

GAM Development

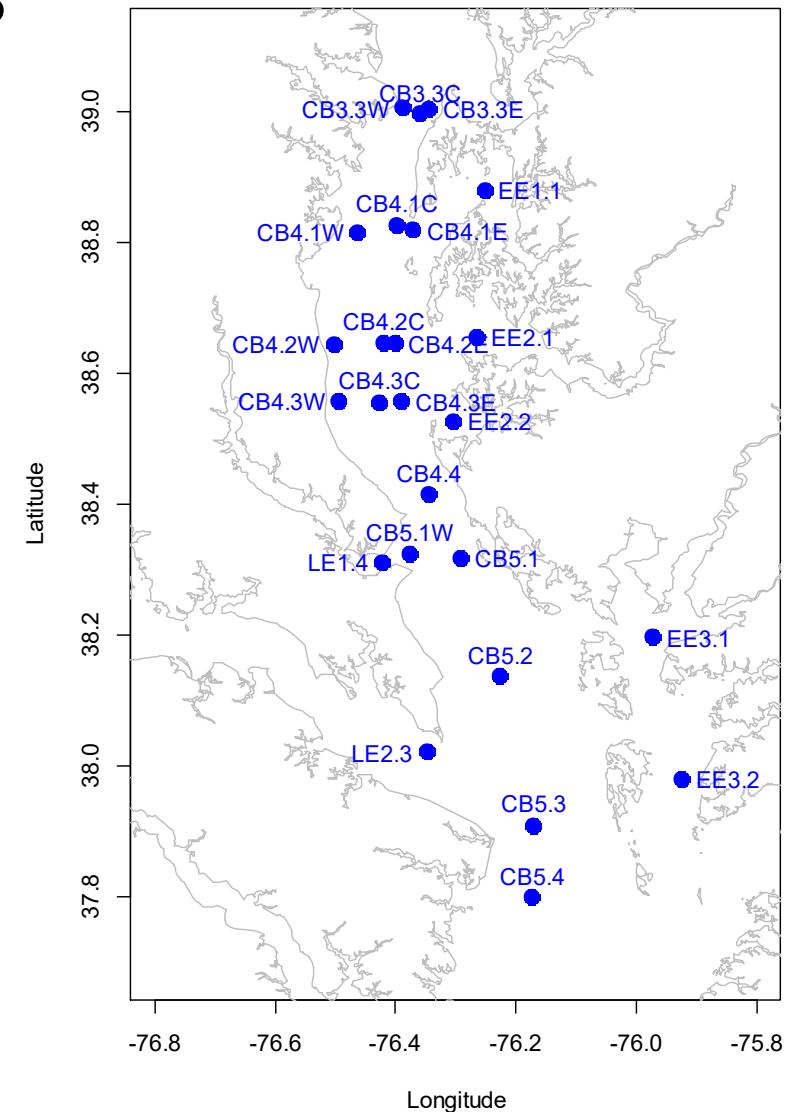
Station hold-out tests
and
Proposed k-value/wiggleness adjustments

Rebecca Murphy

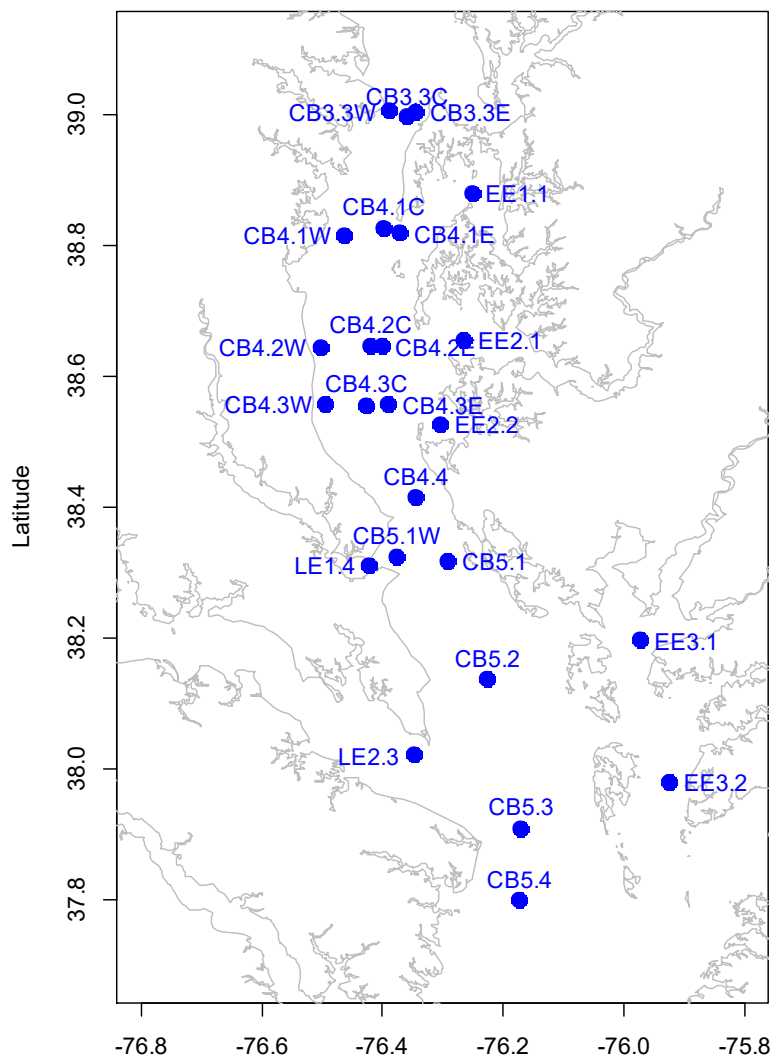
Nov 28, 2022

Station hold-out analysis steps

- One at a time, took out each station's data and refit the GAM without that station = “hold-out GAM”
 - Then used the hold-out GAM to predict at the held-out station for all dates & depths with DO samples.
 - Computed RMSE between the observations and the “hold-out” GAM's estimates.
 - Repeated for each station.
- Compared these to RMSEs computed from the same GAM fit using the entire data set (including that station).

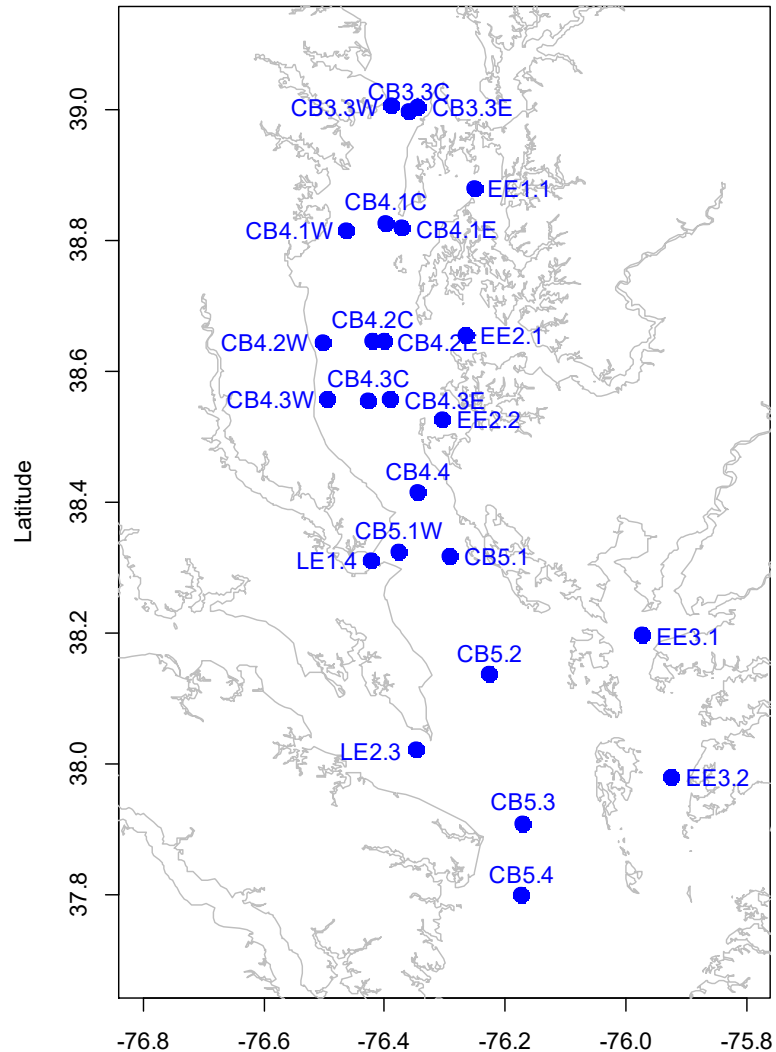


Compare RMSE from DO predictions with and without each station in fitted model



Station	RMSE from gs7 fit with all data	RMSE from gs7.cv model fit holding out this station's data
CB3.3C	1.50	5.06
CB3.3E	1.67	4.87
CB3.3W	1.77	4.17
CB4.1C	1.36	2.43
CB4.1E	1.42	2.50
CB4.1W	1.68	4.42
CB4.2C	1.33	1.35
CB4.2E	1.47	1.60
CB4.2W	1.85	3.00
CB4.3C	1.25	3.36
CB4.3E	1.55	1.73
CB4.3W	1.67	2.35
CB4.4	1.29	1.35
CB5.1	1.33	1.44
CB5.1W	1.10	12.55
CB5.2	1.25	1.72
CB5.3	1.10	1.40
CB5.4	1.22	2.90
EE1.1	1.57	1.71
EE2.1	1.05	1.72
EE2.2	1.79	3.07
EE3.1	1.05	12.99
EE3.2	1.13	6.15
LE1.4	1.23	27.07
LE2.3	1.43	26.59

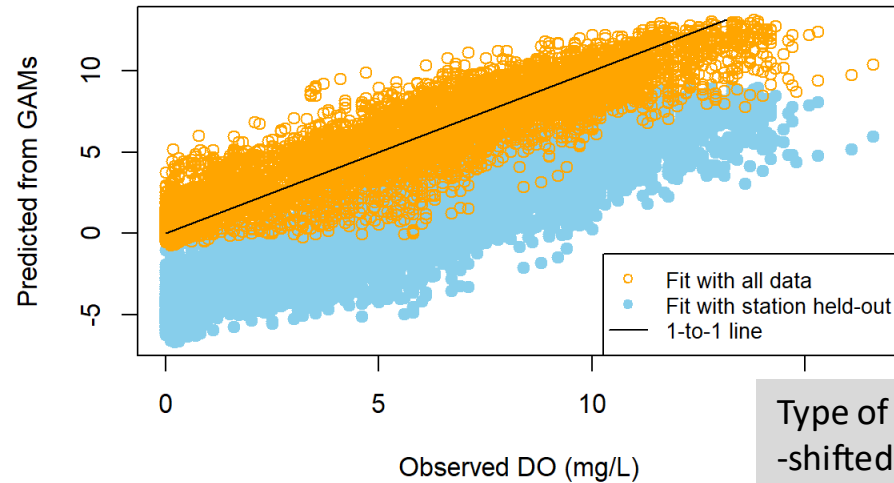
Some very high RMSEs



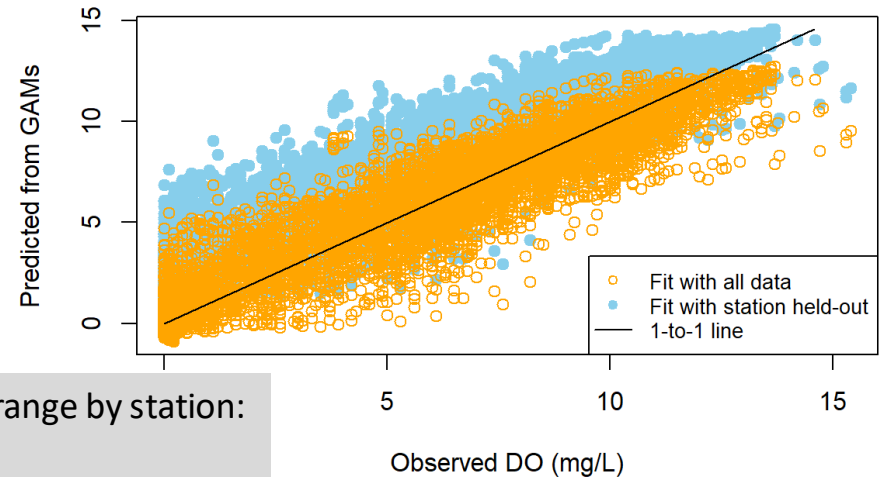
Station	RMSE from gs7 fit with all data	RMSE from gs7.cv model fit holding out this station's data
CB3.3C	1.50	5.06
CB3.3E	1.67	4.87
CB3.3W	1.77	4.17
CB4.1C	1.36	2.43
CB4.1E	1.42	2.50
CB4.1W	1.68	4.42
CB4.2C	1.33	1.35
CB4.2E	1.47	1.60
CB4.2W	1.85	3.00
CB4.3C	1.25	3.36
CB4.3E	1.55	1.73
CB4.3W	1.67	2.35
CB4.4	1.29	1.35
CB5.1	1.33	1.44
CB5.1W	1.10	12.55
CB5.2	1.25	1.72
CB5.3	1.10	1.40
CB5.4	1.22	2.90
EE1.1	1.57	1.71
EE2.1	1.05	1.72
EE2.2	1.79	3.07
EE3.1	1.05	12.99
EE3.2	1.13	6.15
LE1.4	1.23	27.07
LE2.3	1.43	26.59

What do these predictions look like?

Hold-out RMSE = 5.06 **CB3.3C**

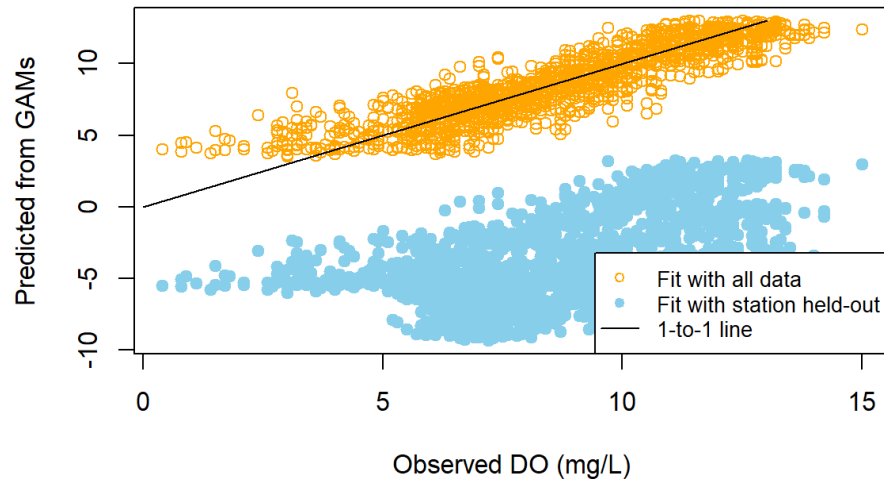


CB4.1C Hold-out RMSE = 2.43

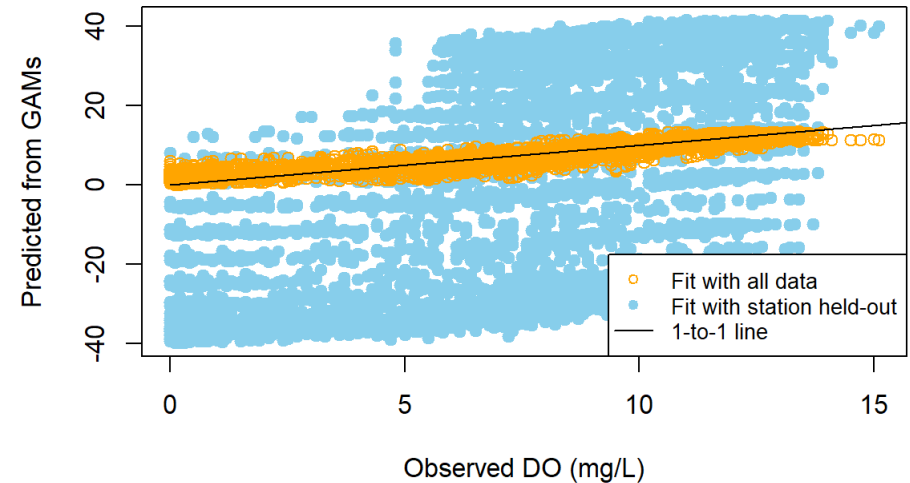


Type of erroneous predictions range by station:
-shifted a few mg/L high or low
-shifted dramatically lower
-generating a huge range

Hold-out RMSE = 12.6 **CB5.1W**



LE2.3 Hold-out RMSE = 26.6



Some possibilities

- Extrapolation at boundaries
- Distribution is allowing predictions <0 for DO
- GAM is overfitting to this specific data
 - Need to investigate if k-values are too high on some of the splines.
- Combination of factors possible of course

About k-values

- In the R package we are using, mgcv, the k-value is a user-specified parameter on each spline which effectively sets an upper limit on flexibility.
- When not specified, defaults for s() is k=10 and ti() terms are k=5.
- Testing is recommended for these values. We had at first suspected we would need to increase k-values to capture more variability.

gs7 model we've been using:

```
value ~ s(cyear, k = 20) + s(doy,k=10) +  
  s(wDepth,k=10) + s(LonKm,k=10) + s(LatKm,k=10) + s(bDepth,k=10) +  
  ti(LatKm, wDepth, LonKm,k=5) + ti(wDepth, bDepth,k=5) +  
  ti(wDepth, cyear,k=5) + ti(LonKm, cyear,k=5) + ti(LatKm, cyear,k=5) + ti(bDepth, cyear,k=5) +  
  ti(wDepth, doy,k=5) + ti(LonKm, doy,k=5) + ti(LatKm, doy,k=5) + ti(bDepth, doy,k=5)+  
  ti(cyear, doy,k=5)
```

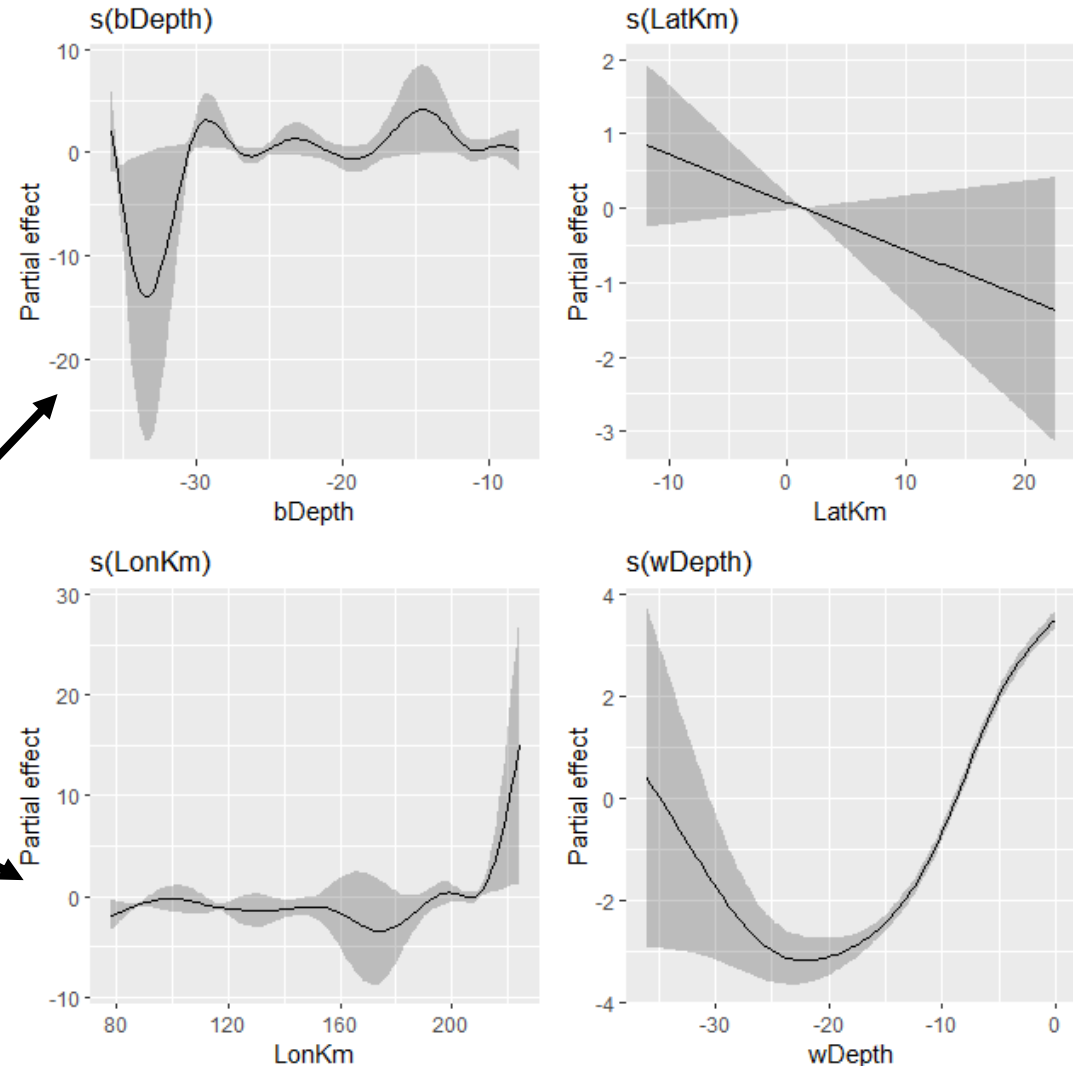
About k-values

Partial effect graphs showing the shape of the spline relationships can help with this analysis:

Do the relationships meet our science-based knowledge of the system?

Higher k allows these component graphs to change direction multiple times/be more wiggly:

Partial effects graphs from spatial single-splines for **gs7** model



Test reductions in spatial-term k-values

Did a suite of tests:

- Increased and decreased the k by 1 on each of the single terms, down to 3 and up to 20
- For interactions, k=3 is the lowest it goes, also tried 4 and several increases
- Did some testing on the interactive temporal terms, and so far see no difference (but may need to hold out months using Elgin's approach)

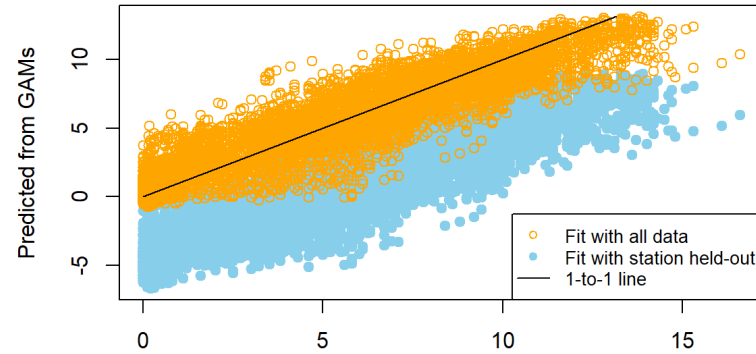
→ **Results: Much better hold-out results (and spline shapes) by reducing the spatial-terms' single splines to $k=5$ and interactions to $k=3$**

value $\sim$ s(cyear, k = 20) + s(doy, k=10) +
s(wDepth, $k=5$) + s(LonKm, $k=5$) + s(LatKm, $k=5$) + s(bDepth, $k=5$) +
ti(LatKm, wDepth, LonKm, $k=3$) + ti(wDepth, bDepth, $k=3$) +
ti(wDepth, cyear, k=5) + ti(LonKm, cyear, k=5) + ti(LatKm, cyear, k=5) + ti(bDepth, cyear, k=5) +
ti(wDepth, doy, k=5) + ti(LonKm, doy, k=5) + ti(LatKm, doy, k=5) + ti(bDepth, doy, k=5) +
ti(cyear, doy, k=5)

Compare
original GAM
with default k's
to “Reduced k-
model”

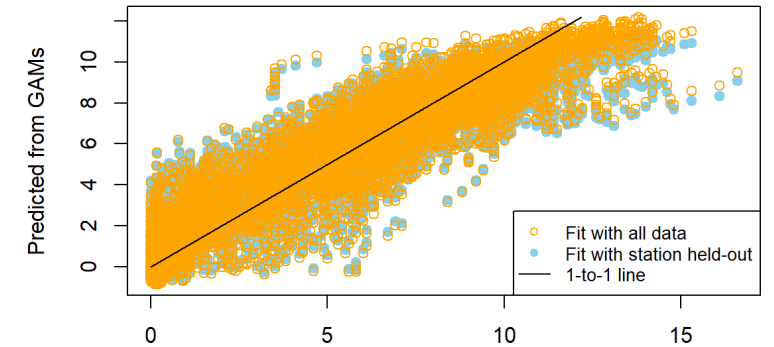
Original GAM

CB3.3C

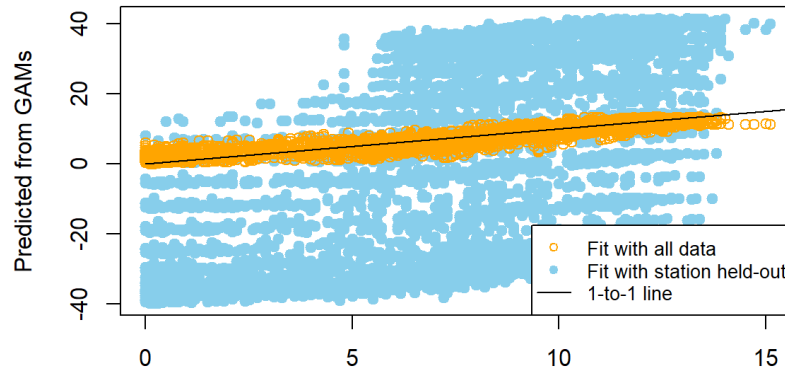


Reduced k-GAM

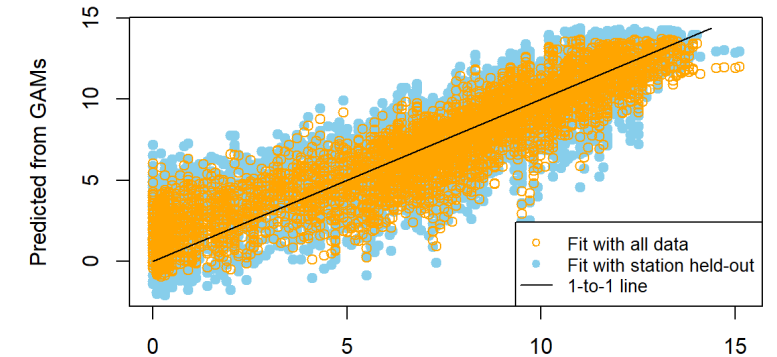
CB3.3C: Reduced k model



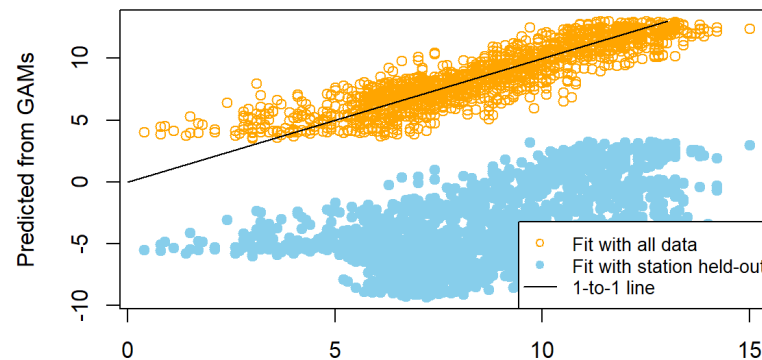
LE2.3



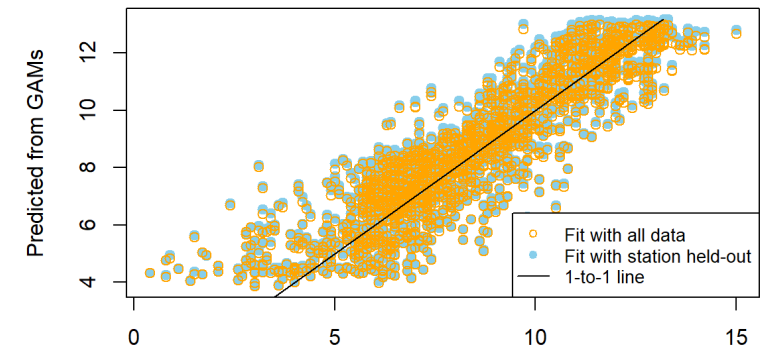
LE2.3: Reduced k model



CB5.1W



CB5.1W: Reduced k model



Observed DO (mg/l)

Station RMSEs from hold-out tests

Model	R2	AIC
Gs7	0.863	414184.2

	Gs7 – original GAM	
Station	RMSE from gs7 fit with all data	RMSE from gs7.cv model fit holding out this station's data
CB3.3C	1.50	5.06
CB3.3E	1.67	4.87
CB3.3W	1.77	4.17
CB4.1C	1.36	2.43
CB4.1E	1.42	2.50
CB4.1W	1.68	4.42
CB4.2C	1.33	1.35
CB4.2E	1.47	1.60
CB4.2W	1.85	3.00
CB4.3C	1.25	3.36
CB4.3E	1.55	1.73
CB4.3W	1.67	2.35
CB4.4	1.29	1.35
CB5.1	1.33	1.44
CB5.1W	1.10	12.55
CB5.2	1.25	1.72
CB5.3	1.10	1.40
CB5.4	1.22	2.90
EE1.1	1.57	1.71
EE2.1	1.05	1.72
EE2.2	1.79	3.07
EE3.1	1.05	12.99
EE3.2	1.13	6.15
LE1.4	1.23	27.07
LE2.3	1.43	26.59

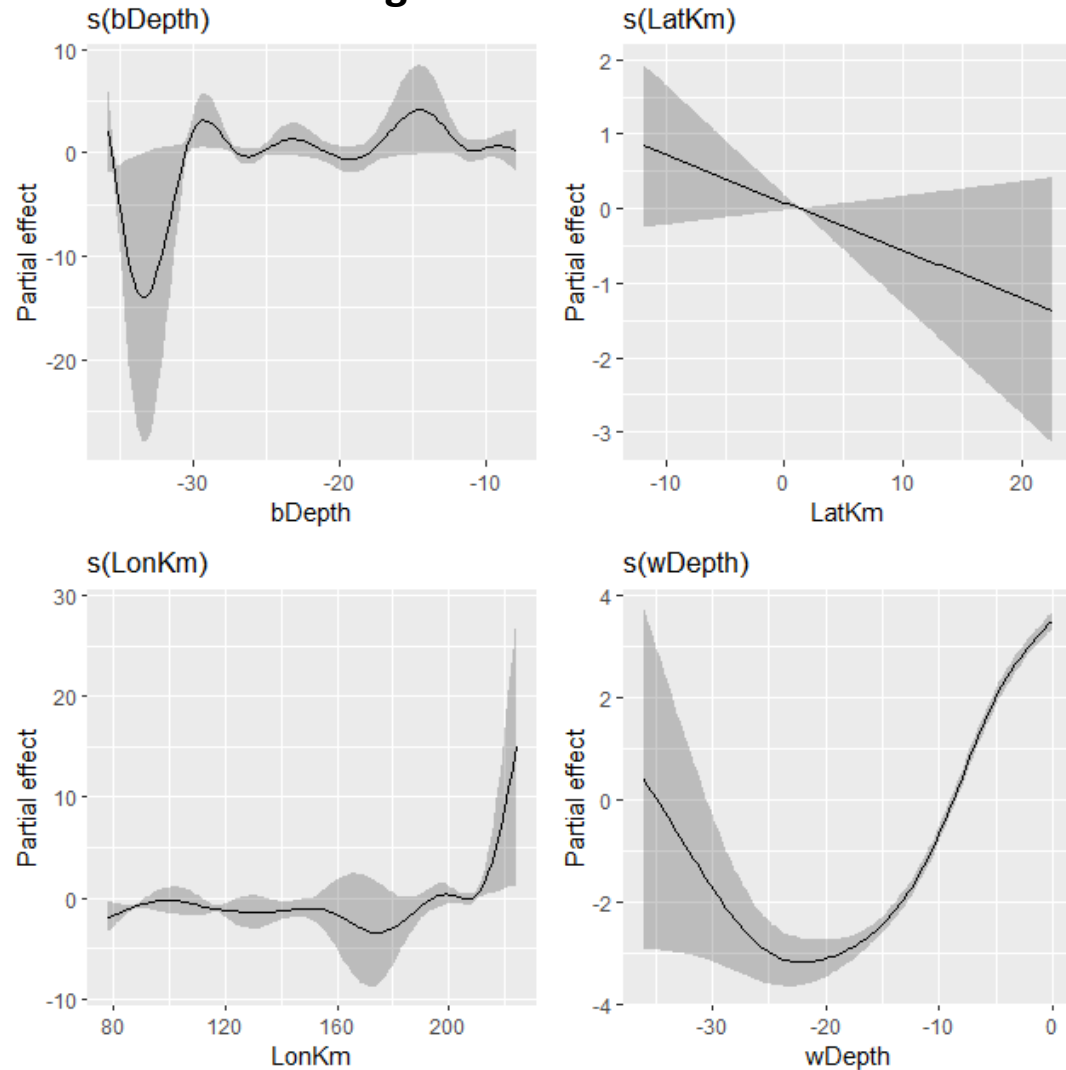
Station RMSEs from hold-out tests

Model	R2	AIC
Gs7	0.863	414184.2
Gs7b (reduce spatial k-values)	0.851	423735.2

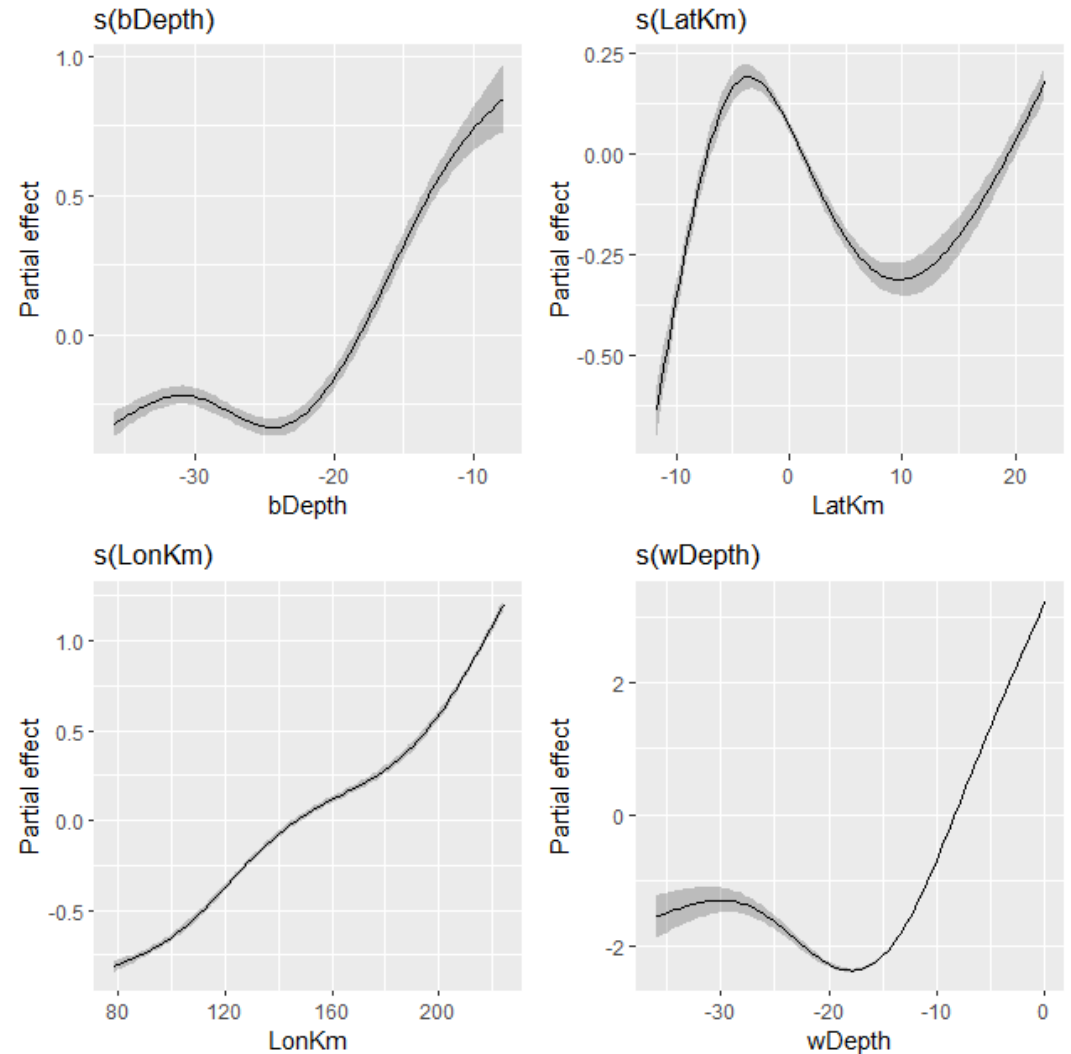
	Gs7 – original GAM		Gs7b – reduced k-GAM	
Station	RMSE from gs7 fit with all data	RMSE from gs7.cv model fit holding out this station's data	RMSE from gs7b fit with all data	RMSE from gs7b.cv model fit holding out this station's data
CB3.3C	1.50	5.06	1.56	1.63
CB3.3E	1.67	4.87	1.77	1.81
CB3.3W	1.77	4.17	1.81	1.82
CB4.1C	1.36	2.43	1.40	1.52
CB4.1E	1.42	2.50	1.44	1.47
CB4.1W	1.68	4.42	1.72	1.74
CB4.2C	1.33	1.35	1.37	1.39
CB4.2E	1.47	1.60	1.45	1.49
CB4.2W	1.85	3.00	1.85	1.86
CB4.3C	1.25	3.36	1.31	1.43
CB4.3E	1.55	1.73	1.57	1.60
CB4.3W	1.67	2.35	1.70	1.71
CB4.4	1.29	1.35	1.30	1.32
CB5.1	1.33	1.44	1.36	1.48
CB5.1W	1.10	12.55	1.13	1.17
CB5.2	1.25	1.72	1.31	1.40
CB5.3	1.10	1.40	1.32	1.50
CB5.4	1.22	2.90	1.33	1.55
EE1.1	1.57	1.71	1.67	1.72
EE2.1	1.05	1.72	1.16	1.26
EE2.2	1.79	3.07	1.83	1.97
EE3.1	1.05	12.99	1.21	1.69
EE3.2	1.13	6.15	1.14	2.60
LE1.4	1.23	27.07	1.27	1.38
LE2.3	1.43	26.59	1.48	1.88

Changes in the graphs

Original model with default k's



New model with reduced k's

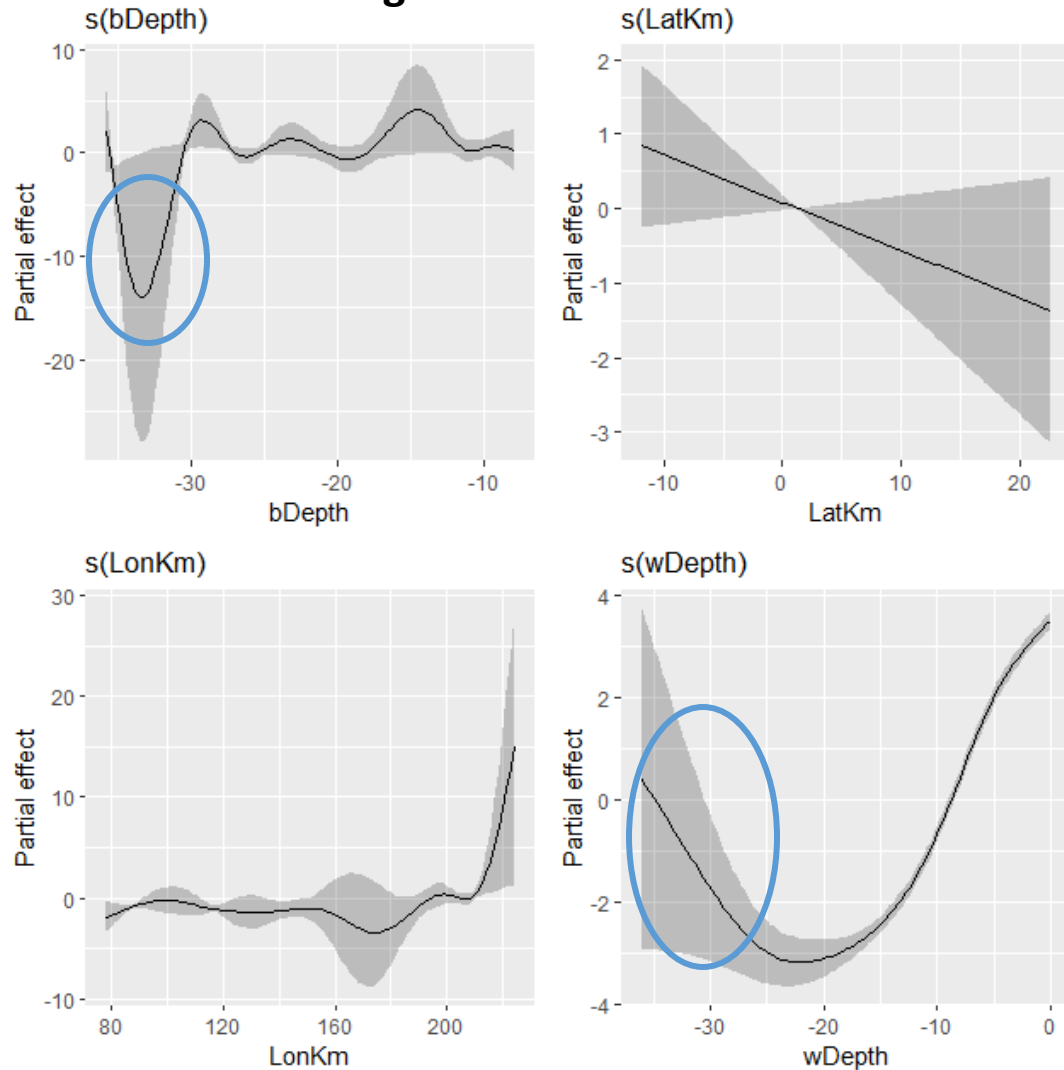


Bigger LonKm = further south

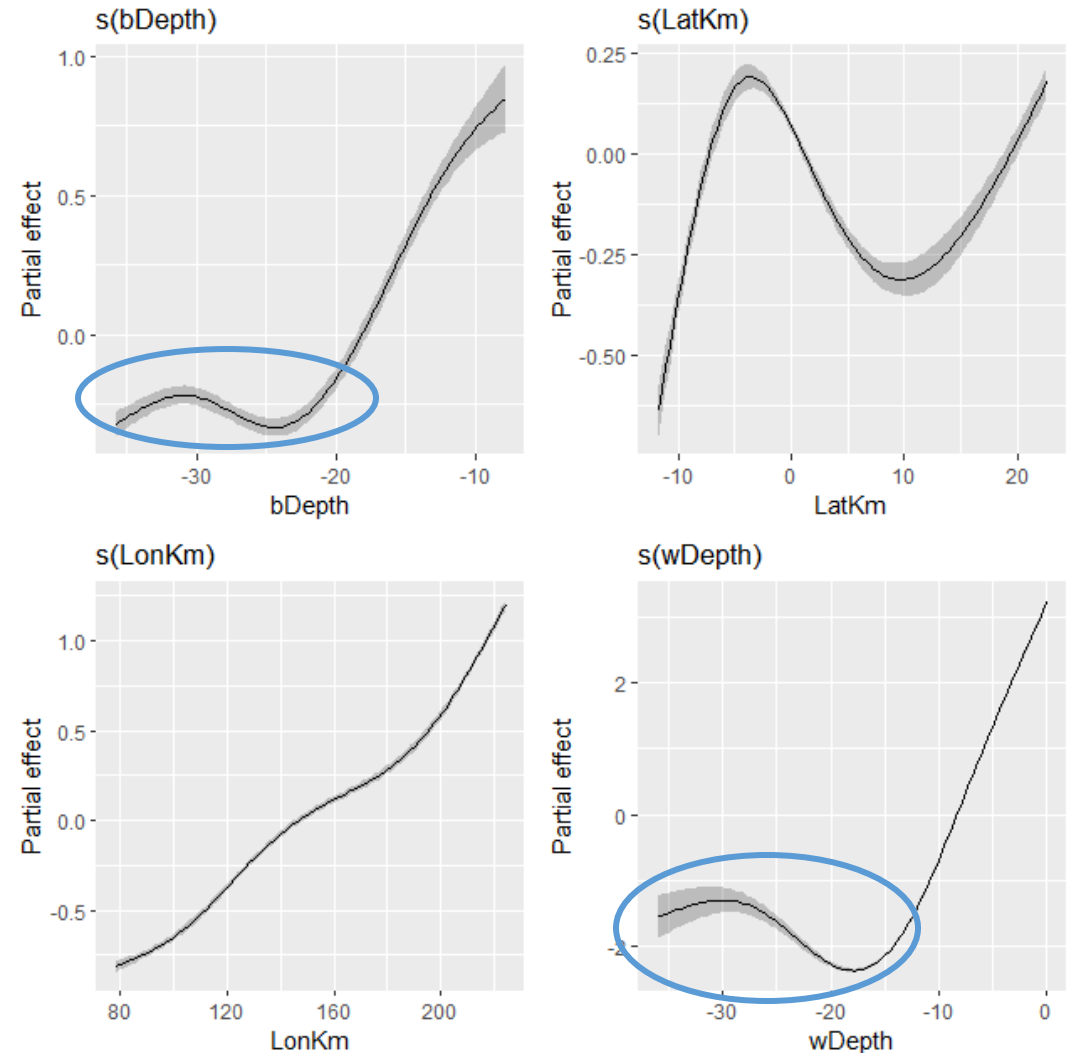
Changes in the graphs

Simplified shapes meet our expectations for DO patterns better

Original model with default k's



New model with reduced k's



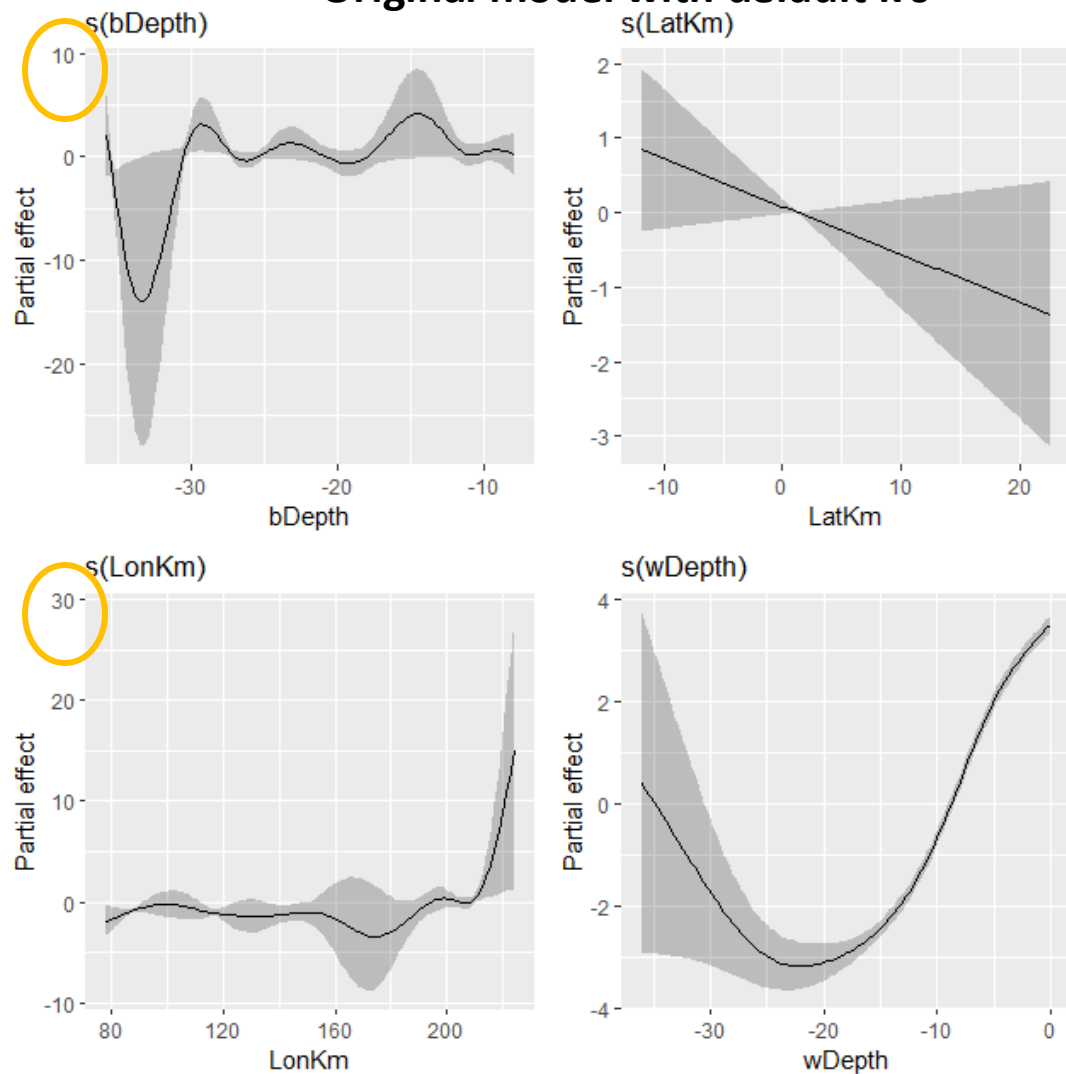
Bigger LonKm = further south

Changes in the graphs

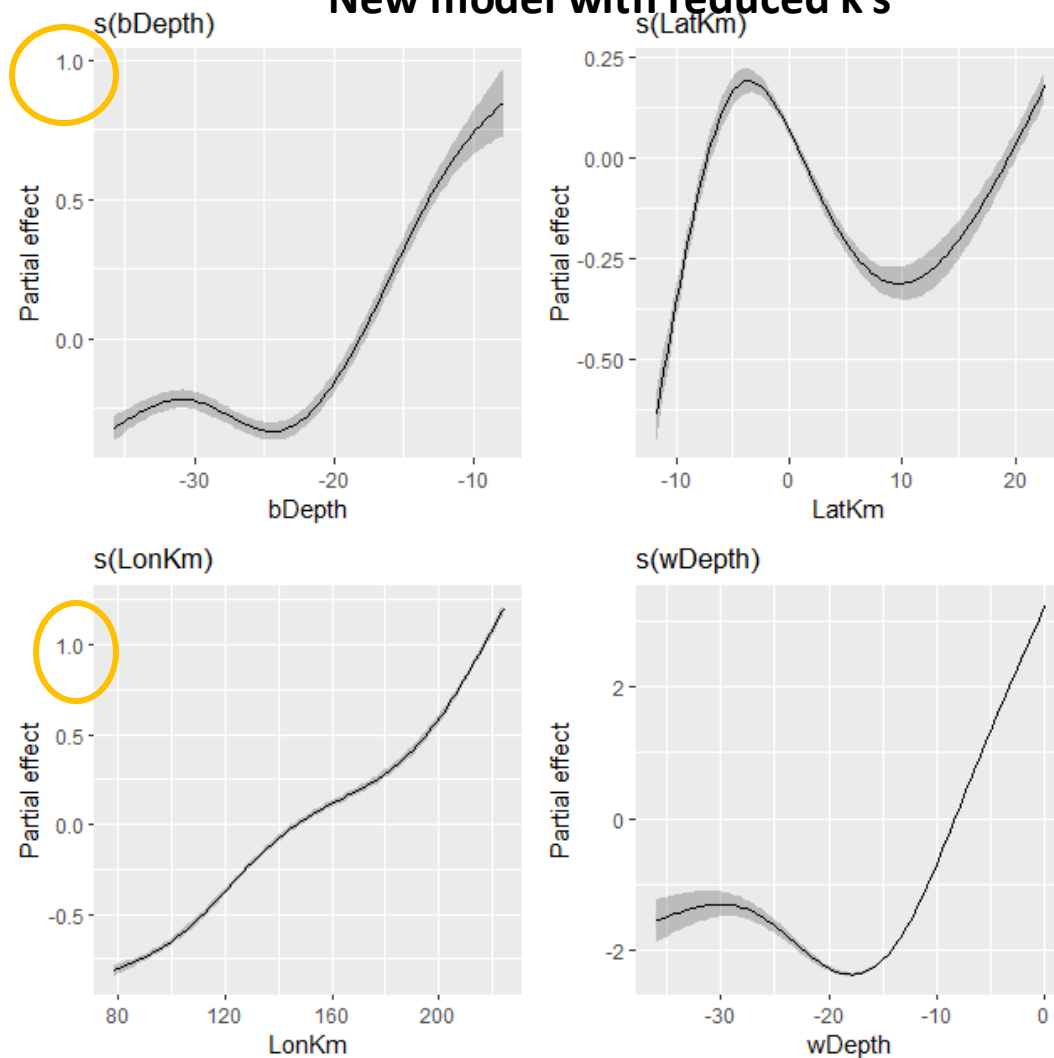
Relative magnitude of the partial effects
got more reasonable



Original model with default k's



New model with reduced k's

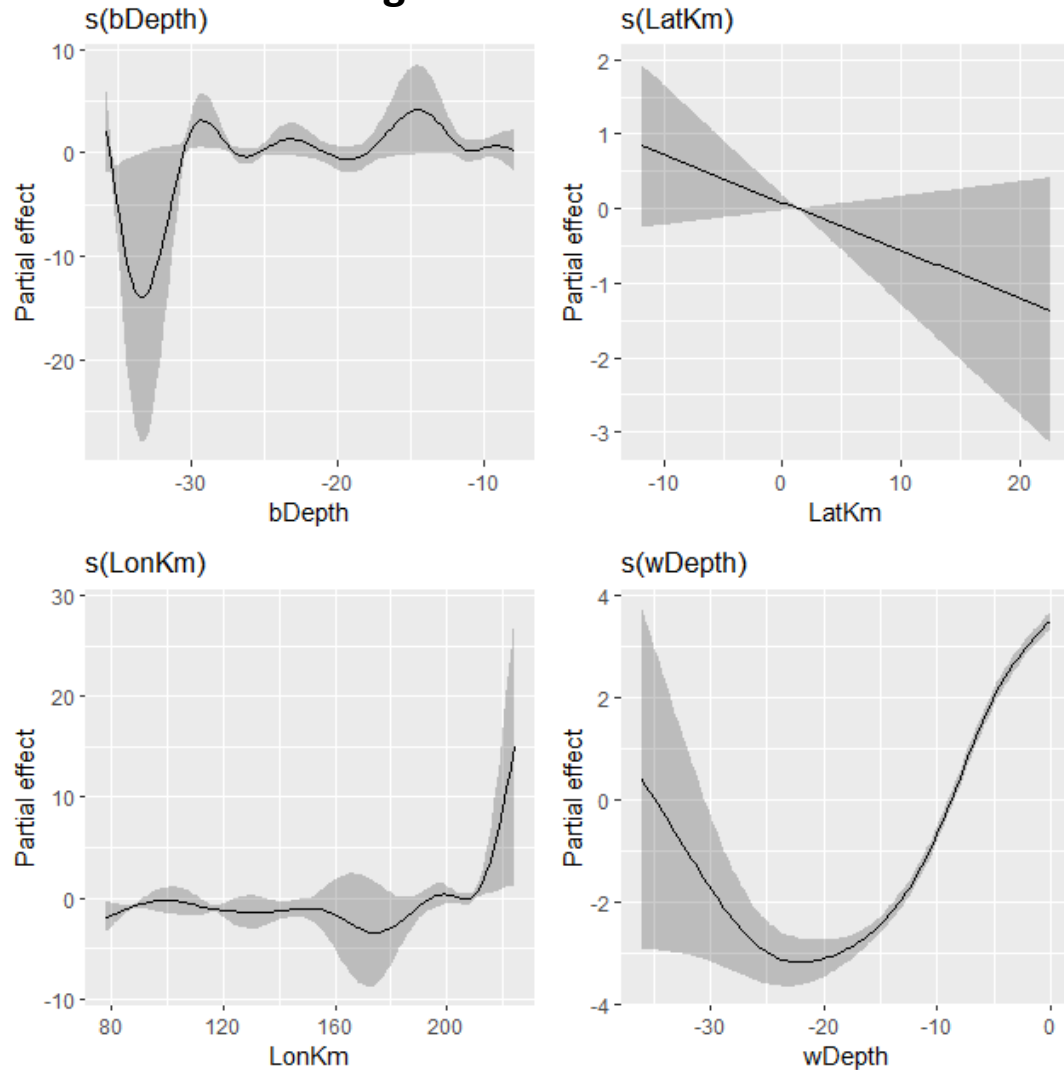


Bigger LonKm = further south

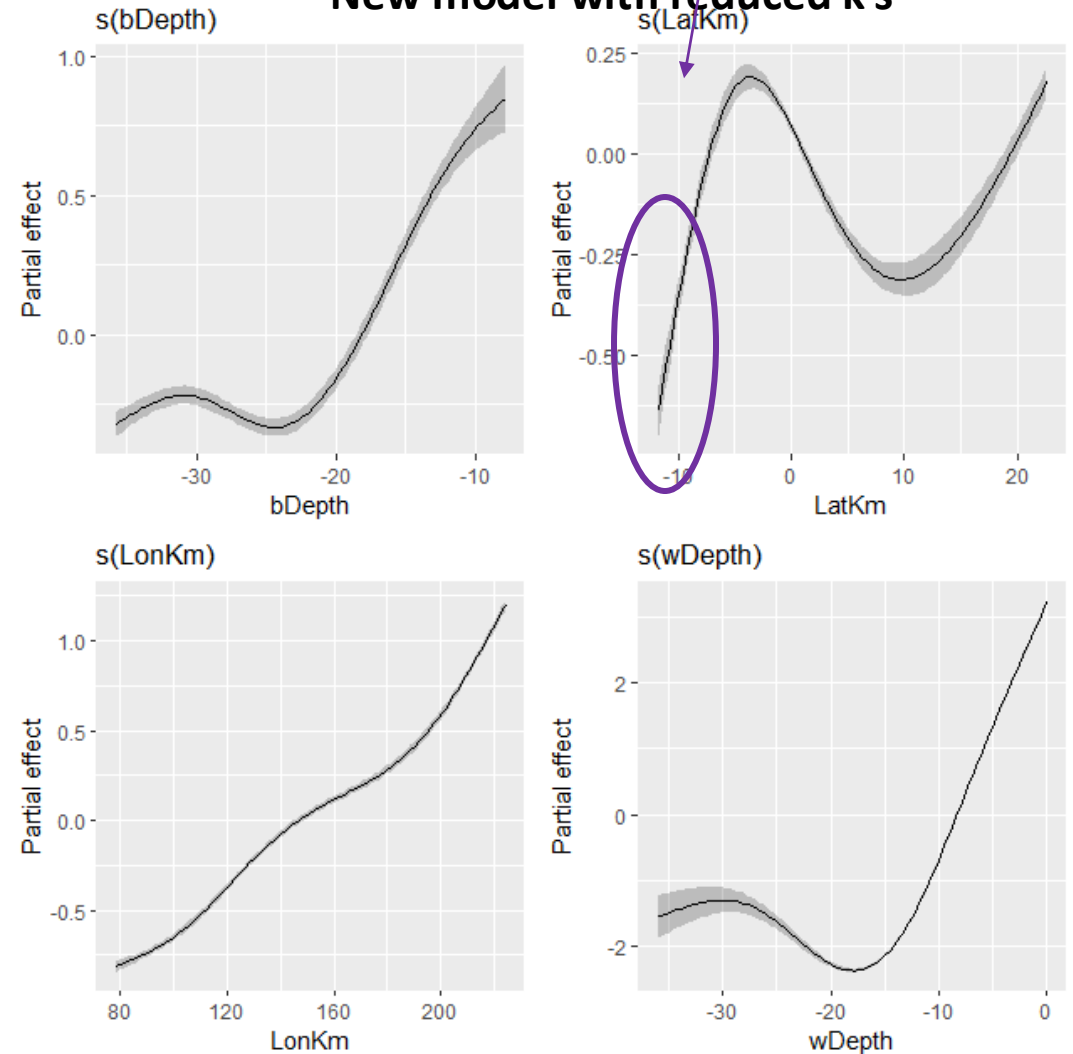
Changes in the graphs

Still some possible extrapolation problems
(but the magnitude of the possible impact
much smaller)

Original model with default k's



New model with reduced k's



Bigger LonKm = further south

Conclusions

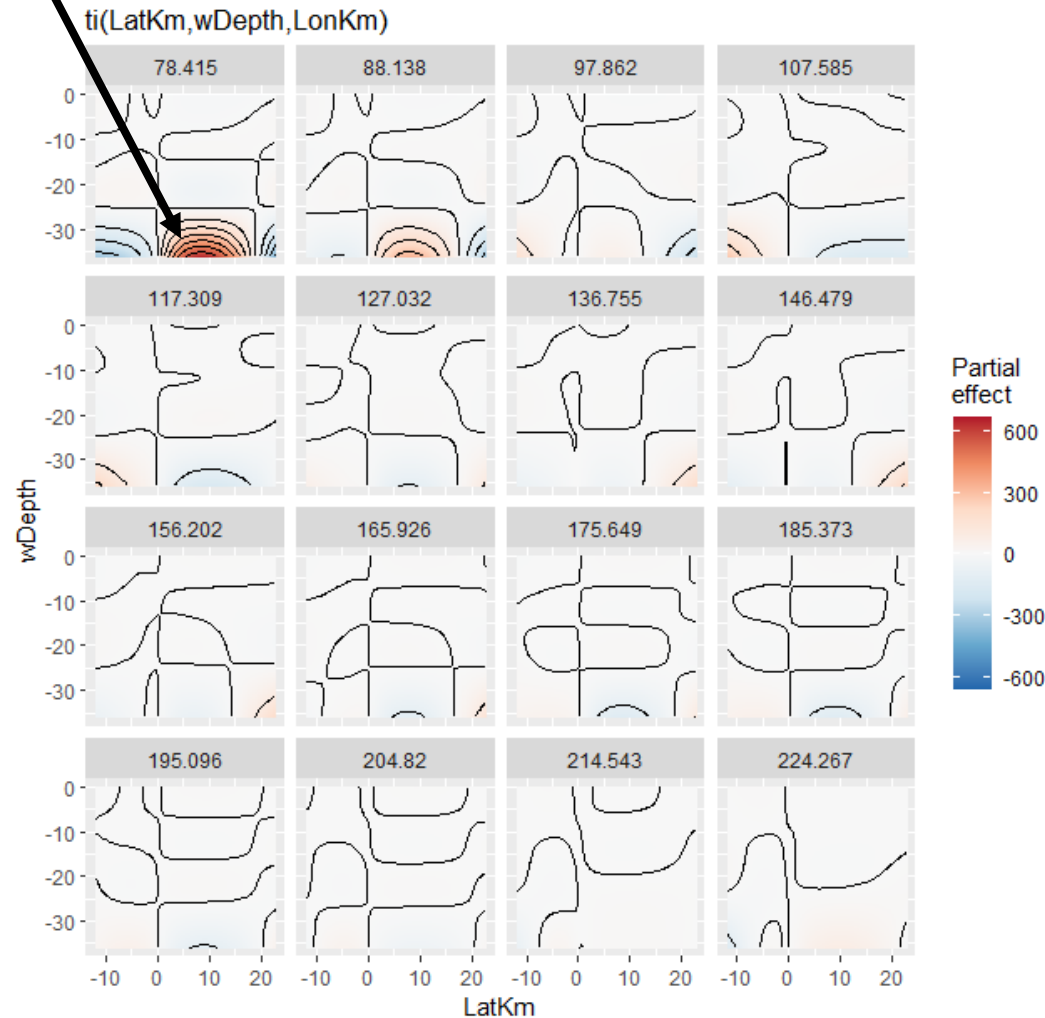
- For the spatial terms, the default k's allowed for too much flexibility, over-fitting to the data.
- Even with slight degradation of R^2 and AIC with reduced k's, this seems like an important modification to make since our purpose is interpolation.
- We now suspect that it is the missing explanatory variables (flow, wind, tide, temperature, etc) that would be needed to account for the unexplained variability, not additional flexibility in the spatial model.
- Remaining steps:
 - Finish exploring impacts of changing k-values in temporal terms.
 - If variables are added to the model, consider if k-value adjustment is needed again.
 - Do some k-testing with expansion into tributaries as well.

extras

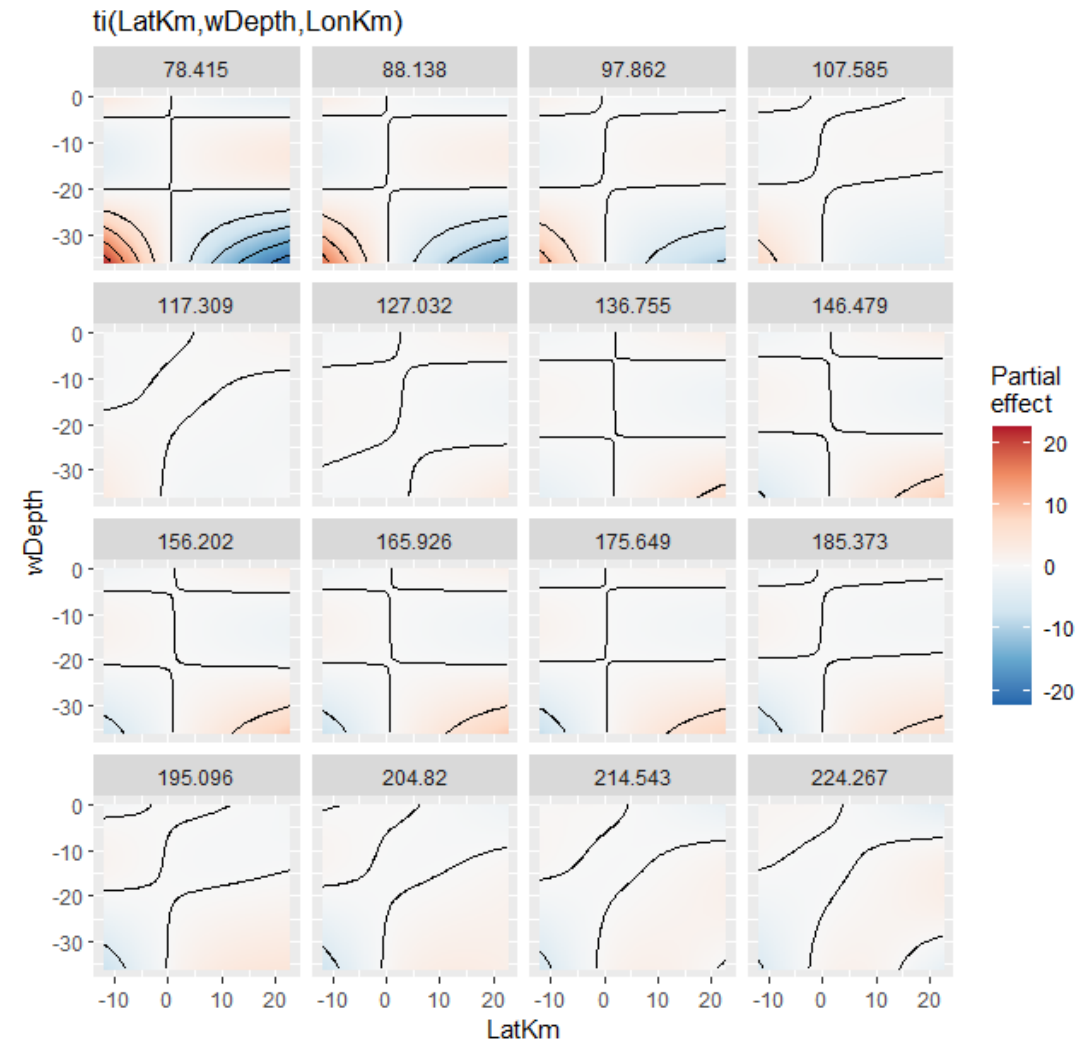
Reduced k change in ti(LatKm,wDepth,LonKm) graphic:

Higher k also allows for regions with large effect to occur in the interactive surfaces. This is deeper than the station here.

Original model with default k's

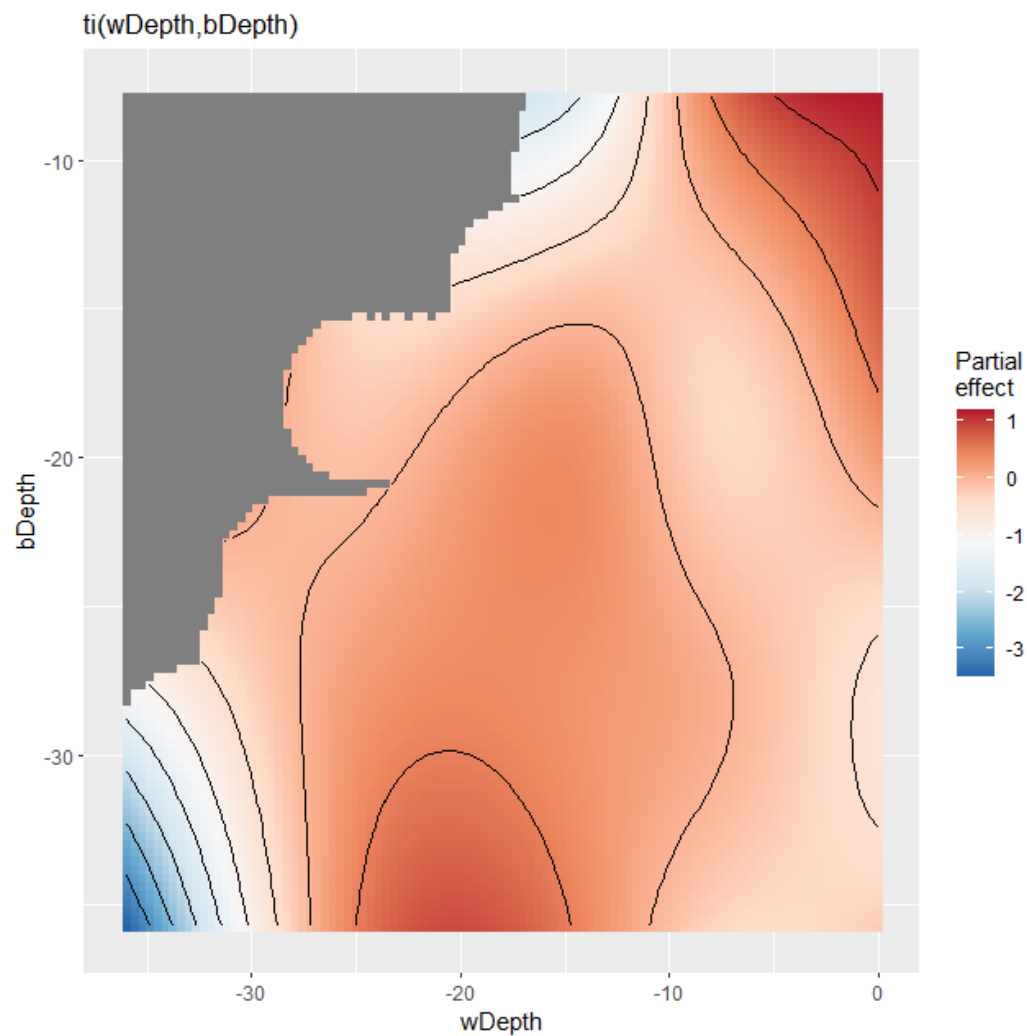


New model with reduced k's

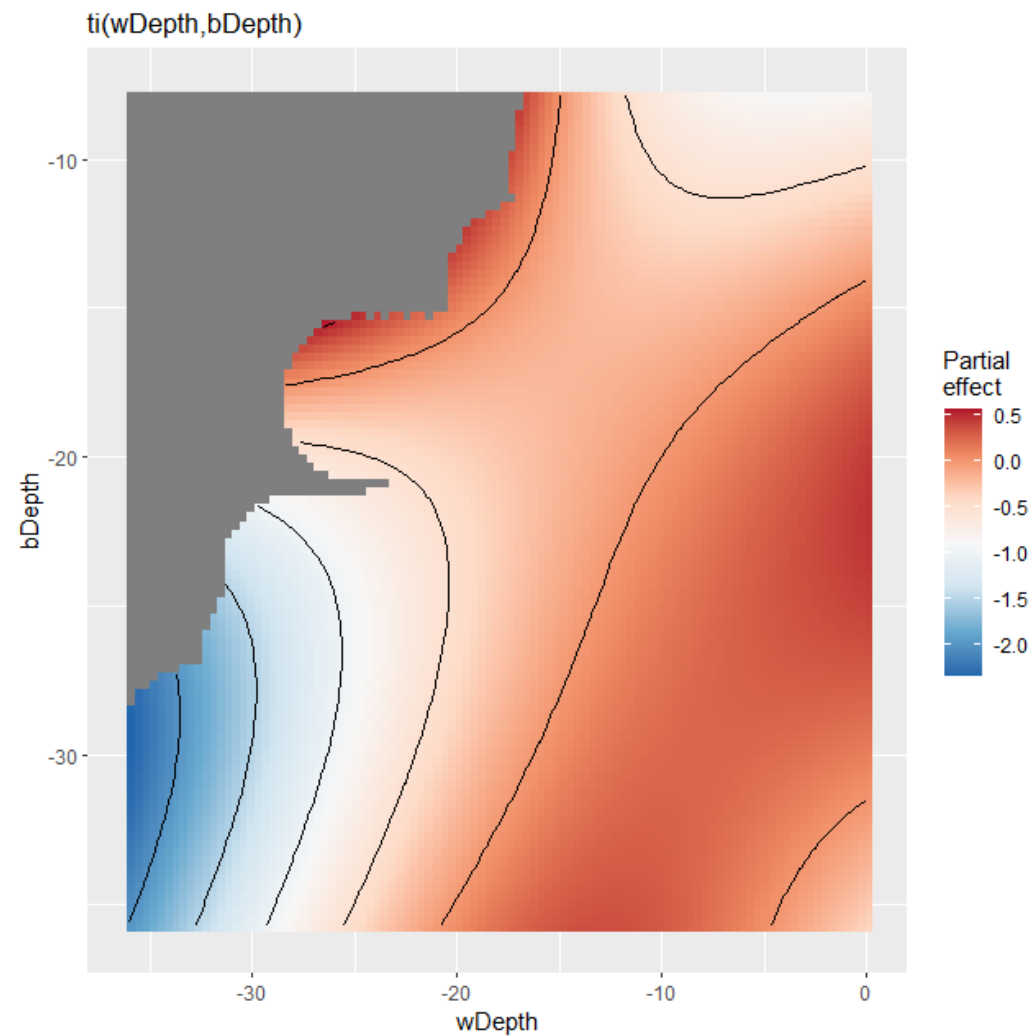


Reduced k change in $ti(wDepth, bDepth)$ graphic:

Original model with default k's



New model with reduced k's



Investigating k-values in GAMs

- K-values control the upper limit of flexibility in the GAM splines. These are set to a default in our currently tested model. All of the documentation stresses that this parameter needs to be evaluated.
- I had been concerned we may be “underfitting” the models in certain times & places and increasing the k-values might be needed.
- Gam.check is a tool for this:
 - `gam.check()` is really just testing for local bias in the prediction surface.
 - indicates we may be setting our upper limit of flexibility, or “k-values” too low in some cases
- Other tools: graphics of the splines AND hold-out analysis

```
> gam.check(gs7)
```

```
Method: GCV Optimizer: magic
Smoothing parameter selection converged after 33 iterations.
The RMS GCV score gradient at convergence was 4.546067e-07 .
The Hessian was not positive definite.
Model rank = 265 / 265
```

Basis dimension (k) checking results. Low p-value (k-index<1) may indicate that k is too low, especially if edf is close to k'.

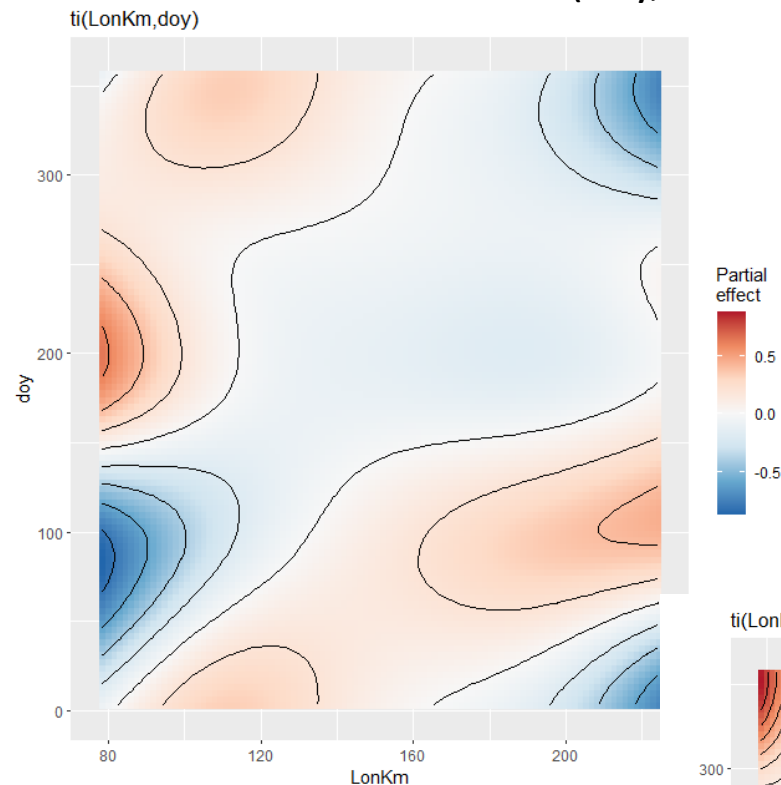
	k'	edf	k-index	p-value
s(cyear)	19.00	18.96	0.76	<2e-16 ***
s(doy)	8.00	7.97	0.91	<2e-16 ***
s(wDepth)	9.00	8.41	0.96	<2e-16 ***
s(LonKm)	9.00	8.45	0.98	0.100 .
s(LatKm)	9.00	1.00	0.98	0.075 .
s(bDepth)	9.00	9.00	0.98	0.080 .
ti(LatKm,wDepth,LonKm)	61.00	60.86	1.01	0.715
ti(wDepth,bDepth)	16.00	9.72	1.02	0.925
ti(wDepth,cyear)	16.00	12.42	0.94	<2e-16 ***
ti(LonKm,cyear)	16.00	14.07	0.84	<2e-16 ***
ti(LatKm,cyear)	16.00	13.57	0.83	<2e-16 ***
ti(bDepth,cyear)	16.00	14.19	0.90	<2e-16 ***
ti(wDepth,doy)	12.00	11.72	0.97	0.025 *
ti(LonKm,doy)	12.00	11.80	0.89	<2e-16 ***
ti(LatKm,doy)	12.00	11.89	0.88	<2e-16 ***
ti(bDepth,doy)	12.00	11.45	0.89	<2e-16 ***
ti(cyear,doy)	12.00	11.92	0.77	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

K's set to default of 5 each in ti(doy,LonKm)

Investigations

- Stepped through increasing k-value in splines and interactions and examined graphics to see if getting more explanation in certain areas
- Sometimes see graphics change, like this:
- Suspect these shapes might indicate over-fitting, check with a hold-out analysis



K's increased to 7 for each

