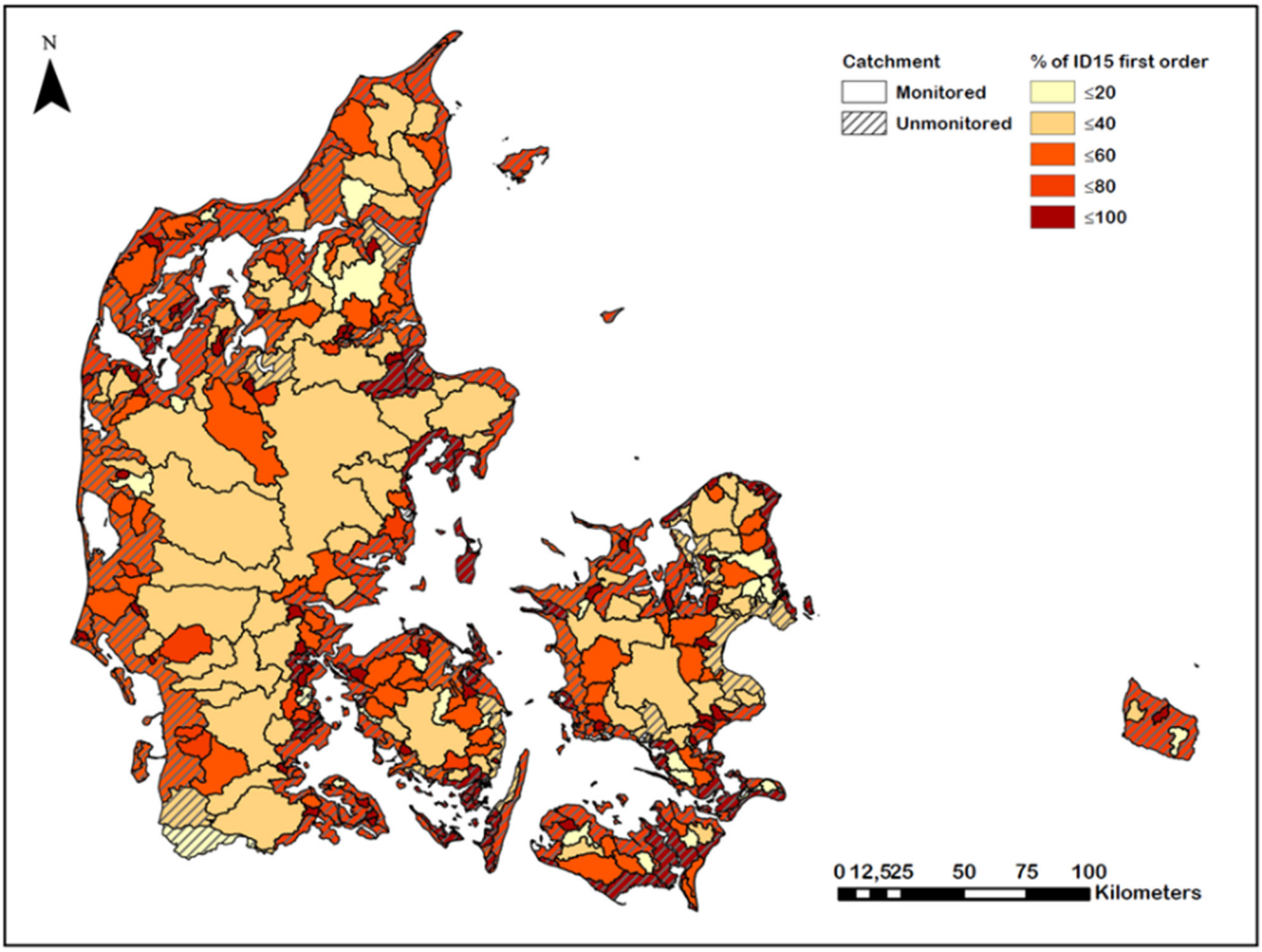
År	DK-P model	Bias-korrigeret DK-P model (DK-P_{bias})	Forskel DK-P og DK-P_{bias}	Ny TP-model Data-Robot	Forskel ny TP-model til DK-P_{bias}
	(Tons P)	(Tons P)	(%)	(Tons P)	(%)
1990	5971	5936	-0,6	5819	-2,0
1991	4874	4867	-0,1	4726	-2,9
1992	4104	3985	-2,9	3949	-0,9
1993	3791	3705	-2,3	3585	-3,2
1994	4608	4624	0,3	4397	-4,9
1995	3324	3288	-1,1	3146	-4,3
1996	1997	1983	-0,7	1925	-2,9
1997	1868	1839	-1,6	1770	-3,8
1998	2729	2640	-3,3	2541	-3,8
1999	3024	3045	0,7	2835	-6,9
2000	2641	2572	-2,6	2456	-4,5
2001	2350	2278	-3,1	2186	-4,0
2002	2885	2711	-6,0	2674	-1,4
2003	1650	1580	-4,2	1542	-2,4
2004	2242	2130	-5,0	2038	-4,3
2005	1879	1780	-5,3	1705	-4,2
2006	2158	1978	-8,3	1919	-3,0
2007	2668	2423	-9,2	2396	-1,1
2008	2279	2132	-6,5	2050	-3,8
2009	1960	1823	-7,0	1781	-2,3
2010	2142	1993	-7,0	1906	-4,3
2011	2455	2266	-7,7	2182	-3,7
2012	2444	2310	-5,5	2209	-4,3
2013	2090	1942	-7,1	1870	-3,7
2014	2308	2130	-7,7	2058	-3,4
2015	2696	2438	-9,6	2394	-1,8
2016	2130	1974	-7,3	1912	-3,1
2017	2283	2076	-9,1	2020	-2,7
2018	1719	1599	-7,0	1536	-3,9
2019	2256	2042	-9,5	1994	-2,4
Gns.	2675	2603	-4,9	2517	-3,3

De umålte kystområder af Danmark har i en GIS-analyse vist sig hovedsageligt at bestå af selvstændige ID15-oplande, der afvander direkte til kystvand (> 60 %) (figur 44). Det samme vil også gælde for de fleste umålte søoplande. Derfor er den udviklede nye ML-06 TP-model netop anvendelig i denne type af umålte områder, da den bygger på et datamateriale fra mindre typeoplande. Andre analyser har vist, at de umålte arealer i deres arealanvendelse og jordtypeforhold er meget lig de målte arealer (bilag 6).



Figur 43. Forskel i TP-tilførsel (i %) mellem nuværende bias-korrigeret DK-P model (1) og den nye ML-06 TP-model udviklet i DataRobot (2), vist som model 2 minus model 1 for 2.-ordens kystafsnit i Danmark i perioden 1990-2019. Negative tal indikerer en mindre og positive tal en større estimeret TP-belastning ved brug af den nye TP-koncentrationsmodel.

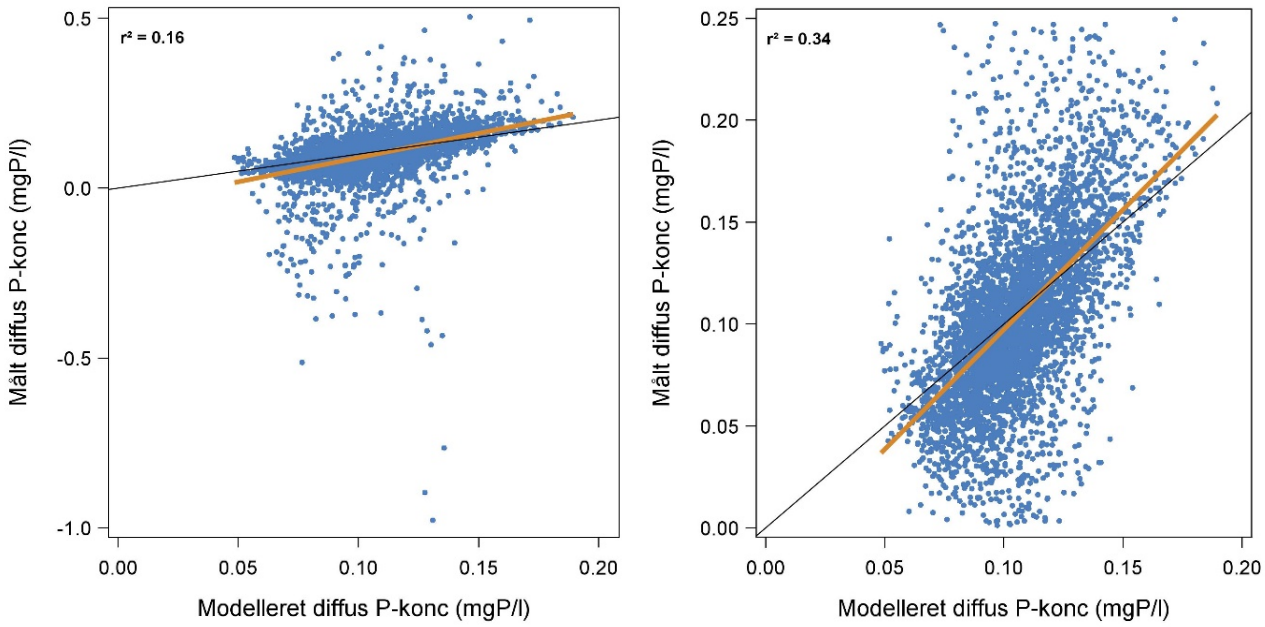
En særskilt sammenligning af den nye TP-models beregnede diffuse vandføringsvægtede TP-koncentrationer ved alle havstationer i NOVANA-programmet er vist i figur 45. Den diffuse TP-koncentration ved målestationerne i NOVANA-havbelastningsprogrammet (Thodsen et al., 2021) er beregnet ved at fratrække udledningen fra punktkilder i oplandet (dog ikke spredt bebyggelse) fra den beregnede årstransport for derefter at beregne den årlige vandføringsfægtede TP-koncentration. Da der ikke pålægges en retention i vandløbssystemet på de udledte mængder af TP fra punktkilder, er der især i starten af overvågningsperioden, hvor udledningen fra punktkilder var stor, en stor usikkerhed ved denne beregning. Derfor er årene 1990-1994 udeladt fra figur 45. Der er dog stadigvæk flere år med beregnede negative TP-koncentrationer, specielt i år med lille TP-transport som 1997.



Figur 44. Andel af ID15-oplande, der afvander direkte til kystafsnit i de umålte og målte områder af Danmark.

Når der laves en beregning af den diffuse TP-koncentration ved havstationerne, fratrækkes bruttoudledningen af P fra punktkilder (dog ikke spredt bebyggelse). Da retentionen i vandløbssystemet af det udledte P fra punktkilder ikke kendes, fratrækkes der alt andet lige en for stor punktkildeandel i forhold til nettobetydningen (inkl. retention). Dette forhold ses også i figur 45, til venstre, hvor der selv efter 1994 i visse år opnås negative koncentrationer af TP i vandløb.

Når koncentrationsintervallet indsnævres til 0-0,2 mg P/l (figur 45 til højre), ses det, at der er en rimelig god overensstemmelse mellem de modelberegnede og de målte TP-koncentrationer ved målestationerne (R^2 = 0,34). Dog ses der en tydelig skævhed i den modellerede diffuse TP-koncentration sammenholdt med de målte diffuse TP-koncentrationer ved havstationerne. Dette forhold skyldes en kombination af påvirkninger fra både usikkerheder på opgørelser af udledningerne fra punktkilder og den manglende inddragelse af fosforretention for både punktkilder og diffuse kilder i større vandløbsoplande. Her kan der være retention af fosfor i både vandløb, vådområder og søer.



Figur 45. Sammenligning af endelig ML-06 TP-model-simuleret gennemsnitlig årlig vandføringsvægtet diffus TP-koncentration mod den målte (TP-transport minus TP-udledning fra punktkilder) ved havstationerne i NOVANA. I figuren til venstre indgår alle data fra perioden 1995-2019. I figuren til højre er data skåret af til intervallet 0-0,2 mg P/l for de modelsimulerede estimater, da modellen er udviklet for data i dette interval på grund af de store usikkerheder, der introduceres ved fratræk af udledningen fra punktkilder.

Usikkerheden på den udviklede nye ML-06 TP-model er beregnet på baggrund af valideringen af modellen på det helt uafhængige valideringsdatasæt af typeoplande (N=1261) samt kalibreringsdatasættet (N=2389). Root Mean Square Error (RMSE) er beregnet til at være lav og modellen derfor god (<0,2) for langt de fleste georegioner (tabel 12). Dermed vil usikkerheden ved simulering af årlig vandføringsvægtet TP-koncentration med den endelige ML-06 TP-model være relativ lille i otte af de ni georegioner (tabel 12). Mean Absolute Error (MAE) er ligeledes beregnet og vist i tabel 12. MAE er generelt mindre for kalibreringsdatasættet, end for det uafhængige validerings datasæt (tabel 12).

Det er kun i georegion 8, at der er beregnet en høj RMSE og MAE og dermed en stor usikkerhed på modelsimuleringen af den årlige vandføringsvægtede total fosfor-koncentration med den nye ML-06 TP-model.

Tabel 12: Beregnet Root Mean Square Error (RMSE) og gennemsnitlig absolut fejl for den nye ML-06 TP-model for de ni georegioner i Danmark, beregnet på henholdsvis kalibreringsdatasættet og det uafhængige valideringsdatasæt.

Georegion	Beregnet Root Mean Square Error for kalibrerings datasættet	Beregnet gennemsnitlig absolut fejl (mg P/l)	Beregnet Root Mean Square Error for det uafhængige valideringsdatasæt	Beregnet gennemsnitlig absolut fejl (mg P/l)
1	0,048	0,023	0,070	0,055
2	0,055	0,036	0,039	0,031
3	0,002	0,015	0,026	0,017
4	0,030	0,019	0,038	0,026
5	0,019	0,015	0,019	0,019
6	0,027	0,016	0,038	0,028
7	0,069	0,034	0,068	0,047
8	0,28	0,072	1,37	0,56
9	0,050	0,024	0,050	0,003

5 Konklusion og perspektiver

De gennemførte analyser viser, at den i trin 6 endeligt udviklede maskinlæringsmodel (ML-06) i DataRobot for estimering af den årlige vandføringsvægtede koncentration af diffus total fosfor (TP) i vandløb er god, da den har en relativt høj forklaringsgrad (R^2 = 0,67-0,71). Selv på et uafhængigt valideringsdatasæt opnås der er høj forklaringsgrad før transformation (R2 = 0,62) og efter transformation (R^2 = 0,41). Den nye TP-model udviklet på landsdækkende stikprøver har derfor en langt højere forklaringsgrad end den gamle TP-model udviklet på intensivdata.

Tilsvarende er modellen god for de fleste georegioner når den bedømmes ud fra Root Mean Square Error (RMSE), da den er mindre end 0,2 for kalibrerings- og valideringsdatasættet i otte af de ni georegioner i Danmark. Desuden er Mean Absolute Error (MAE) mellem 0,003-0,055 i otte af de ni georegioner for det uafhængige valideringsdatasæt. Det er således kun i georegion 9, at der er meget stor usikkerhed ved anvendelse af TP-modellen til at estimere TP koncentrationen. Her kan der således efterfølgende være behov for at foretage en bias-korrektion af modellen.

I modeludviklingen er variable, som kovarierer, så vidt muligt udeladt (dvs. hvis korrelationskoefficienten R > 0,5). Modellen har høje forklaringsgrader, både når den testes på et 'holdout'-datasæt i DataRobot (R^2 = 0,67), og når den udsættes for et uafhængigt valideringsdatasæt, som indeholder vandløb over hele landet (R^2 = 0,62). Desuden har den nye udviklede ML-06 TP-model en meget lille bias, både når der testes inden for georegioner og mellem år i kalibrerings- og valideringsresultaterne.

Konsekvenserne ved anvendelse af den nye ML-06 TP-model i umålte oplande til kystafsnit viser, at der i gennemsnit opnås en ca. 3 % mindre årlig beregnet diffus TP-belastning til kystvandene i hele landet med den nye ML TP-model set i forhold til den nu anvendte bias-korrigerede DK-P model i NOVANA. Der er et enkelt år, hvor forskellen når op på ca. 7 %. Desuden er der visse regioner, hvor den nye TP-model giver større ændringer end for landet som helhed.

Den endelige konklusion fra projektet er, at det er muligt at inddrage den nye TP-model direkte i de årlige opgørelser uden at medtage en bias-korrektion, som pt. udføres på den nuværende NOVANA TP-model. Den nye ML TP-model skal afvikles i DataRobot hvert år, og inputdata til modellen skal for flere temaers vedkommende opdateres for alle ID15-oplande (f.eks. inputdata til modellen om klimavariable og arealanvendelse).

I operationaliseringen af den nyudviklede ML-06 TP-model til NOVANA-beregningerne skal der desuden udvikles en metode til at omregne den årlige modelsimulerede TP-belastning fra umålte oplande til en månedlig TP-belastning. Månedlige eller endog finere tidsskridt af TP-belastninger anvendes nemlig som inputdata i f.eks. marine modeller og sømodeller.

Dette skal ske ved at etablere et referencenet af målestationer inden for udvalgte regioner, hvorfra den gennemsnitlige månedsandel af den årlige målte TP-transport og/eller afstrømning overføres til de(t) umålte oplande inden for den pågældende region. Dette arbejde skal gennemføres i NOVANA-regi

ved indkøring af modellen til den årlige NOVANA-beregning af fosfortilførsel til kystvande.

Den nye model bygger på datasæt fra stikprøvetagning, og derfor vil modellen alt andet lige underestimere TP-koncentrationen og dermed TP-transporten i vandløb, når ML-06 modellen anvendes på mindre oplande, der grænser op til kystvande i de umålte områder af landet. Hvis der i fremtiden skal korrigeres for denne forskel, vil det kræve flere intensivstationer i NOVANA-stationsnettet, da det pt. kun er i fem LOOP-vandløb, at der gennemføres intensivprøvetagning (Blicher-Mathiesen et al., 2021). Der er dog flere åbenlyse fordele ved at skifte til anvendelse af den nye ML-06 model i NOVANA TP-belastningsopgørelserne:

- Den nye TP-model udviklet i DataRobot har en god forklaringsgrad i både kalibreringsdatasæt, holdout-datasæt og et helt uafhængigt datasæt.
- Den nye TP-model simulerer den årlige TP-koncentration uden bias i de fleste områder af landet.
- Den nye TP-model simulerer årlige TP-koncentrationer, som generelt er på niveau med dem, som måles i både typeoplands- og havstationsnettet, hvor der anvendes stikprøvetagning.

Der arbejdes i DCE, AU, i et erhvervs-PhD-projekt og et projekt med Miljøstyrelsen på at teste nye sensorer til onlinemålinger af både kvælstof- og fosforkoncentrationer i vandløb (se van't veen et al., 2019). Derfor vil der i fremtiden opnås mere viden om betydningen af højfrekvente målinger af næringsstofkoncentrationer i vandløb.

I de større oplande, hvor der er målt TP-koncentration med stikprøver i havbelastningsnettet, viser vores analyser, at forskellen mellem intensiv prøvetagning og stikprøvetagning er af mindre betydning end i mindre typeoplande. Dette forhold indikerer, at retentionsprocesser ved transporten af fosfor gennem vandløbssystemer er med til at reducere dynamikken af især partikelbundet fosfor, hvilket forventeligt vil påvirke koncentrationen af opløst uorganisk fosfor pga. optag og frigivelse fra planter i vandløb og søer. Denne interne retention af fosfor i vandløb, ådale, vådområder og søer er derfor af stor betydning, hvis der i fremtiden skal opstilles en oplandsmodel for fosfor. Der er i konsekvensanalyserne af anvendelsen af den nye TP-model konstateret problemer ved fratræk af udledning af TP fra alle punktkilder i et opland fra den beregnede TP-transport ved en vandløbsstation. Dette forhold skyldes formentlig både en usikkerhed ved de eksisterende opgørelser af udledninger fra punktkilder kombineret med, at der ikke pålægges en retention i vandløbssystemet på disse udledninger. Der mangler også en mere præcis viden om udledningen fra punktkilder til ferskvand gennem året (månedligt), hvilket er vigtigt i forhold til at kunne lave månedsfordeling af fosforbelastningsopgørelser til kystvande og søer.

Viden om, hvad der sker med fosfor i vandløbssystemerne, såvel abiotiske processer (sorption, sedimentation, resuspension) som biotiske processer (optag og frigivelse), er fortsat meget mangelfuld. Derfor kan en mere samlet forståelse af fosfordynamikken og –transporten gennem et vandløbssystem først modelleres, når der forefindes data og beskrivelser af disse processer.

Forslag til mulige fremadrettede forbedringer i relation til opgørelser af fosforbelastning af kystvande og søer er følgende:

Forbedret grundlag for at opgøre fosforudledningen fra punktkilder i et opland til i hvert fald månedsniveau og betydning af retention for tilførsel af TP fra punktkilder til vandløb.

Oprettelse af et antal NOVANA-stationer med sensorer til onlinemålinger af fosforkoncentrationer og -transport.

Udvikling af en selvstændig model for opløst orthofosfatkoncentrationer i vandløb (månedsmodel), der kan linkes til bioindikatorer i vandløb (bentiske alger).

Opnåelse af mere viden om abiotiske og biotiske fosfor retentionsprocesser i vandløb.

På sigt udvikling af en oplandsmodel for fosforkoncentration og -transport i vandløb baseret på intensivt målte fosforkoncentrationer og høj opløsning på nedbør og afstrømning.

6 Referencer

Andersen, H.E., Kronvang, B. & Larsen, S.E. 2005, Development, validation and application of Danish empirical phosphorus models. Journal of Hydrology 304: 355-365

Andersen, H.E. & Heckrath, G. (redaktører). 2020. Fosforkortlægning af dyrkningsjord og vandområder i Danmark. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 340 s. – Videnskabelig rapport nr. 397. http://dce2.au.dk/pub/SR397.pdf

Blicher-Mathiesen, G., Houlborg, T., Petersen, R.J., Rolighed, J., Andersen, H.E., Jensen, P.G., Wienke, J., Hansen, B. & Thorling, L. 2021. Landovervågningsoplande 2020. NOVANA. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 260 s. – Videnskabelig rapport nr. 472

Breiman, L. 1997. Arcing the edge. Technical Report 486, Statistics Department, University of California at Berkeley.

Breiman, L. 2001. Random Forest. Machine Learning 45: 5.32.

Chen, T. & He, T. 2015. Higgs boson discovery with boosted trees. Cowan et al., editor, JMLR: Workshop and Conference Proceedings. No. 42.

Datarobot 2021. Datarobot inc. manual.
https://docs.datarobot.com/en/docs/index.html

Ferguson, R.I. 1986. River loads underestimated by rating curves. Water Resources Research 22: 74-76.

Freund, Y. & Schapire, R.E. 1997. A decision-theoretic generalization of online learning and an application to boosting. Journal of Computer and System Sciences 55.1: 119-139.

Friedman, J.H. 2001. Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics 29: 1189-1232.

Géron, A. 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly / Wiley, 2019, 2.ed. ISBN: 9781492032649

Hastie, T., Tibshirani, R. & Friedman, J. 2009. Elements of Statistical Learning. Springer.

Højberg, A.L., Stisen, S., Olsen, M, Troldborg, L., Uglebjerg T.B & Jørgensen, L.F. 2015. DK-model 2014 - Model opdatering og kalibrering. GEUS-rapport 2015/8, København. Bilag til GEUS rapport 2015/8.

Kaggle, Avito, Context and Clicks 2015, Kaggle competitions.

Larsen, S.E., Windolf, J. & Kronvang, B. 2016. Validering af fosformodellen. DCE, Aarhus Universitet notat 2016, 16 s.

Larsen, S.E., Tornbjerg, H., Søndergaard, M., Thodsen, H. & Blicher-Mathiesen, G. 2020. Forskelle i målt koncentration af totalkvælstof og totalfosfor i fersk-vand ved at anvende de to oplukningsmetoder til organisk stof; autoklave- og UV-metode. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 53 s. – Fagligt notat nr. 2020|38 https://dce.au.dk/fileadmin/dce.au.dk/Udgivelser/Notatet_2020/N2020_38.pdf

Ovesen, N.B. Larsen, S.E., Bøgestrand, J. & Kronvang, B. 2009. Fosforbelastningen af havet. Kapitel 7 i: Bøgestrand, J. (red.) 2009: Vandløb 2007. NOVANA. Danmarks Miljøundersøgelser, Aarhus Universitet. 108 s. – Faglig rapport fra DMU nr. 711 http://www.dmu.dk/Pub/FR711.pdf

Svendsen, L.M. & Jung-Madsen, S. (red.) 2020. Homogenitetsbrud og potentielle fejl i nedbørsdata. Eksempler på konsekvenser for myndighedsbetjeningen. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 28 s. - Fagligt notat nr. 2020|51 https://dce.au.dk/fileadmin/dce.au.dk/Udgivelser/Notatet_2020/N2020_51.pdf

Thodsen, H., Tornbjerg, H., Rasmussen, J.J., Bøgestrand, J., Larsen, S.E., Ovesen, N.B., Blicher-Mathiesen, G., Kjeldgaard, A. & Windolf, J. 2019. Vandløb 2018. NOVANA. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 70 s. – Videnskabelig rapport nr. 353.
http://dce2.au.dk/pub/SR353.pdf

Thodsen, H., Tornbjerg, H., Rolighed, J., Baattrup-Pedersen, A., Larsen, S.E., Ovesen, N.B., Blicher-Mathiesen, G. & Kjeldgaard, A. 2021. Vandløb 2020. NOVANA. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 82 s. – Videnskabelig rapport nr. 473. http://dce2.au.dk/pub/SR473.pdf

Tornbjerg, H., Windolf, J., Larsen, S.E. & Kronvang, B. 2019. Notat om opgørelse af fosfortransport i NOVANA – målinger og modeller. – Fagligt notat fra DCE – Nationalt Center for Miljø og Energi, 32 s. http://dce.au.dk/fileadmin/dce.au.dk/Udgivelser/Notater_2019/Opgoerelser_af_fosfortransport_i_NOVANA.pdf

Van't Veen, S., Holm, H. & Kronvang, B. 2020. Undersøgelse af anvendelse af sensorer i overvågningen- og test af en nitratsensor i vandløb. Aarhus Universitet, DCE – Nationalt Center for Miljø og Energi, 79 s. – Fagligt notat nr. 2020|48 https://dce.au.dk/fileadmin/dce.au.dk/Udgivelser/Notatet_2020/N2020_48.pdf

Vejen, F., Schmith, T., Scharling, M. & Vedel, H. 2021. Betydning af instrumentskift ved overgang fra manuel til automatisk måling af nedbør. DMI-rapport nr. 21-38, 16 s.

Bilag 1: Oversigt over temaer og referencer for inputdata anvendt i de forskellige modeller

Dokumentation, inddata til P-model

Id15-opland version 2.1

IMK-markkort 2010-2019
2010-2016 downloaded fra https://kortdata.fvm.dk/download/ 20190401
2017-2019 modtaget via Statens filkasse.
De mange afgrøder er samlet i to overordnede klasser, intensivt/ekstensivt dyrket ifølge tabellen afgrode_simp.dbf kolonne *Drift*.

Markblokkort 2006-2019
2006-2017 hentet fra AUIT_Geodatamodtaget via Statens filkasse.
Før klip tilføjes ny kolonne *Area_blok* med markblokkens totalareal.

Basemap, arealanvendelseskortlægninger fra 2011, 2016, 2018
produceret af Gregor Levin. De i alt 36 arealanvendelsesklasser oversættes til Corine-klasser vha. nøgletabellerne:

- Basemap_2011_2016_legend.dbf
- Basemap_2018_legend.dbf

Corine-klasserne grupperes ifølge nøgletabellen AIS_legend.dbf.

Corine Land Cover, arealanvendelsekortlægninger fra 1990, 2000, 2006, 2012, 2018
downloadet fra https://land.copernicus.eu/pan-european/corine-land-cover.
De i alt 32 arealanvendelsesklasser grupperes i overklasser ifølge nøgletabellen AIS_legend.dbf.

JB-jordtypekort 2014 produceret af Foulum.
Kortene ligger som rasterlag i fire horisonter: a: 0-30 cm, b: 30-60 cm, c: 60-100 cm, d: 100-200 cm.
Før klip konverteres lagene til vektor.

Kulstof_3-6_12 produceret af Foulum. Laget er landsdækkende med kulstofindholdet inddelt i fire klasser 3: <3 %, 6: 3-6 %, 12: 6-12 %, 60: >12 %.
Før klip konverteres laget til vektor, og klasserne grupperes til under/over 3 % kulstofindhold.

Drained_v3 produceret af Foulum.
Kortet ligger som et rasterlag med cellestørrelse 30,4 x 30,4 m celler beskrivende sandsynligheden for dræning udtrykt som værdier fra 0 til 1 floating point. Før klip er laget omregnet til heltal procent 0-100 % og konverteret til vektor.
Efter klip er der summeret op til en samlet dræningsprocent for det dyrkede areal i hvert id15-opland således:
[(Shape_area * 100% / Area_dyrket) = Arealprocent]
[(Arealprocent * gridcode %) = Total_pct] og Total_pct summeres over id15-opland

Alternativt er der summeret op til < 50 % = drænet, > 50 % = ej drænet og ej dyrket =0.

Pkilder_ID15 produceret af BIOS / Foulum 2020.
Laget indeholder følgende parametre beregnet til id15-opland version 1.5. Laget indeholder kolonnerne:

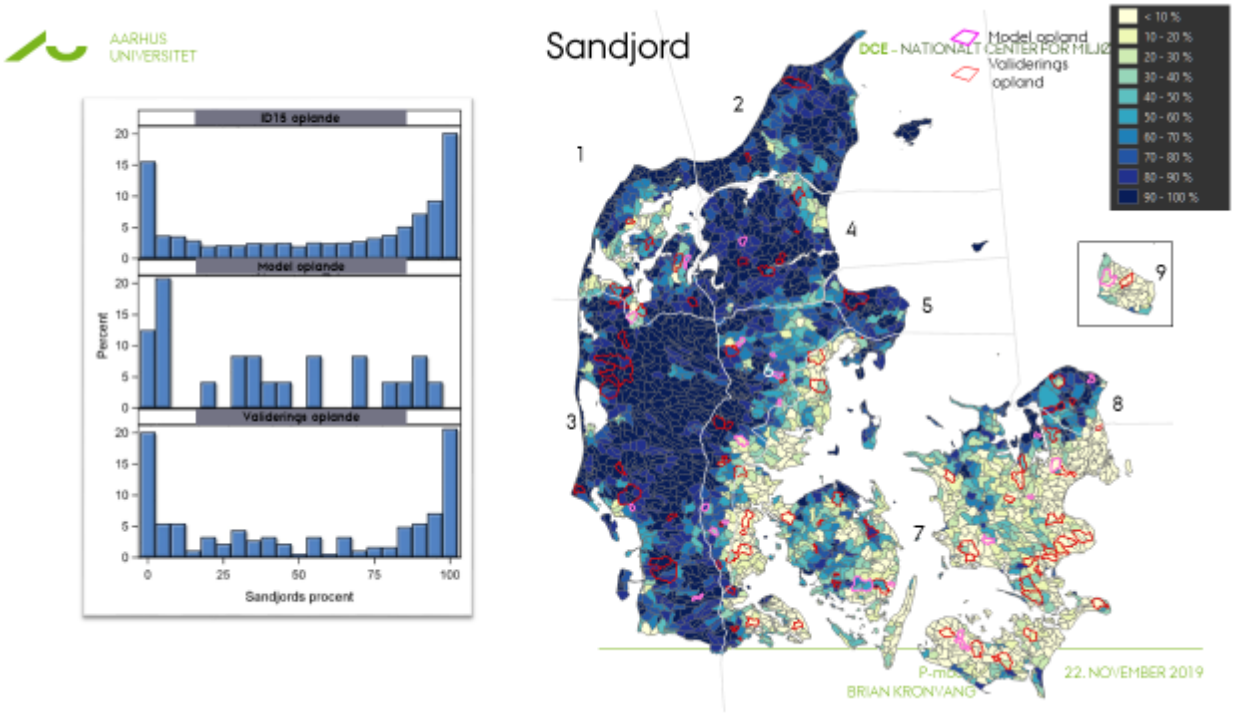
- BE_kg / ha Brinkerosion
- Org_kg / ha Dyrket organisk lavbundsjord
- MP_kg / ha Makropore-transport
- Ma_kg / ha Matrice-udvaskning
- VE_kg / ha Vanderosion
- Diffu_kg / ha Diffus transport
- Lndbr_kg / ha Landbrugsbidrag

Efter klip omregnes til id15opland version2.1 således:
(Ver1.5_klip Shape_area * 100% / ver2.1 total_area) = Arealprocent
(Arealprocent * kg / ha / 100%) = kg / ha.

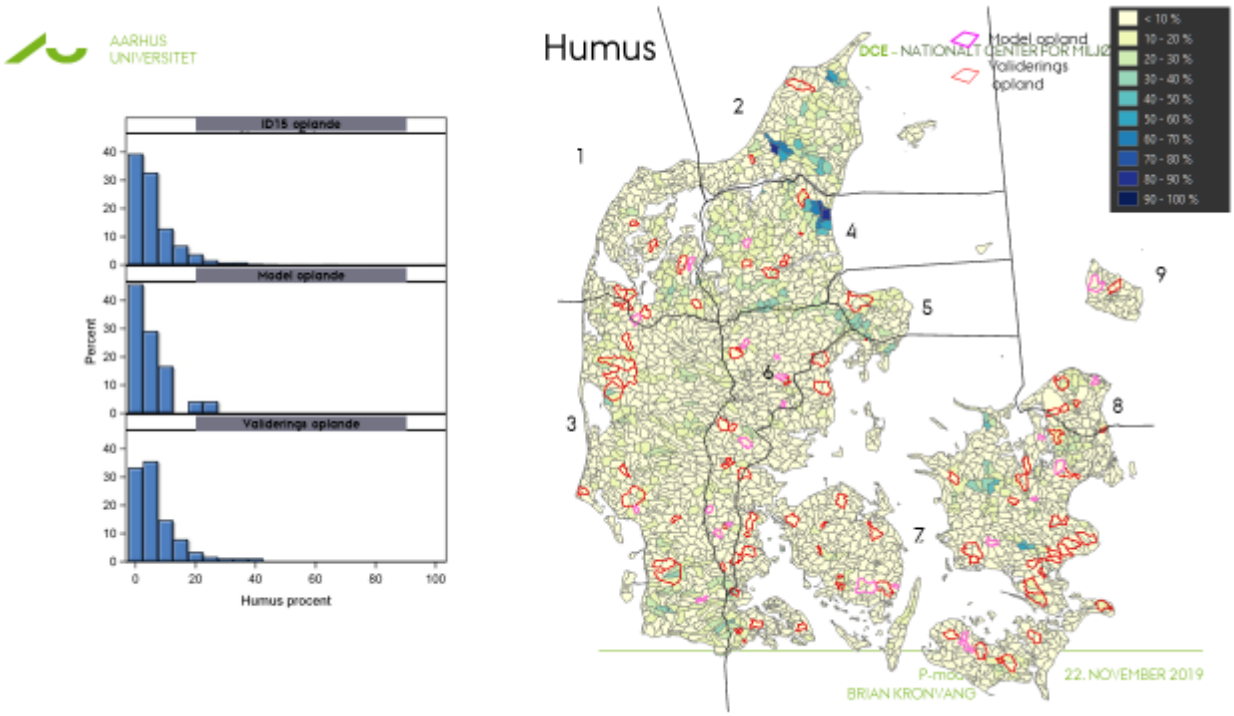
Georegioner produceret af Foulum, hentet fra O:\AUIT_Geodata\Denmark\Natural_ressources\Soil_geology\Georegions.

Nedbør: Hentet fra DMI grid-nedbør (10 x 10 km grids) og nedskaleret til jordoverfladen.

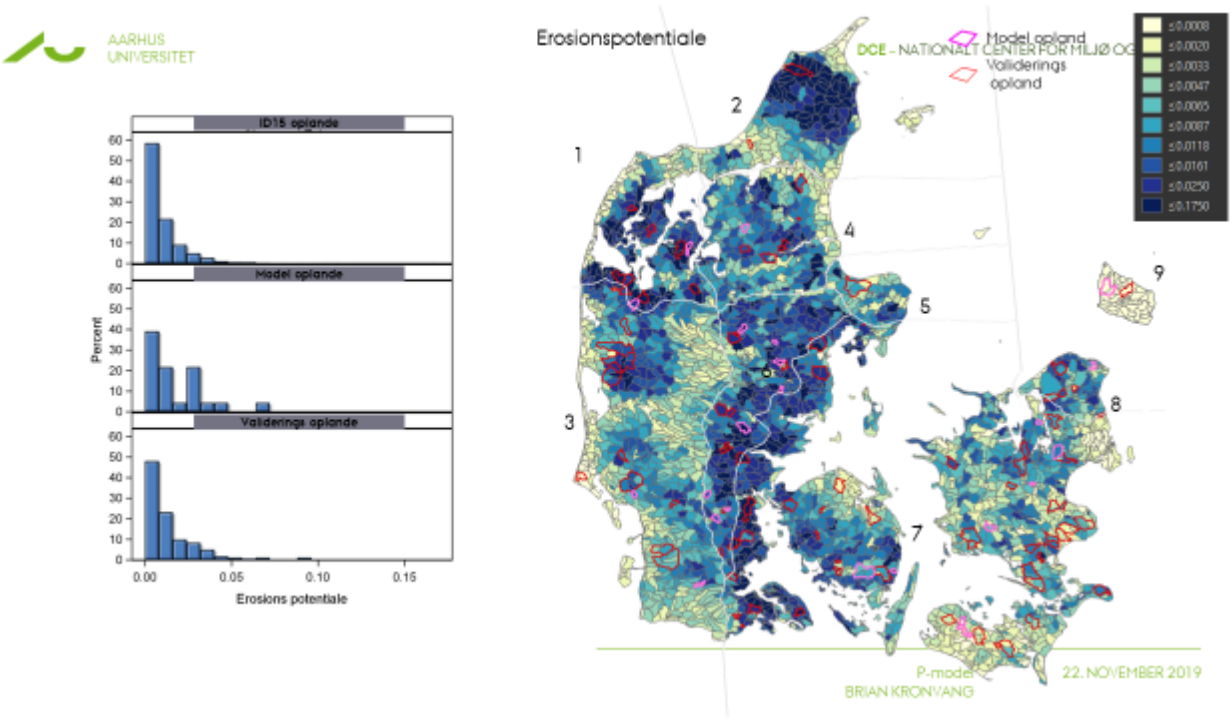
Bilag 2: Oversigt over mulige modelparametre fra oplandene



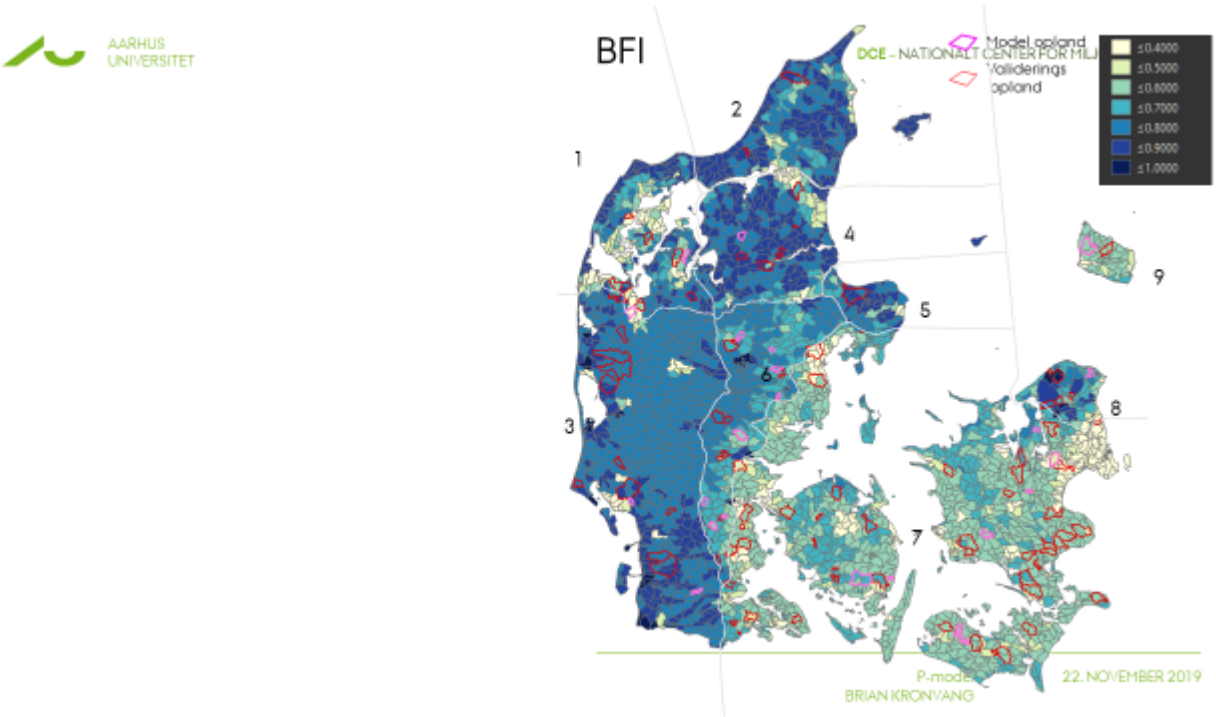
Variation i sandjordsprocenten i ID15-oplande i Danmark, kalibreringsoplande og valideringsoplande.



Variation (%) for humusjorder i ID15-oplande i Danmark, kalibreringsoplande og valideringsoplande.

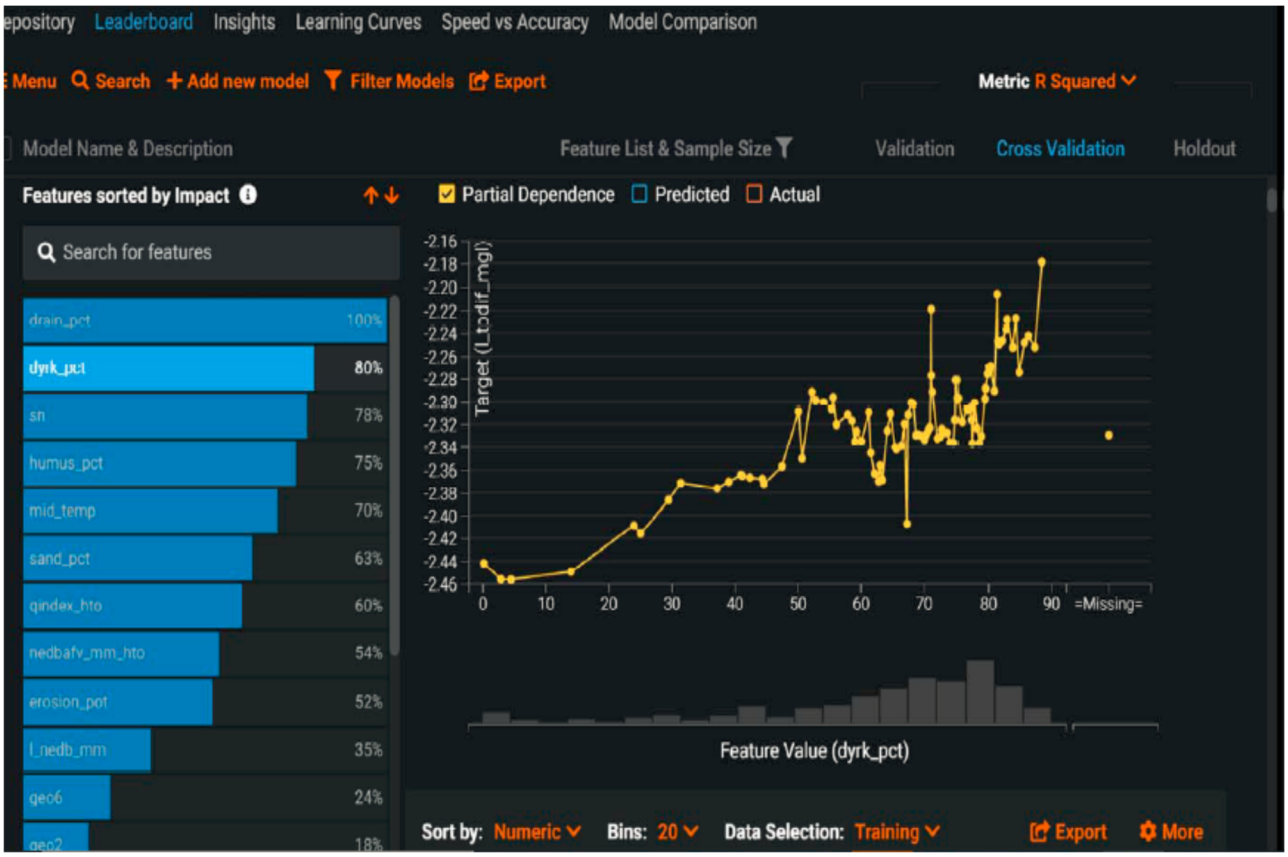


Variation (%) i potentiale for erosion og overfladisk tilstrømning af sediment til vandløb (tons/år) i ID15-oplande i Danmark, kalibreringsoplande og valideringsoplande.

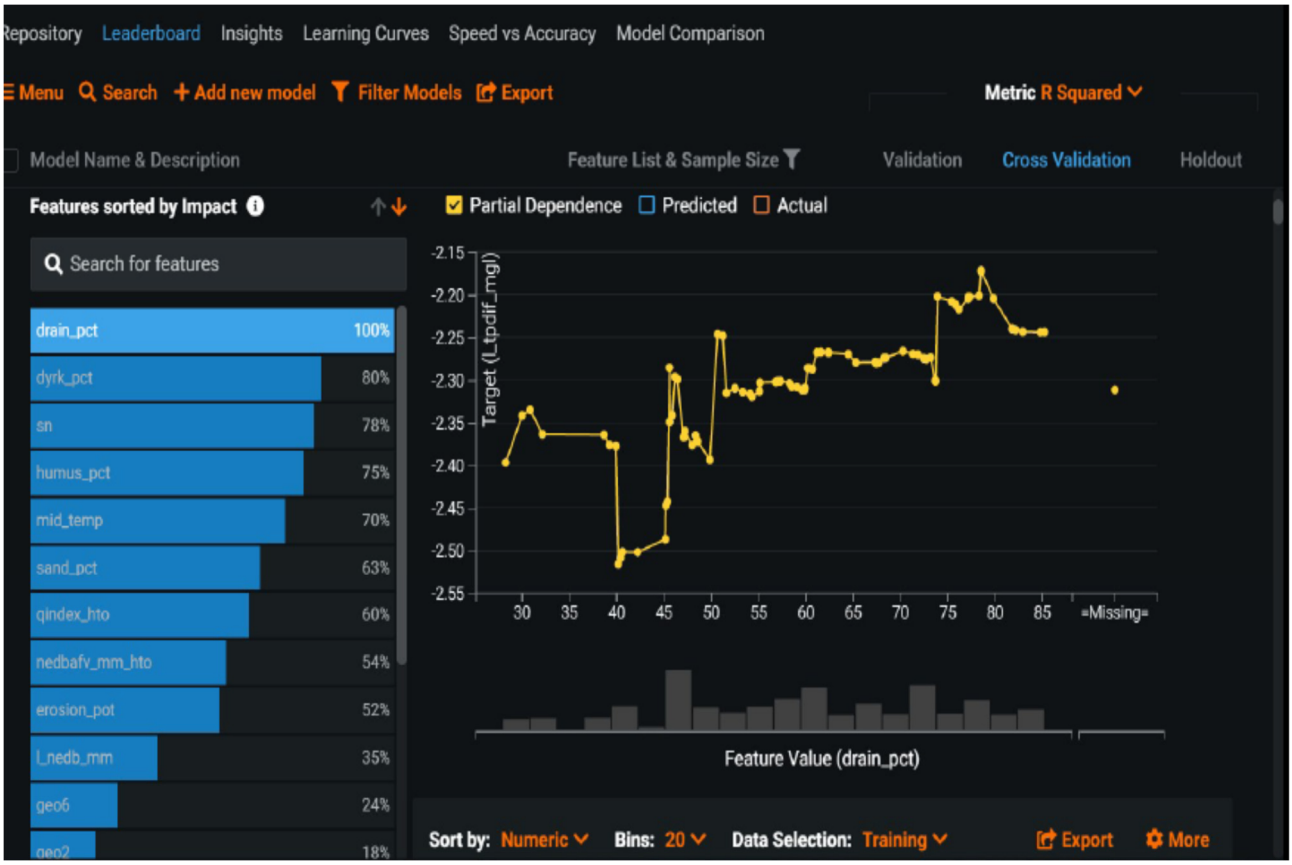


Variation i NOVANA-baseflow-indeks (BFI) i ID15-oplande i Danmark beregnet med ligning 2.

Bilag 3: Prototype ML TP-model – betydning af de mest forklarende variable vist i de tværgående blå søjler fra screen dump fra DataRobot



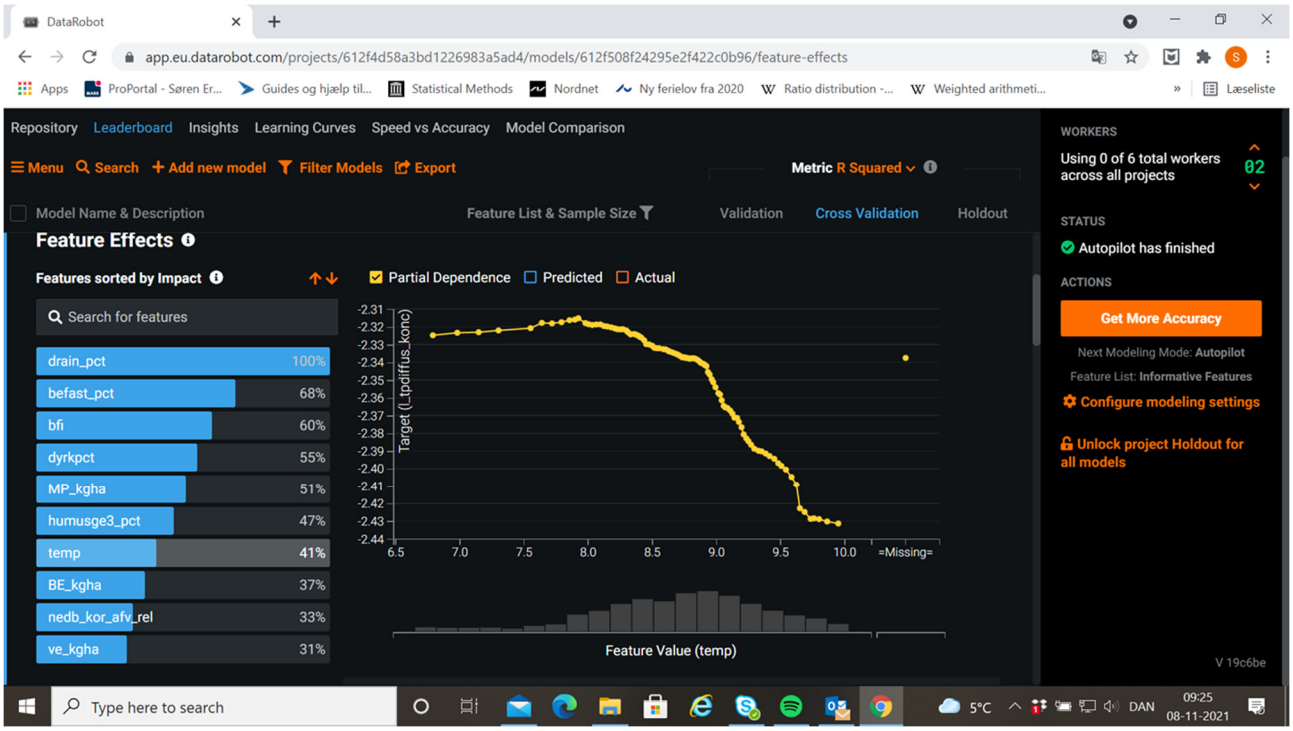
Oversigt over, hvordan den forklarende variabel 'Dyrkningsgrad i oplandet' på x-aksen indgår i prototypemodellen (y-aksen er ln(total P-koncentration) (screen dump fra DataRobot).

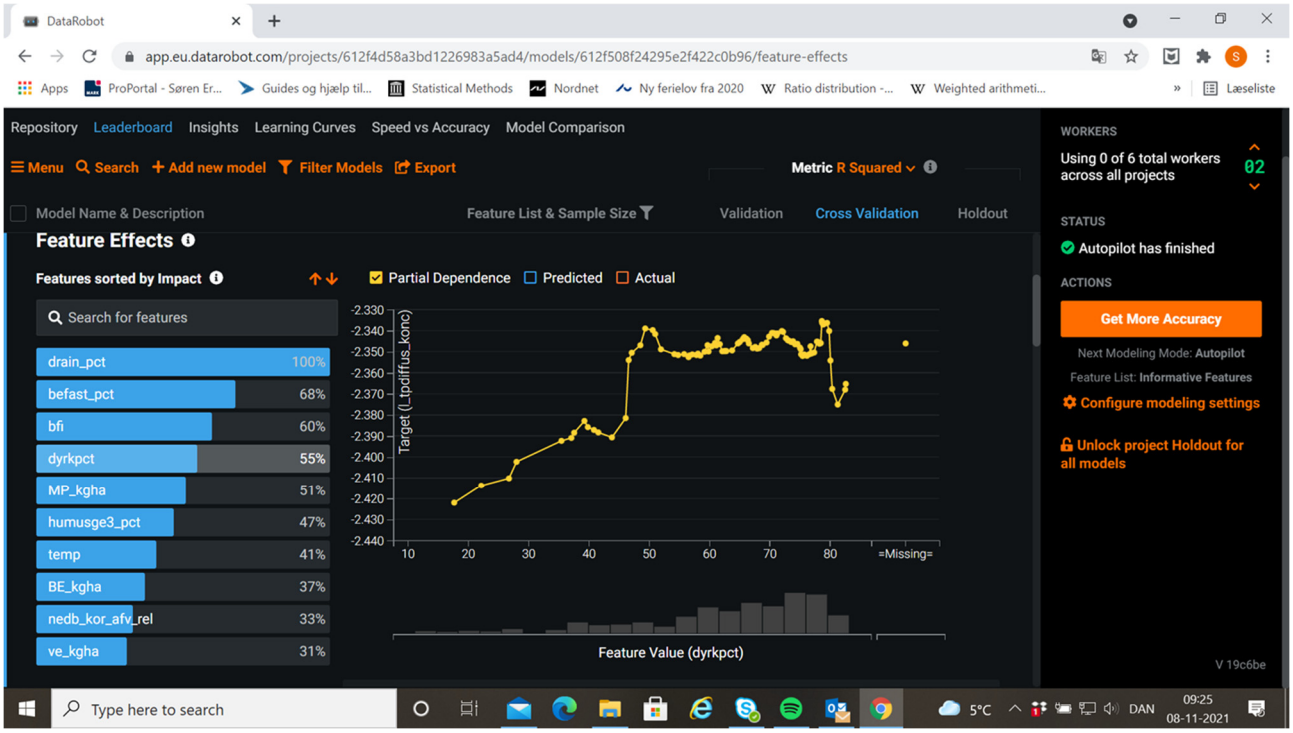


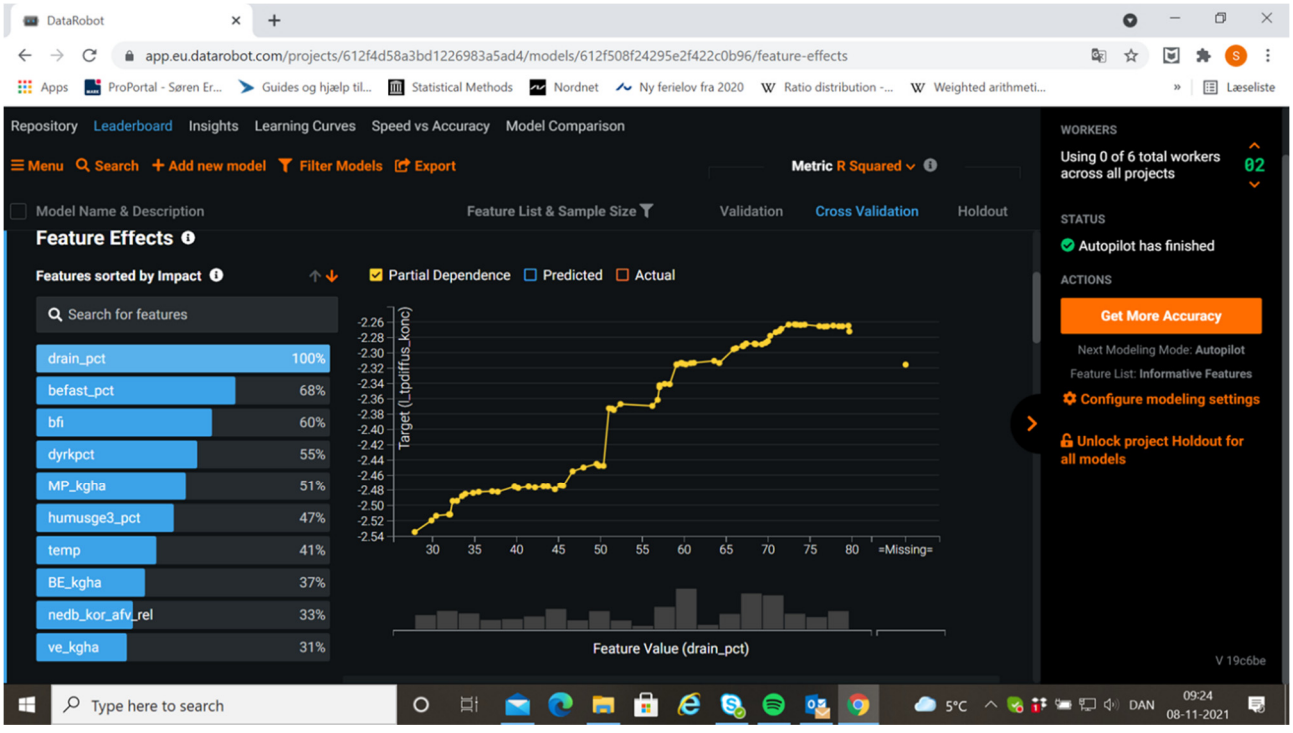
Eksempel på, hvordan den forklarende variabel 'Dræningsprocent i oplandet' på x-aksen (x-aksen viser også antal observationer i intervaller af dræningsgrad; y-aksen er ln(TP-koncentration), indgår i prototypemodellen (screen dump fra DataRobot).

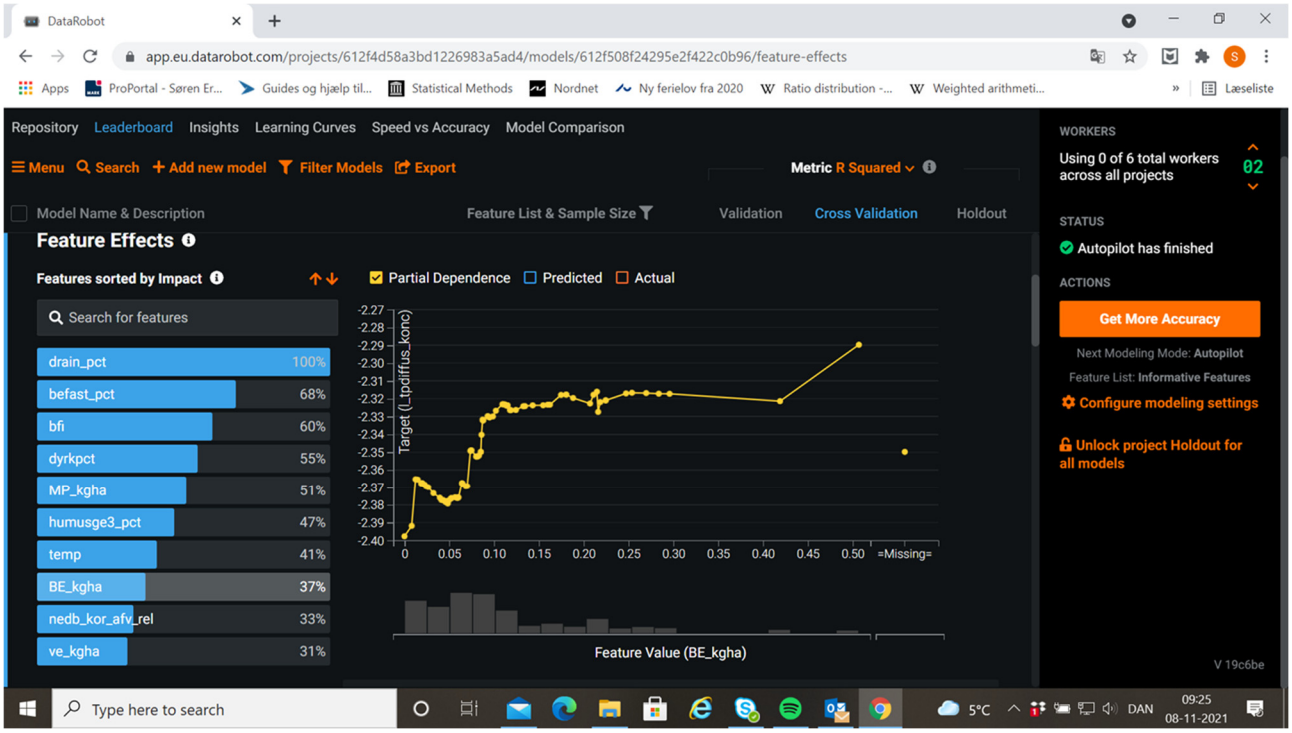
Bilag 4: Oversigt over respons for variable som indgår i ML-01 TP-modellen

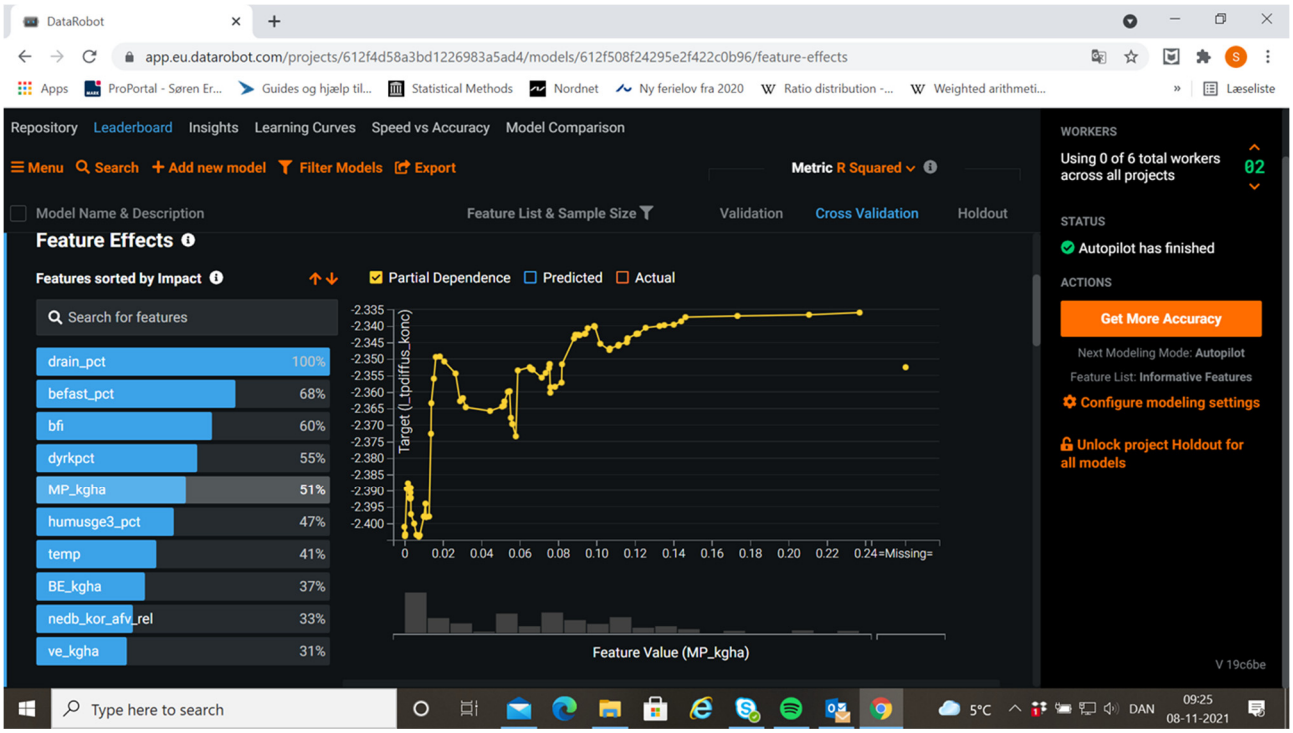
I figurer, som er screen dump fra DataRobot, vises betydningen af de enkelte variable i ML-01 modelversionen (X-aksen) for ln (TP-koncentrationen), som er vist på y-aksen. Temp=temperatur (C), dyrkpct=dyrkningsgrad (%), drain_pct= dræningsgrad (%), BE_kg_ha=brinkerosion (kg / ha), MP kgha= makroporetab,VE_kgha=vanderosion (kg / ha), befast_pct=befæstningsgrad (%), humusg3_pct=humusindhold (%), BFI=baseflow-indeks, nedb_kor_afv_rel =relativ nedbørsafvigelse (korrigeret nedbør).

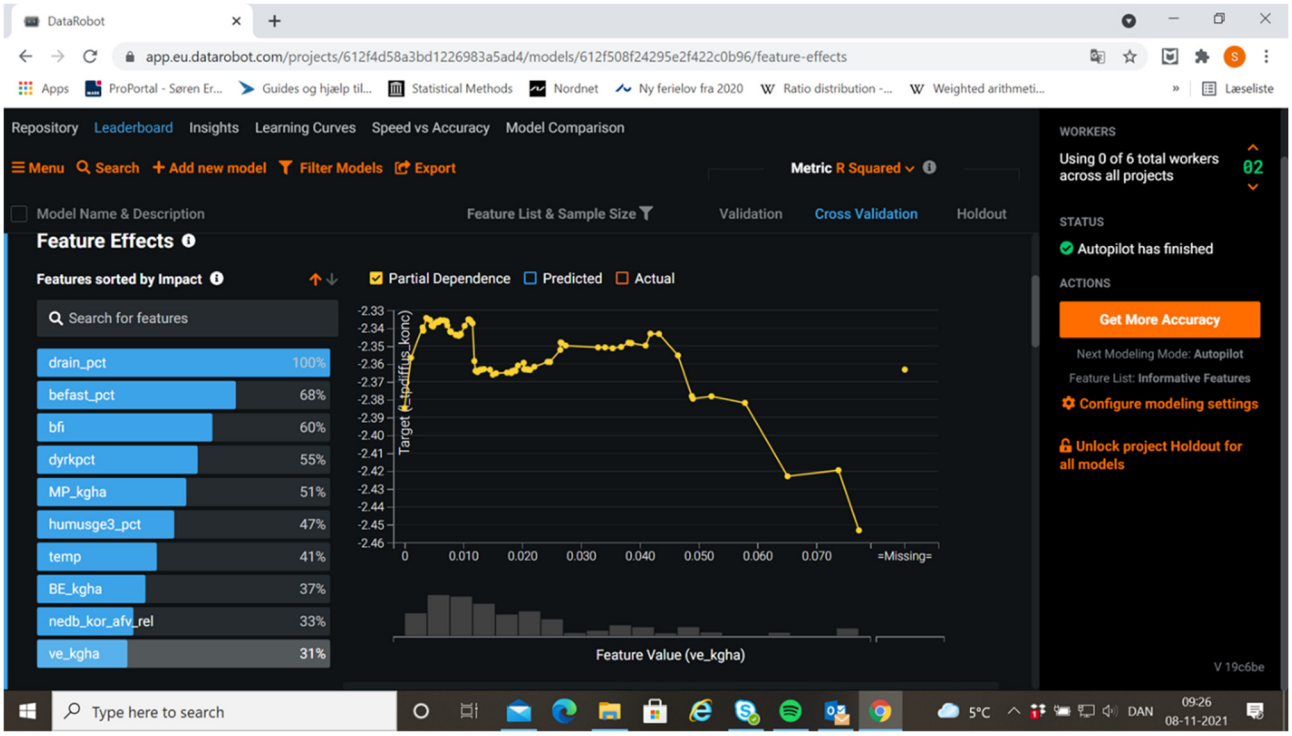


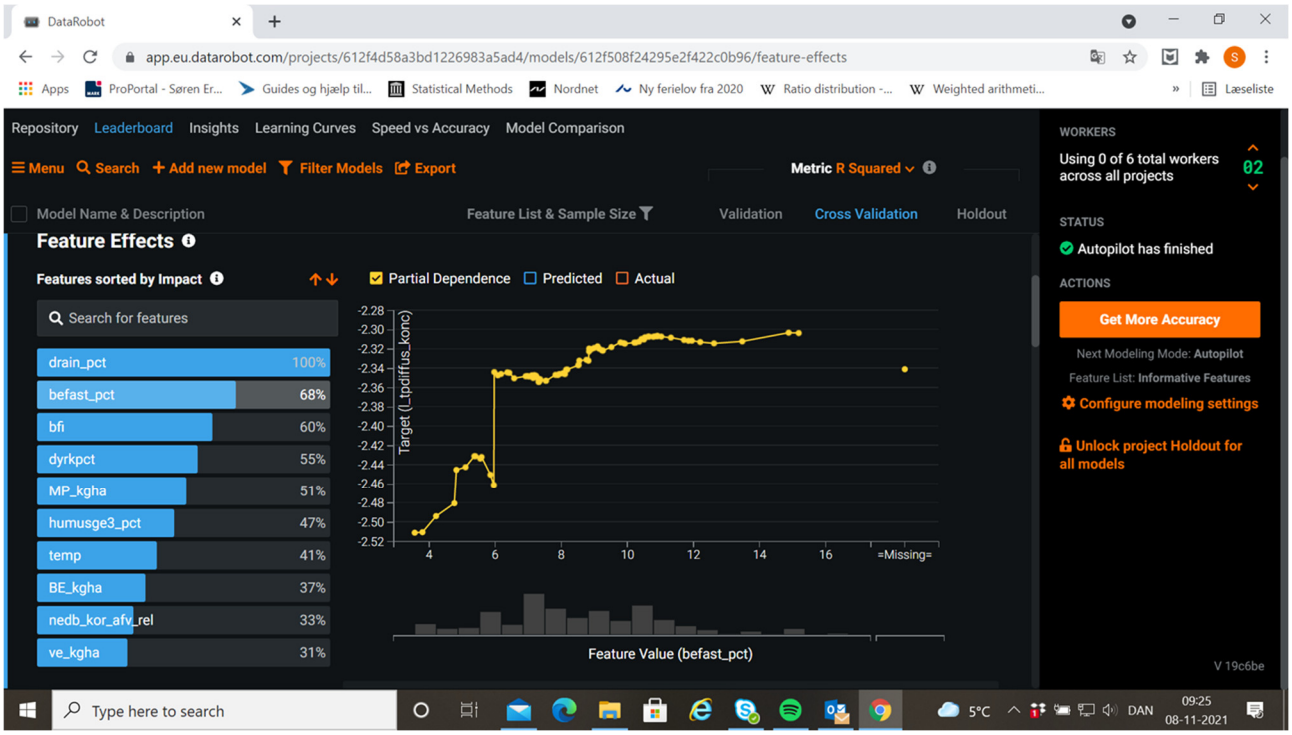

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository Leaderboard Insights Learning Curves Speed vs Accuracy Model Comparison
Menu Search Add new model Filter Models Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (dyrkpct)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:25
08-11-2021

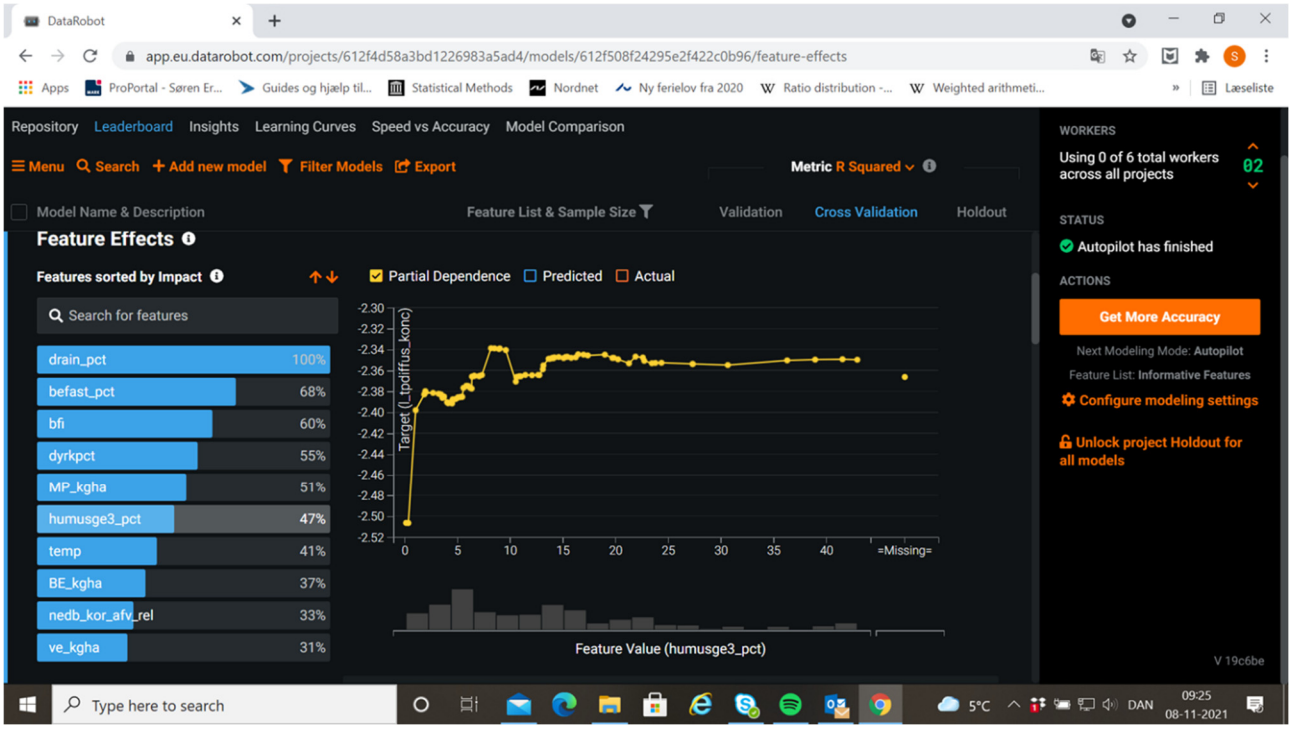

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository Leaderboard Insights Learning Curves Speed vs Accuracy Model Comparison
Menu Search Add new model Filter Models Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (drain_pct)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:24
08-11-2021

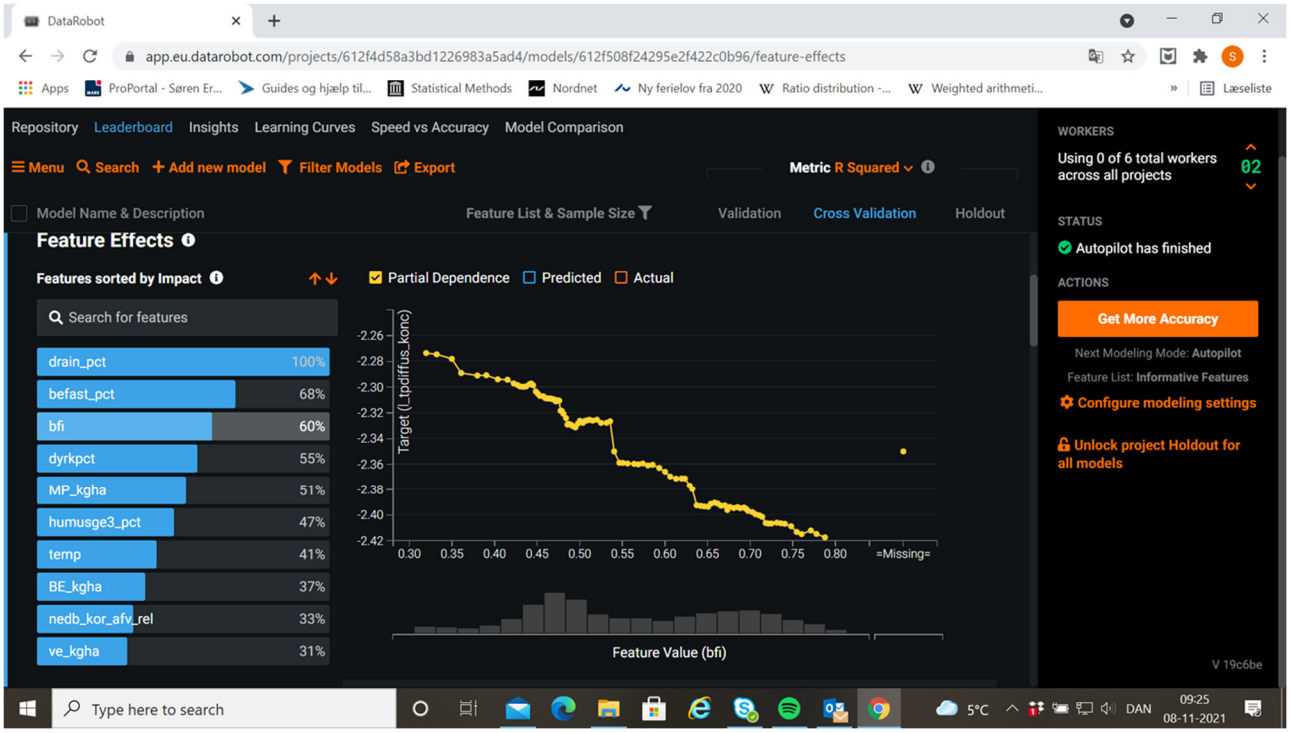

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (BE_kgha)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:25
08-11-2021

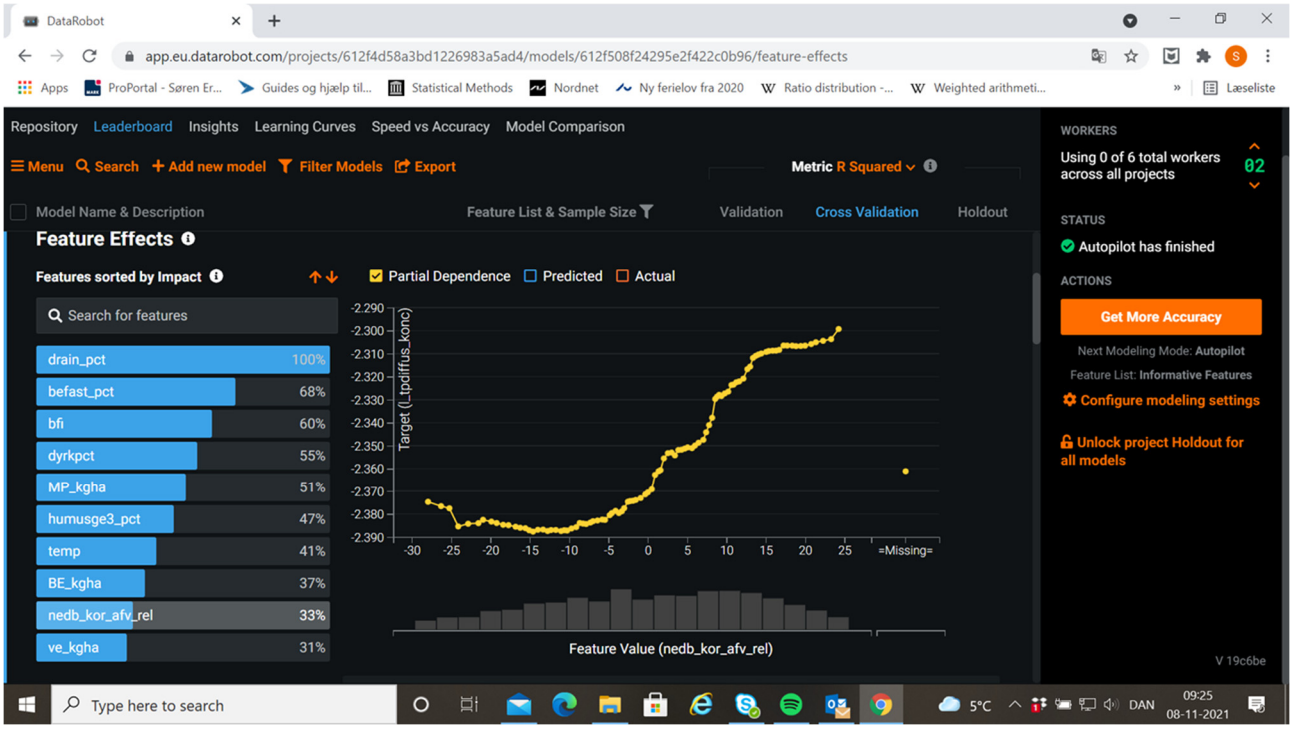

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (MP_kgha)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:25
08-11-2021


DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (ve_kgha)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:26
08-11-2021


DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (befast_pct)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:25
08-11-2021

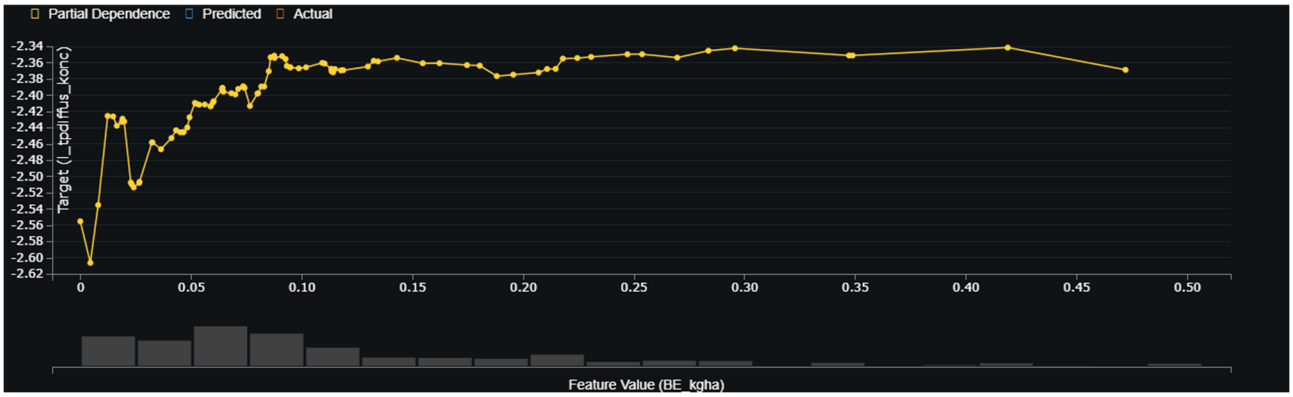

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (humusge3_pct)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
5°C
DAN
09:25
08-11-2021

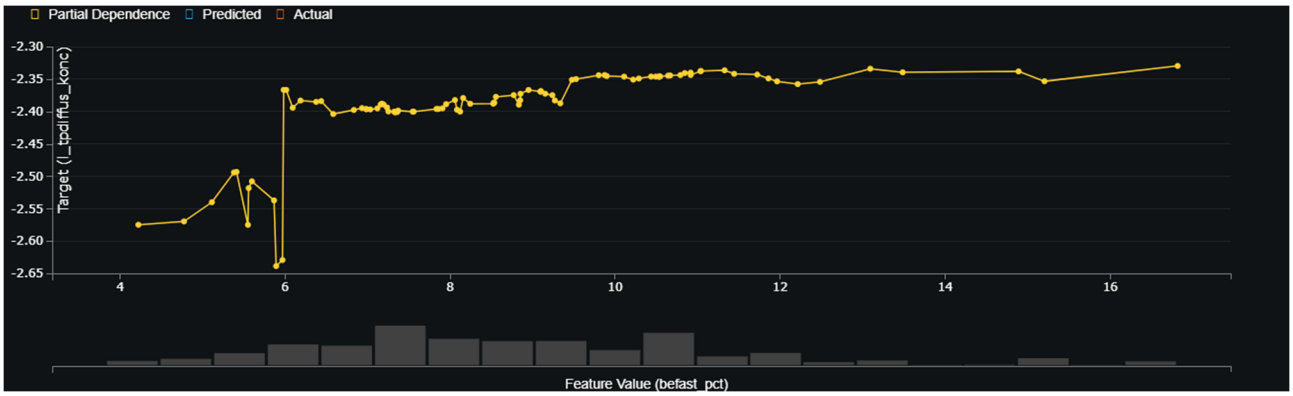

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (l_tpdiffus_konc)
Feature Value (bfi)
WORKERS
Using 0 of 6 total workers across all projects
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
5°C
DAN
09:25
08-11-2021

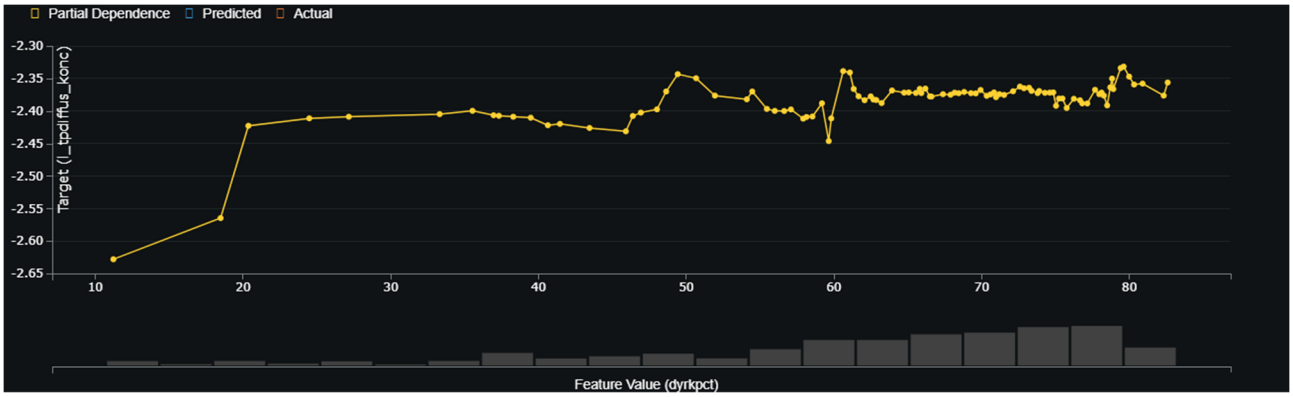

DataRobot
app.eu.datarobot.com/projects/612f4d58a3bd1226983a5ad4/models/612f508f24295e2f422c0b96/feature-effects
Repository
Leaderboard
Insights
Learning Curves
Speed vs Accuracy
Model Comparison
Menu
Search
Add new model
Filter Models
Export
Metric R Squared
Model Name & Description
Feature List & Sample Size
Validation
Cross Validation
Holdout
Feature Effects
Features sorted by Impact
Partial Dependence
Predicted
Actual
Search for features
drain_pct 100%
befast_pct 68%
bfi 60%
dyrkpct 55%
MP_kgha 51%
humusge3_pct 47%
temp 41%
BE_kgha 37%
nedb_kor_afv_rel 33%
ve_kgha 31%
Target (L_tpdlfus_konc)
Feature Value (nedb_kor_afv_rel)
WORKERS
Using 0 of 6 total workers across all projects
02
STATUS
Autopilot has finished
ACTIONS
Get More Accuracy
Next Modeling Mode: Autopilot
Feature List: Informative Features
Configure modeling settings
Unlock project Holdout for all models
V 19c6be
Type here to search
09:25
08-11-2021

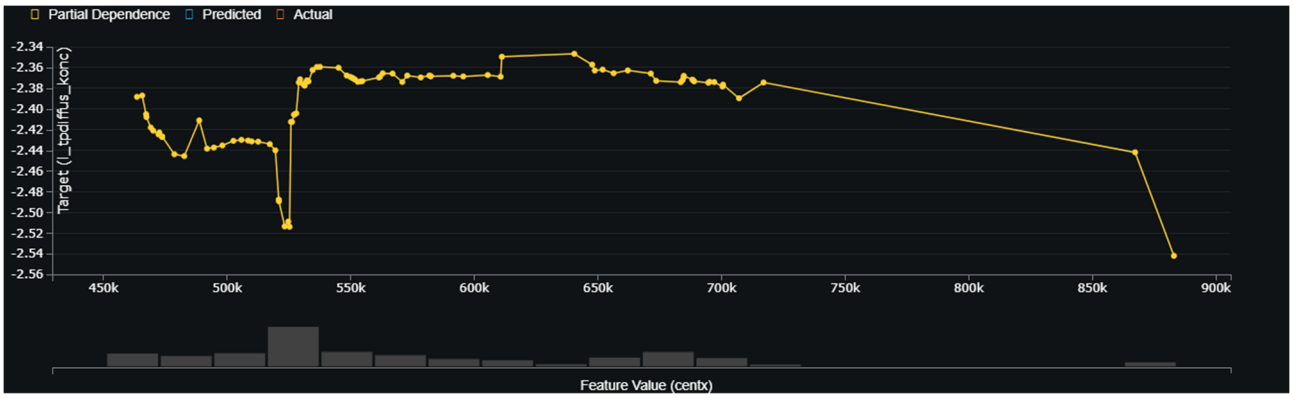
Bilag 5: Oversigt over plots fra den endelige ML-06 TP-model i DataRobot

Resultater for betydningen af de enkelte variable (jf. tabel 10) i den endelige ML-06 TP-koncentrationsmodel udviklet i DataRobot. Resultater er screen dumps fra DataRobot og viser på X-aksen variablen (med antal observationer i datasættet som grå søjler) og på y-aksen ln (TP-koncentrationen).

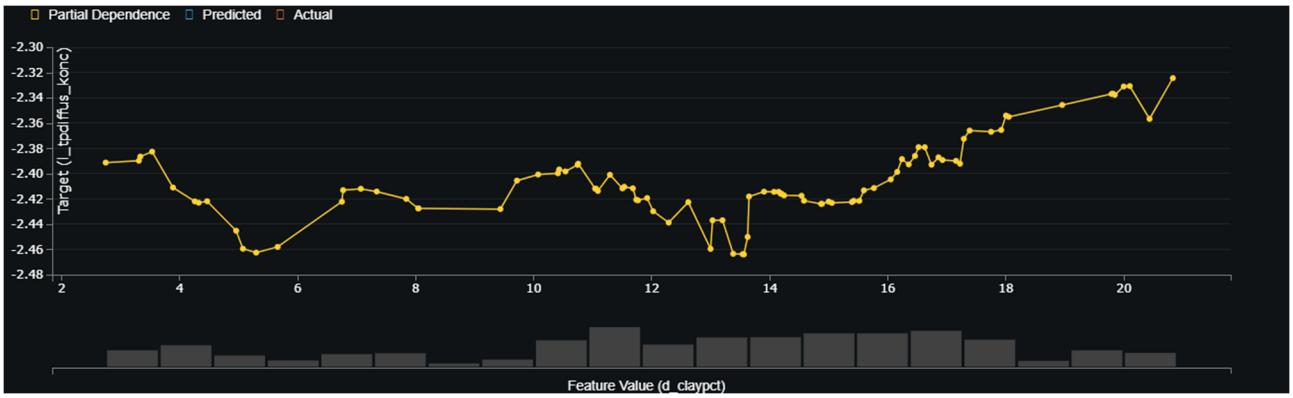


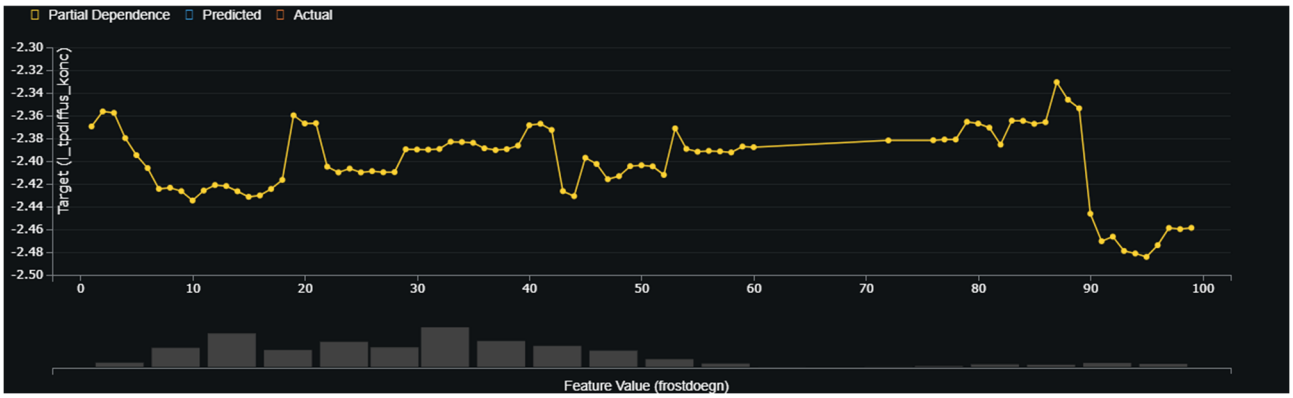


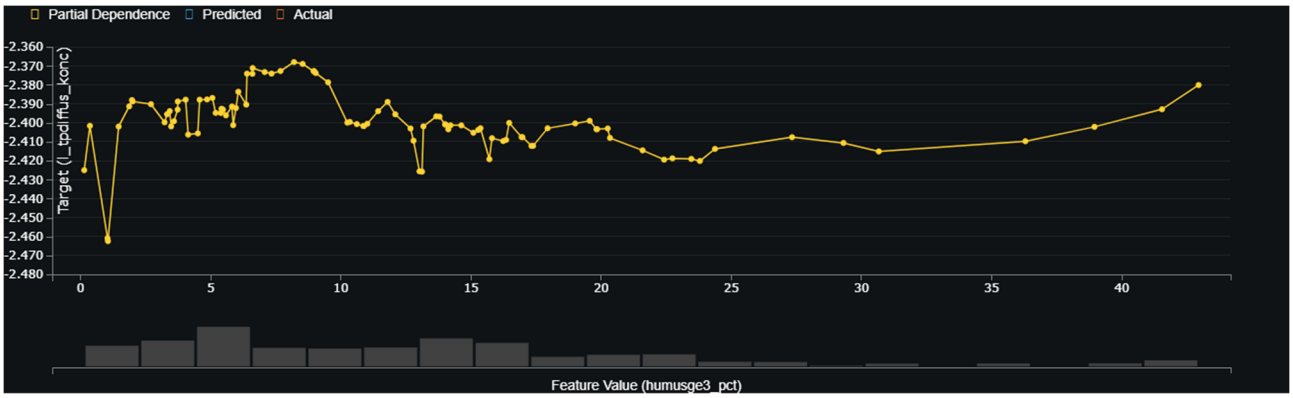



Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
-2.50
-2.52
-2.54
-2.56
450k
500k
550k
600k
650k
700k
750k
800k
850k
900k
Feature Value (centx)

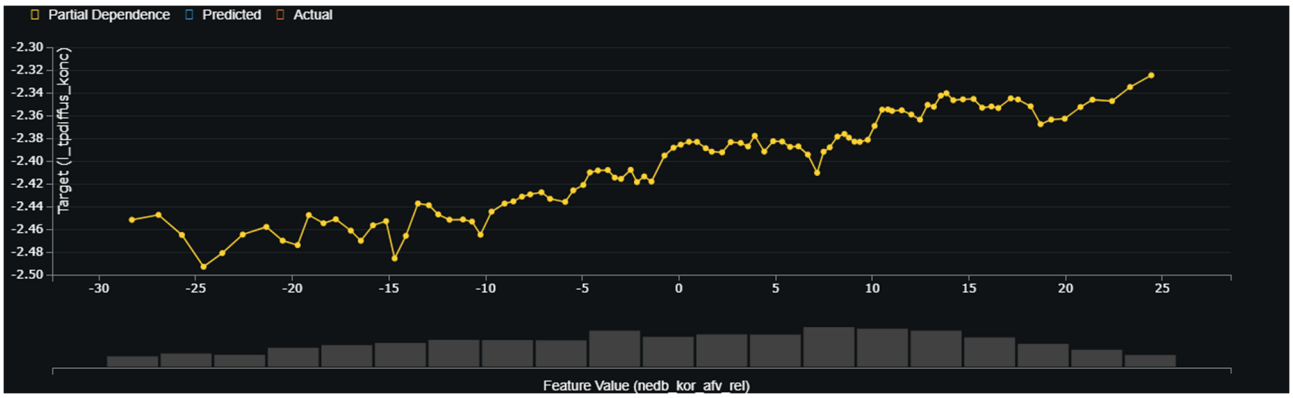
Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.22
-2.24
-2.26
-2.28
-2.30
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
6.06M
6.08M
6.1M
6.12M
6.14M
6.16M
6.18M
6.2M
6.22M
6.24M
6.26M
6.28M
6.3M
6.32M
6.34M
6.36M
Feature Value (Centy)

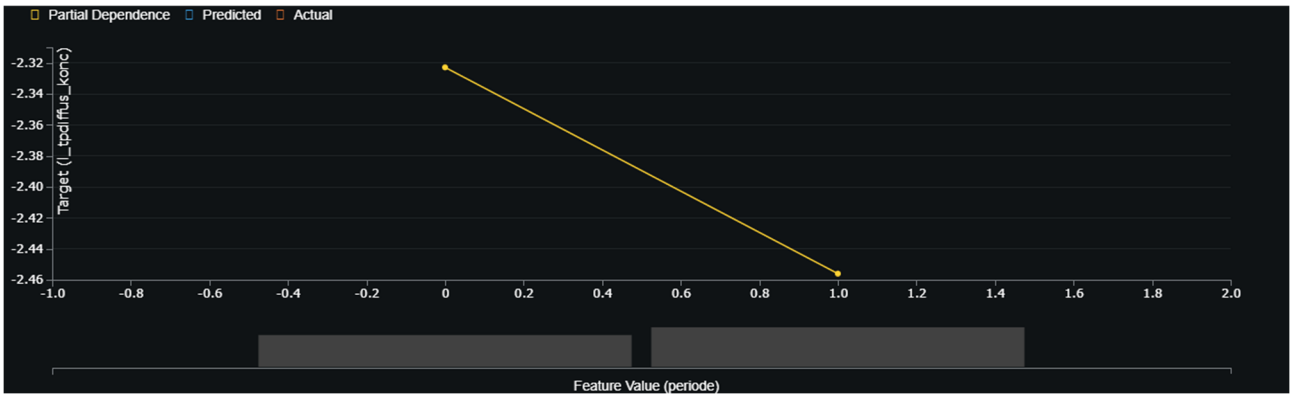

Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.30
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
2
4
6
8
10
12
14
16
18
20
Feature Value (d_claypct)

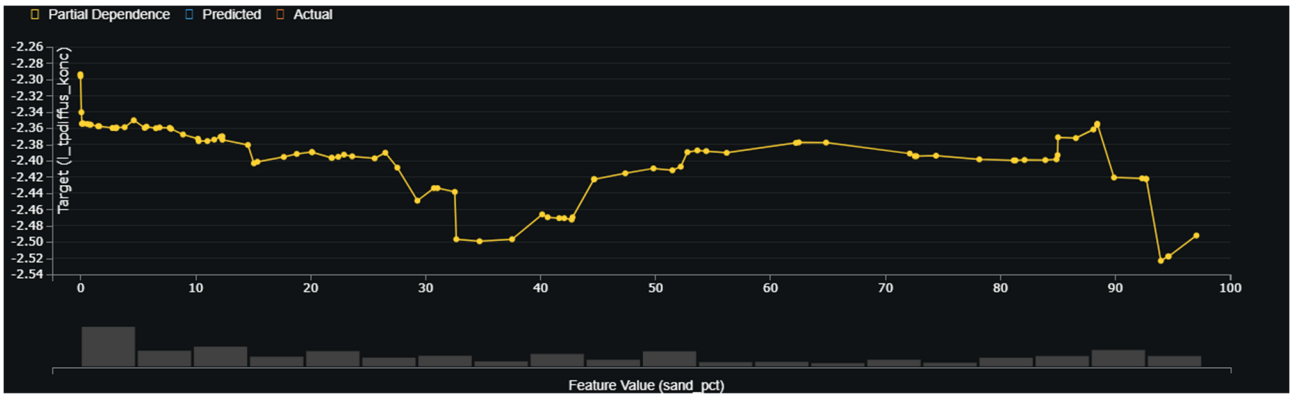

Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.30
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
-2.50
0
10
20
30
40
50
60
70
80
90
100
Feature Value (frostdoegn)


Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.360
-2.370
-2.380
-2.390
-2.400
-2.410
-2.420
-2.430
-2.440
-2.450
-2.460
-2.470
-2.480
0
5
10
15
20
25
30
35
40
Feature Value (humusge3_pct)

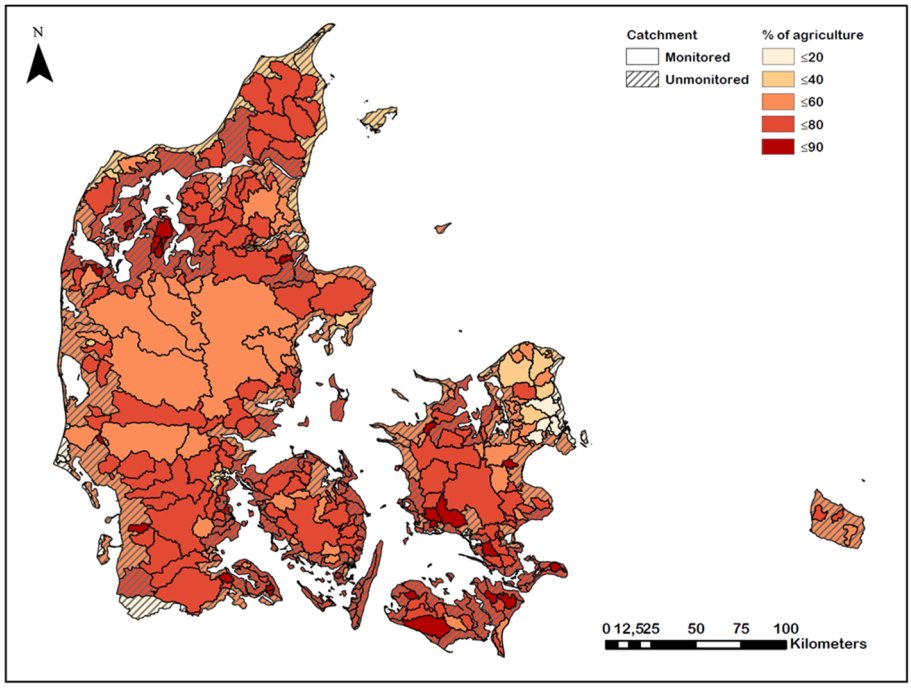
Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.35
-2.36
-2.37
-2.38
-2.39
-2.40
-2.41
-2.42
-2.43
-2.44
-2.45
-2.46
6.30
6.40
6.50
6.60
6.70
6.80
6.90
7.00
7.10
Feature Value (l_nedb_kor_mm)

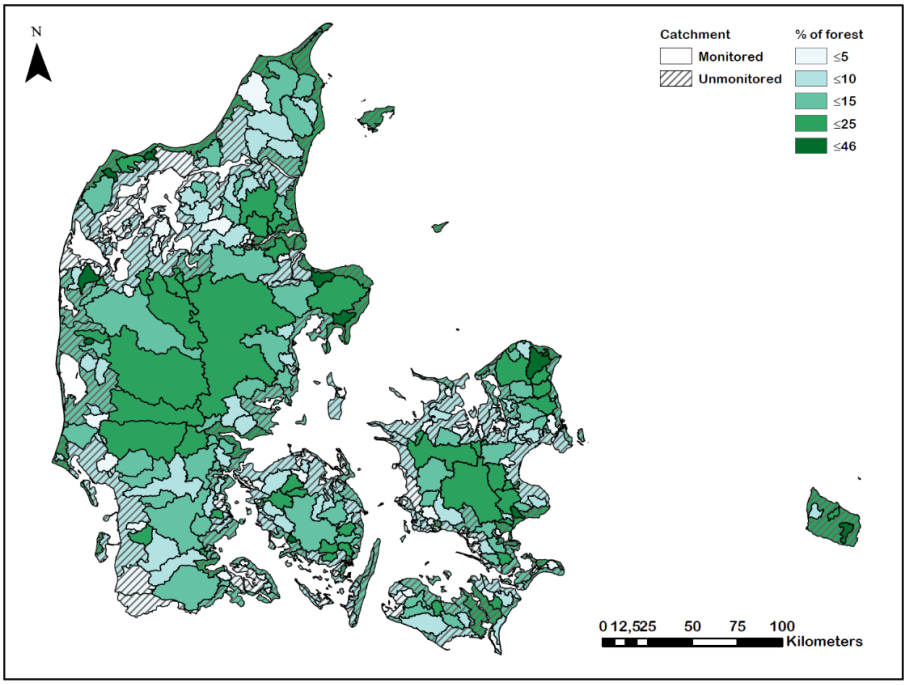

Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.30
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
-2.50
-30
-25
-20
-15
-10
-5
0
5
10
15
20
25
Feature Value (nedb_kor_afv_rel)

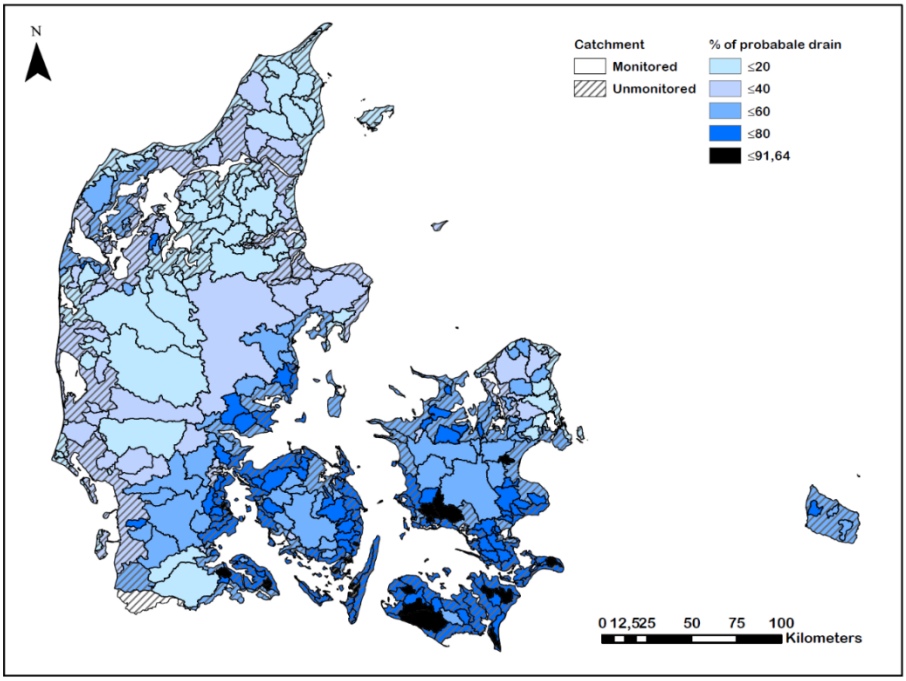

Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-1.0
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Feature Value (periode)

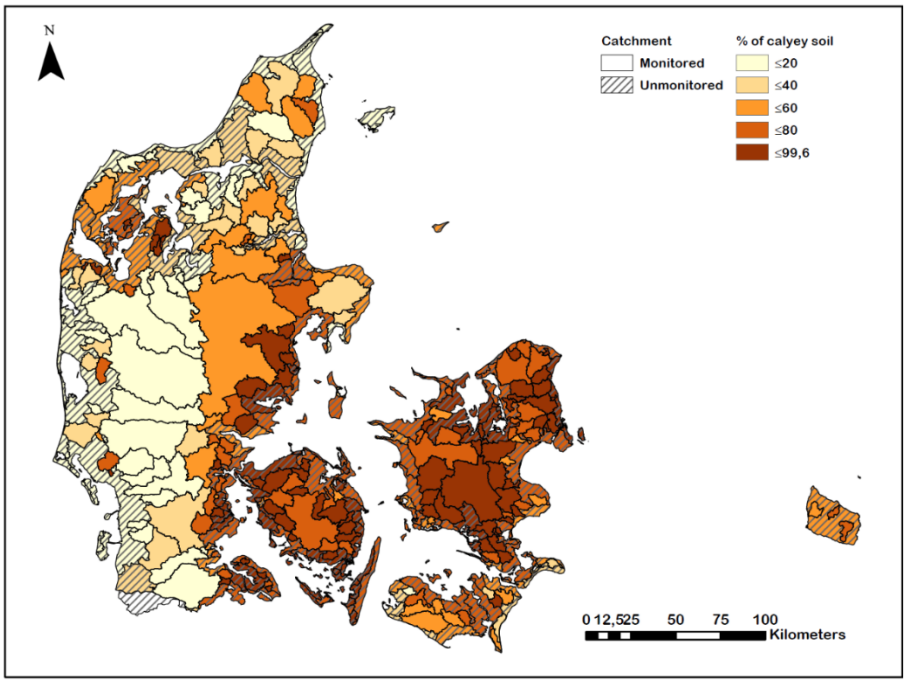

Partial Dependence
Predicted
Actual
Target (l_tpdiffus_konc)
-2.26
-2.28
-2.30
-2.32
-2.34
-2.36
-2.38
-2.40
-2.42
-2.44
-2.46
-2.48
-2.50
-2.52
-2.54
0
10
20
30
40
50
60
70
80
90
100
Feature Value (sand_pct)

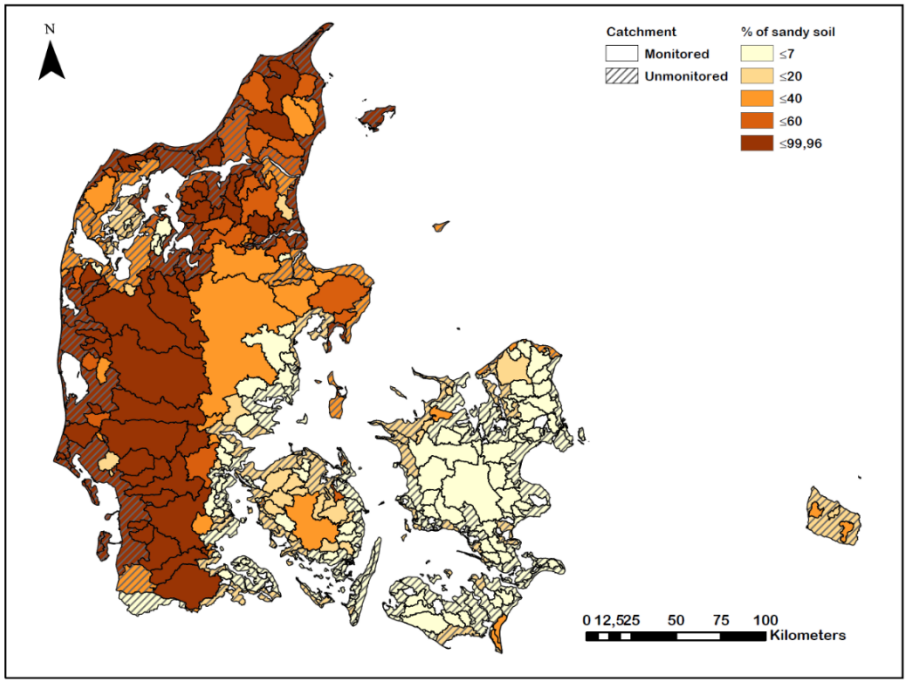
Bilag 6: Sammenligning mellem arealanvendelse og jordtyper for målte og umålte arealer defineret ud fra NOVANA-programmets havbelastnings-stationsnet

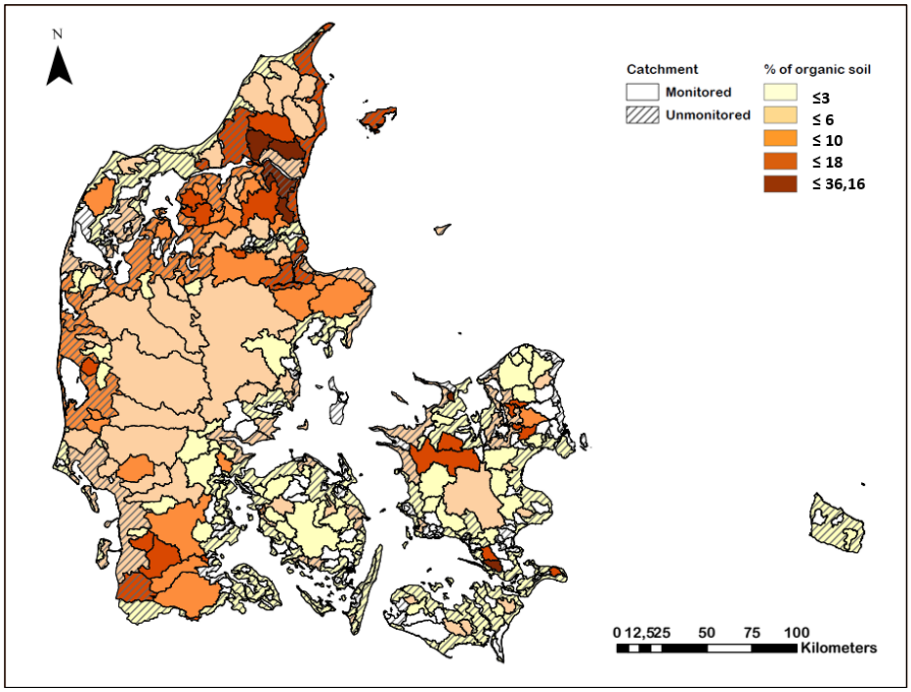





N
Catchment
Monitored
Unmonitored
% of probabale drain
≤20
≤40
≤60
≤80
≤91,64
0 12,525 50 75 100
Kilometers


N
Catchment
Monitored
Unmonitored
% of calyey soil
≤20
≤40
≤60
≤80
≤99,6
0 12,525 50 75 100
Kilometers


N
Catchment
Monitored
Unmonitored
% of sandy soil
≤7
≤20
≤40
≤60
≤99,96
0 12,525 50 75 100
Kilometers


N
Catchment
Monitored
Unmonitored
% of organic soil
≤3
≤ 6
≤ 10
≤ 18
≤ 36,16
0 12,525 50 75 100
Kilometers

NY FOSFORMODEL TIL ESTIMERING AF ÅRLIG VANDFØRINGSVÆGTET KONCENTRATION AF TOTAL FOSFOR FRA DIFFUSE KILDER I ID15-OPLANDE

En ny model til estimering af den årlige vandføringsvægtede total fosfor (TP) koncentration i vandløb i ID15-oplande er udviklet på stikprøvedata ved brug af maskinlærings softwaren DataRobot på baggrund af data fra 207 typeoplande. Den nye model er af typen 'eXtreme Gradient Boosted Trees Regressor with early stopping' har en forklaringsgrad (R^2) på 0,67-0,71 på træningsdatasættene i DataRobot. Resultater fra et uafhængigt valideringsdatasæt giver en forklaringsgrad (R^2) på 0,62 for modellen og på 0,41 efter tilbagetransformering, hvilket er en klar forbedring ift. den nuværende model, der kun bygger på data fra 24 vandløb med intensiv prøvetagning. Ved skift til den nye model vil den årlige fosfortilførsel til kystvandene i gennemsnit blive omkring 3 % lavere, end opgjort med den nuværende anvendte og biaskorrigerede TP-model.

ISBN: 978-87-7156-697-0
ISSN: 2244-999X