SC/67B/EM/03

No substantial change in Antarctic minke whale condition during the JARPA years

McKinlay, J. P., de la Mare, W. K., Welsh, A. H.

INTERNATIONAL
WHALING COMMISSION

No substantial change in Antarctic minke whale condition during the JARPA years

J.P. McKinlay[1], W.K. de la Mare[1] and A.H. Welsh[2].
1. Australian Antarctic Division, Channel Highway, Kingston, Tasmania, Australia 7050.
2. Mathematical Sciences Institute, the Australian National University, Canberra ACT, Australia 2601

Abstract

Minke body condition has been posited as relevant to the development of Antarctic ecosystem models. A number of previous analyses of data collected under the JARPA special permit whaling programme have focussed on whether there is evidence of a statistically significant linear decline in body condition. The focus here is broadened to examine the likely shape and variability in any time trends in body condition without restricting the analyses to linear trends and measures of statistical significance.

We analyse trends in Antarctic minke whale body condition using generalised additive models of body weight and blubber thickness (BT11). We take into account that data arising from each of the West and East sampling regions are confounded in space and time, that females (all pregnant) and males may require appreciably different model structures, and that some ambiguity exists concerning how long animals have spent in Antarctic waters feeding prior to capture. We also focus on the change in body condition that can be attributed to summer feeding in Antarctica.

Our results indicate very little trend or variability in the improvement in body condition attributable to the period of summer feeding sampled by JARPA, for males or females from the West or East sampling regions. As the patterns of change we detected are not consistent between regions or sexes, we conclude that past studies estimating a single trend for both sexes over the entire region are implausible. In particular, we show that longer animals have proportionately lower blubber thickness than shorter animals, but that this relationship is only revealed by including body weight as a covariate in models. We additionally identify a time-trend for increasing body lengths in the JARPA catches. Model misspecification by omitted body weight as a covariate in models of blubber thickness is likely to be interacting with increasing lengths in the JARPA catches to induce false signals for changing condition. By inference, this issue would apply equally to analyses of fat weight.

While our analyses have improved our understanding of the features of the data relevant to estimating body condition, we remain concerned about the veracity of any results arising from analyses of the JARPA dataset. This is due to the apparent large-scale year-to-year differences in key characteristics of the data, such as the proportion of low and high diatom animals in catches and the increase in average lengths over time. These differences could indicate that the segment of the total population being sampled has changed year-to-year, that subjective data recording protocols have been inconsistent, and that fleet operations have changed over time. We do not believe it is possible to fully resolve the influence of these potential sources of bias using the data available.

KEYWORDS: Antarctica, body condition, blubber thickness, minke whale, biased sampling, JARPA Special Permit Whaling, lurking variable.

No substantial change in Antarctic minke whale condition during the JARPA years

J.P. McKinlay, W.K. de la Mare[*] *and*
A.H. Welsh[†]

Contents

[*]Australian Antarctic Division
[†]Australian National University

.

1 Introduction

Two teams of analysts have been working towards improving our understanding of minke whale body condition, as reflected in the data collected under the JARPA Special Permit whaling program, 1989-2005. During IWC-SC67A, McKinlay, de la Mare and Welsh (hereafter MDW) presented models of body weight using additive models (McKinlay et al., 2017) and fat weight using linear mixed effects models (LMM) (de la Mare et al., 2017). At the same meeting, Cunen, Walloe and Hjort (hereafter CWH) examined several condition related variables and presented analyses that brought together several different techniques, including a jackknife-like technique to examine the importance of spatial effects, LMM for formal model development, and applied the focused information criterion (FIC) for model selection (Cunen et al., 2017). Conclusions drawn from the results presented by the two teams differed; CWH assert there has been a single, global decline in body condition during JARPA, while MDW contended that most of the data (83%) showed no decline in condition at all. Each analysis team raised queries about the assumptions, analysis decisions and statistical techniques employed by the other team.

In this work we address several queries or criticisms of our analysis raised during IWC-SC 67A, and introduce several extensions to our previous work. In summary, in this work we:

- develop additive models using blubber thickness (`BT11`) as a response, in order to compare results with those from models based on body weight (`BWt`). Note that in this work we focus exclusively on `BT11` as the blubber measurement thought to be most responsive to changes in condition.
- incorporate age as a covariate in models of body weight and blubber thickness.
- include body weight as a covariate in models where blubber thickness is the response. Body weight proves to be an important predictor of blubber thickness, and we compare these models with and without the addition of body weight as a covariate.
- consider separate models for the West and East regions of JARPA, as our preferred method of accommodating the fact that these time-areas are completely confounded with respect to space and time. We compare results with models that ignore this discontinuity entirely.
- establish that diatom score is largely unnecessary in models of body condition for females, provided that the variable fetus length is included in models.
- address the criticism that analyses presented in McKinlay et al. (2017) subset the data unnecessarily. In reviewing our previous decision to subset, we modify our subsetting strategy in light of further consideration of the sampling design. We maintain that separate analyses for each sex and each West-East sampling region are likely to provide the best inferences to address the question of changing body condition.

Throughout we refer to the total weight of animals (tonnes) as body weight (or simply weight), corresponding to the variable `BWt` in the JARPA data. Similarly, body length (m) (or simply length) of animals corresponds to the variable `BLm`, blubber thickness (cm) at a lateral point under the dorsal fin is `BT11`, year is `YearNum2` (1989-2005), fetus length (cm) is used synonomously with its variable name `FetusLength`, sampling day within each season (days since 30 November) corresponds to the variable `DateNum`, the longitude of catch positions is `LongNum`, and the categorical variable `Ice` denotes the proximity of catches to the ice edge (either `near` or `far`). Most other variable names are self-explanatory, but some require discussion concerning their interpretation (such as diatom score), and for those we defer definitions to the methods section.

We conclude by introducing an important caveat to the results presented in this work. The JARPA program was not designed to collect data for assessing changes in minke whale condition, and so the data are limited in several respects. We have previously discussed some of these limitations, including spatial and temporal confounding, biased sampling, and imbalance in relation to many of the important design characteristics (de la Mare, 2012, 2011; de la Mare et al., 2014; McKinlay et al., 2017; Wotherspoon et al., 2014). We have done our best to accommodate these deficiencies where possible (e.g. by analysing completely confounded data separately), but some we have had no choice but to ignore (e.g. imbalance and biased sampling). We therefore judge the results in this work to be the best we could do in the circumstances, but caution our audience to consider our previous work in relation to the limitations of the program when assessing the results presented here.

2 Methods

The statistical methods we employ are again grounded in the additive model methods of Wood (2017, 2011), as described in McKinlay et al. (2017). However, in response to feedback received during SC67A, and in light of further investigation, we modify and expand our methods to include:

i) our rationale for excluding 14.6% of cases that are missing data on variables considered essential for analysis;
ii) a more detailed explanation of why we advocate separate analyses for some partitions of the data;
iii) consideration of whale age in models of body weight and blubber thickness;
iv) assessment of the importance of spatial terms in models of condition;
v) modelling blubber thickness (`BT11`) as a response, for the purposes of comparison with models of body weight.

The remainder of this section discusses several of these issues in more detail, and is concluded by an outline of the structure of the results section.

2.1 Data preprocessing

The original data contain 4781 cases, comprising 1814 female and 2967 male samples. However, not all the original data are suitable for analysis.

There are 40 cases for females that are missing `FetusLength`. This proves an important covariate in explaining female condition, so these records are dropped from the analysis. This resulted in no more than 6 cases dropped in any one year, and reduced the female sample size across all years from 1814 to 1774.

A further four female cases were dropped due to outlying values on fetus length (Figure 1). These values are unusual in that they indicate close to full-term fetus lengths from mothers sampled near the beginning of the Austral summer (Dec/Jan). Additionally, these mothers had been present in Antarctic waters for some time (i.e. had diatom score 3 or 4, defined later in this section). We believe these large values are likely legitimate measurements, but that they arise from female animals that fell pregnant relatively late in the previous season, arrived on the feeding grounds late, and so over-wintered near the ice. This hypothesis is consistent with the data and the known gestation period for minke whales. These samples are clearly atypical compared with all other females in the data, indicating that this type of behaviour (if we are correct in our assessment) is reflected only rarely in the JARPA data. Dropping these four records reduces the number of female cases to 1770.

A further 653 cases were dropped from analyses due to missing values on `BT11` (15 cases), body weight (192 cases), diatom score (5 cases) or age (441 cases), all considered critical as either response or covariate. We believe that omitting cases due to missing values on `BT11` or diatom score should be uncontroversial, but that omissions due to body weight or age need further explanation.

In this work we examine changes in condition by developing models for two different response variables, body weight and blubber thickness (`BT11` only). Models of blubber thickness account for the deposition of subcutaneous fat, while models of body weight additionally account for weight gain in muscle, bone and viscera, including the deposition of fat in those tissues. Some members of the Working Group on Ecosystem Modelling argued during SC67A that models of `BT11` should be preferred over models of body weight since the latter is not a reliable indicator of condition. However, we remain uncertain about this issue since the literature is equivocal on this point, if not entirely lacking. To our knowledge, there are no studies of baleen whales with sufficient sample sizes taken across a feeding season, and with concurrent measurements of body weight and blubber thickness, to allow any kind of meaningful comparative study. We would welcome pointers to these types of studies, if they exist. It seems to us that most studies of condition in rorquals utilise blubber thickness as the primary measure of condition because obtaining body weights of such large animals is usually impracticable. Since the JARPA data contain body weights, we develop models for both body weight and blubber thickness. We avoid any argument about which response might be the better indicator of condition by presenting results from models of each response. Interestingly, in our best models the covariates available

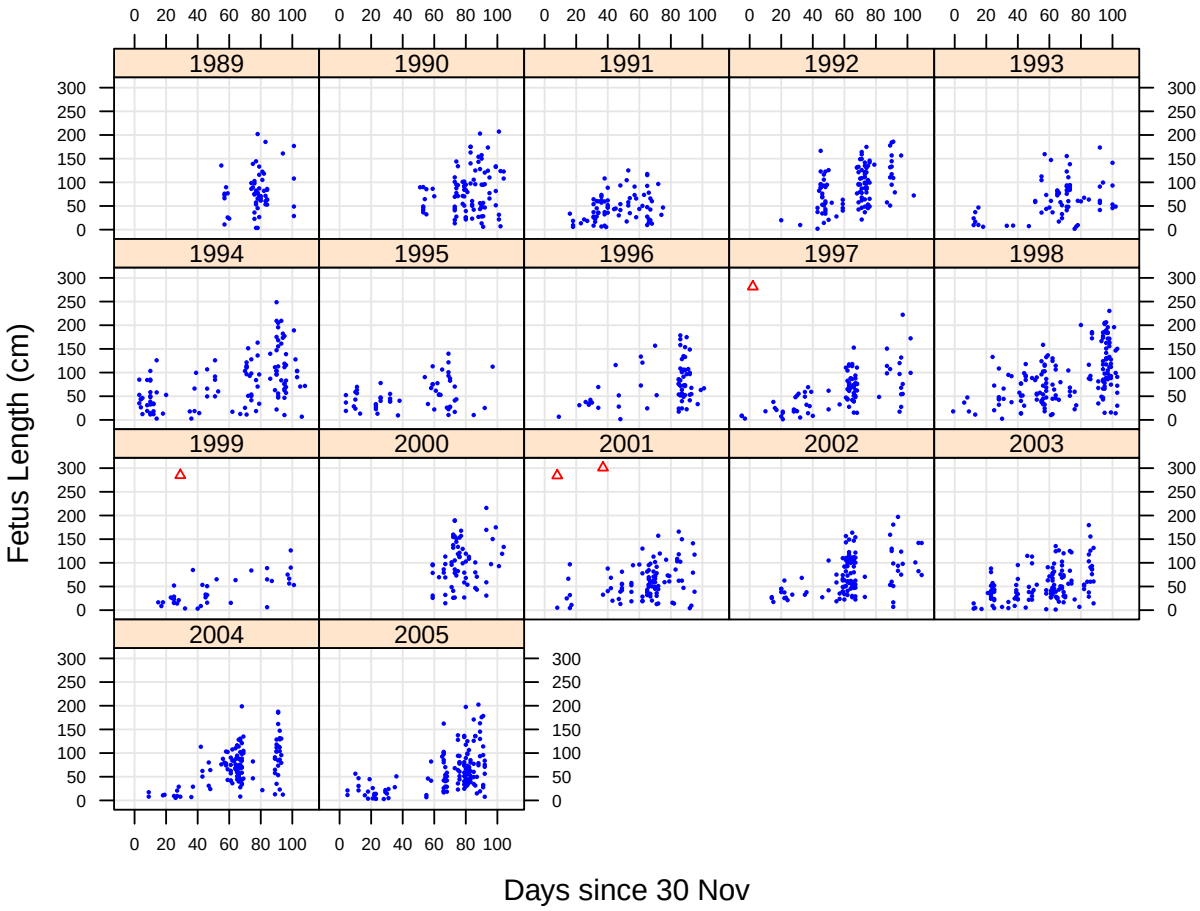

Figure 1: Scatterplot of fetus lengths according to the day within season (DateNum) on which the mother was sampled, separately for each year (panels). Four unusually large fetus lengths are evident in the early part of the season (triangles).

to this study typically account for 60-70% of the variance in body weight, but only 40-50% of the variance in BT11.

We included age as a covariate in models of body weight on the basis of advice received from members of the Working Group on Ecosystem Modelling during SC67A, and age indeed proved to be an important predictor of body weight. We also considered age and body weight as covariates in models of blubber thickness, with both proving to be important predictors. We therefore find age and body weight necessary for analyses of BT11, and so enforce the requirement that data be complete on these variables.

During SC67A some members expressed concern over our decision to subtract stomach weight from body weight in order to remove extraneous variation that might mask or otherwise dilute any signal in body condition. The main issue raised concerned the reduction in sample sizes this adjustment would cause, since many cases were missing stomach weight. These analyses do not include any correction for stomach weight, and weights used reflect the body weight of animals including stomach contents.

In summary, a total of 4084 cases, 1560 female and 2524 male, remain after subtracting those cases that are incomplete on important variables. These remaining data comprise 85.4% of the animals sampled under JARPA.

2.2 Sample Sizes

It is relevant to consider how the 4084 cases available for analysis are distributed across important design variables, since their distribution has informed some of our analysis decisions. Recall that sampling under JARPA alternates between broad geographic regions each year; 'even' years are sampled in the Western region (40-130°E), while samples from 'odd' years arise from an Eastern region (130°E-145°W) (Table 1). We hereafter refer to these regions simply as West and East. Each region describes vast tracts of ocean, with West and East accounting for around 4200 km and 3800 km around the Antarctica continent, respectively[1]. This facet of the JARPA sampling design — namely, that samples from alternate years are geographically and temporally distinct — is an import aspect of the program that needs to be considered in analyses, a point we return to shortly.

Table 1: Sample sizes by year in each West-East sampling region.

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Sum
West	0	208	0	170	0	190	0	274	0	216	0	259	0	269	0	260	0	1846
East	135	0	242	0	249	0	210	0	299	0	235	0	289	0	283	0	296	2238

It is also informative to consider the number of samples available by sex and diatom load within each region, West and East (Tables 2 and 3). We argue in the next section that, in order to obtain reliable inferences about condition, separate analyses should be conducted on subsets of the data split by sex and region. For now, we simply make the point that the per-year sample sizes for these cross-classifications of the data are small, often only 20-50 cases per year, and that this provides a challenging starting point for disentangling the within-season improvement in body condition from any long-term trend.

Table 2: Sample sizes by sex (F, M) and diatom load (Low, High) for each year in the West sampling region.

	1990	1992	1994	1996	1998	2000	2002	2004	Sum
Low.F	57	34	20	34	23	35	61	51	315
High.F	12	22	31	47	9	60	62	80	323
Low.M	58	55	32	55	67	54	56	43	420
High.M	81	59	107	138	117	110	90	86	788

[1]Measured as the approximate Greater Circle Distance around the Antarctic continent between the appropriate longitudinal boundaries at 65°S.

Table 3: Sample sizes by sex (F, M) and diatom load (Low, High) for each year in the East sampling region, including the Ross Sea (Stratum 'VES').

	1989	1991	1993	1995	1997	1999	2001	2003	2005	Sum
Low.F	50	45	33	29	48	11	36	62	45	359
High.F	30	57	73	30	93	46	69	48	117	563
Low.M	21	44	29	45	61	43	56	84	45	428
High.M	34	96	114	106	97	135	128	89	89	888

Finally, we examine the distribution of samples by stratum[2] and diatom load for each year of the sampling program, separately for male and female animals (Tables 4 and 5). These cross-tabulations reveal several features that should be considered when developing models:

- per-year sample sizes within individual strata are particularly low (i.e. < 10 animals in many years), with likely consequences for the ability of models to detect spatial differences in condition.
- the Western-most (IIIE) and Eastern-most (VIW) strata were only sampled for half the number of years compared with other strata.
- not every stratum is sampled according to the usual biennial sampling pattern; several strata are sometimes only sampled every four years.
- for males (Table 4), a majority of animals sampled were of high diatom load.
- for females (Table 5), for several years in several strata a majority of animals sampled had low diatom load.
- in the first year of the program to includes all the covariates, 1989, all catches arose from strata VEN or VES.
- samples are overall skewed towards male animals (62%), the notable exception being the Ross Sea (VES) where females dominate catches. In fact, 32% of all females landed were taken from the single stratum representing the Ross Sea.
- although not readily apparent from these tables, exploratory work has previously shown that catches from individual strata arise from short, isolated periods of time within a season (Wotherspoon et al., 2014, Figure 3).

2.3 Separate analyses by subsetting

During SC67A, analyses presented in McKinlay et al. (2017) were criticised on the basis that data were split by sex and diatom load (as a 2-level factor, low and high), with separate analyses conducted on these subsets. Opposition centred on the idea that conducting separate analyses would reduce the sample size available to each analysis, and that this would reduce overall power to detect an effect should one exist. Our approach contrasted to that of other analysts that considered the entire JARPA data set in a single analysis (Cunen, 2017; Konishi and Walløe, 2015). We have considered the criticism of our subsetting approach seriously, and acknowledge that with better data, and perhaps with different questions of interest, a single model would also be our preferred position. However, on balance we believe that for questions on the nature and extent of variability in body condition to inform the development of ecological models, more robust inferences can be obtained from analyses of coherent data subsets. We also find that there are some aspects of the JARPA design that can only be accommodated by subsetting. After further analysis, we have refined the criteria for defining subsets compared with our previous work. In the following sections we recap and extend the rationale for subsetting put forward in McKinlay et al. (2017), and detail the criteria we adopt in the current work.

The criticism that analyses of subsets will reduce the statistical power to detect an effect is strictly correct in a setting of Neyman-Pearson hypothesis testing. However, rather than a p-value attached to a linear trend,

[2]The JARPA data are geographically stratified to account for differences in effort from the planned tracklines. The strata divide the traditional IWC management Areas in half (West-East, not to be confused with West and East sampling regions) and each of these into North-South strata. An additional stratum was included for Prydz Bay. The Ross Sea is stratum VES. A post-survey stratification for whales taken closer to the ice-edge is also used in some analyses.

Table 4: Sample sizes for males in each Stratum for each year, differentiating low and high diatom score animals.

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	**Sum**
West																		
IIIE.Low	.	.	.	.	.	.	.	15	.	18	.	14	.	15	.	18	.	**80**
IIIE.High	.	.	.	.	.	.	.	26	.	36	.	23	.	28	.	30	.	**143**
PB.Low	.	3	.	3	.	1	.	1	.	6	.	2	.	.	.	4	.	**20**
PB.High	.	1	.	6	.	6	.	17	.	14	.	3	.	4	.	9	.	**60**
IVWN.Low	.	37	.	24	.	5	.	28	.	9	.	7	.	13	.	8	.	**131**
IVWN.High	.	41	.	17	.	12	.	68	.	27	.	18	.	16	.	11	.	**210**
IVWS.Low	.	4	.	14	.	7	.	3	.	11	.	1	.	11	.	2	.	**53**
IVWS.High	.	18	.	19	.	35	.	9	.	16	.	11	.	12	.	20	.	**140**
IVEN.Low	.	12	.	4	.	18	.	6	.	14	.	8	.	11	.	6	.	**79**
IVEN.High	.	11	.	1	.	40	.	9	.	13	.	18	.	7	.	5	.	**104**
IVES.Low	.	2	.	10	.	1	.	2	.	9	.	22	.	6	.	5	.	**57**
IVES.High	.	10	.	16	.	14	.	9	.	11	.	37	.	23	.	11	.	**131**
East																		
VWN.Low	.	.	31	.	10	.	24	.	10	.	3	.	8	.	13	.	.	**99**
VWN.High	.	.	51	.	24	.	68	.	17	.	4	.	15	.	15	.	3	**197**
VWS.Low	.	.	4	.	8	.	.	.	2	.	17	.	18	.	6	.	6	**61**
VWS.High	.	.	9	.	59	.	.	.	9	.	50	.	27	.	11	.	24	**189**
VEN.Low	13	.	9	.	6	.	14	.	15	.	17	.	18	.	31	.	11	**134**
VEN.High	15	.	29	.	8	.	24	.	29	.	62	.	34	.	35	.	12	**248**
VES.Low	8	.	.	.	5	.	7	.	7	.	.	.	1	.	11	.	7	**46**
VES.High	19	.	7	.	23	.	14	.	15	.	.	.	21	.	8	.	23	**130**
VIW.Low	.	.	.	.	.	.	.	.	27	.	6	.	11	.	23	.	21	**88**
VIW.High	.	.	.	.	.	.	.	.	27	.	19	.	31	.	20	.	27	**124**
Sum	55	139	140	114	143	139	151	193	158	184	178	164	184	146	173	129	134	**2524**

our work is concerned with obtaining estimates — linear or non-linear, as the data dictate — of any long-term trend in condition, along with estimates of precision for those estimates. We feel that exploring the *a priori* assumption of linearity adopted by other analyses is an important step in our work, since imposing linear relationships from the outset can sometimes result in quite perverse interpretations of data. For instance, it is well known that linear relationships can be adversely affected by high leverage, influential points near the extremes of the predictor space. In the case of the JARPA data, just one or two "good" feeding years towards the beginning of the JARPA sampling period, or one or two "bad" years towards the end, could lead us to erroneously conclude that there has been a significant negative trend in condition, when for most years it was simply business as usual. Our preferred approach is therefore to obtain our best estimates of the signal in the data using methods that can accommodate influential points without unduly imposing trends where arguably none exist. Our emphasis on estimating the precision of (potentially non-linear) relationships, instead of the significance of a linear trend, changes the importance of sample sizes in analyses. At best, using all the data in a single model may improve estimates of precision, but we argue below that this comes at the cost of muddied inferences.

2.3.1 Split on West and East regions

We first consider the biggest issue, the West-East split in the JARPA sampling scheme. Recall from Section 2.2 that the West and East regions do not overlap, having centres separated by several thousand kilometers, and are each sampled in alternate years. Recent models, by both MDW and CWH, have included data from both regions concurrently, giving no particular emphasis to the fact that data from West and East are each completely separated in time and space. For convenience we categorise this split using the term *region*, but this is not simply a spatial split, it is a *time-space* split. This means that these data are completely confounded with respect to time and space. The fact that the West-East split has received little attention

Table 5: Sample sizes for females in each Stratum for each year, differentiating low and high diatom score animals.

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	**Sum**
West																		
IIIE.Low	.	.	.	.	.	.	.	12	.	13	.	5	.	19	.	14	.	**63**
IIIE.High	.	.	.	.	.	.	.	10	.	.	.	2	.	6	.	2	.	**20**
PB.Low	.	11	.	6	.	4	.	10	.	1	.	1	.	2	.	2	.	**37**
PB.High	.	2	.	9	.	8	.	26	.	4	.	5	.	13	.	7	.	**74**
IVWN.Low	.	10	.	4	.	1	.	10	.	5	.	5	.	13	.	6	.	**54**
IVWN.High	.	.	.	3	.	1	.	6	.	2	.	2	.	3	.	1	.	**18**
IVWS.Low	.	7	.	11	.	3	.	1	.	2	.	.	.	1	.	17	.	**42**
IVWS.High	.	4	.	5	.	8	.	1	.	1	.	10	.	3	.	53	.	**85**
IVEN.Low	.	8	.	2	.	9	.	.	.	1	.	1	.	4	.	2	.	**27**
IVEN.High	.	1	.	1	.	8	.	.	.	1	.	3	.	2	.	.	.	**16**
IVES.Low	.	21	.	11	.	3	.	1	.	1	.	23	.	22	.	10	.	**92**
IVES.High	.	5	.	4	.	6	.	4	.	1	.	38	.	35	.	17	.	**110**
East																		
VWN.Low	.	.	18	.	11	.	7	.	6	.	.	.	3	.	4	.	.	**49**
VWN.High	.	.	11	.	6	.	2	.	10	.	1	.	.	.	.	.	1	**31**
VWS.Low	.	.	6	.	11	.	2	.	10	.	6	.	4	.	7	.	3	**49**
VWS.High	.	.	12	.	30	.	1	.	19	.	36	.	15	.	18	.	33	**164**
VEN.Low	9	.	3	.	3	.	2	.	7	.	5	.	5	.	7	.	1	**42**
VEN.High	1	.	1	.	2	.	3	.	9	.	9	.	1	.	4	.	2	**32**
VES.Low	41	.	18	.	8	.	18	.	16	.	.	.	17	.	35	.	28	**181**
VES.High	29	.	33	.	35	.	24	.	49	.	.	.	48	.	26	.	79	**323**
VIW.Low	.	.	.	.	.	.	.	.	9	.	.	.	7	.	9	.	13	**38**
VIW.High	.	.	.	.	.	.	.	.	6	.	.	.	5	.	.	.	2	**13**
Sum	80	69	102	56	106	51	59	81	141	32	57	95	105	123	110	131	162	**1560**

in previously analyses of the JARPA data speaks to the difficulty of dealing with what is one of the most notable features of the sampling design.

For anyone who doubts the importance of the West-East split we provide just one example of the potential to be misled by ignoring this characteristic of the design. Consider fetus lengths for high diatom load (scores 2, 3 and 4 combined) female animals sampled in February each year (Figure 2). For both West and East combined there appears to be an appreciable decline in fetus length with time (Figure 2a), however the 'saw-tooth' pattern in the averages year-to-year (red line) provides grounds for concern. Examining this apparent trend for West and East separately (Figure 2b) reveals little-to-no trend in the West, but perhaps a notable trend in the East[3]. Importantly, the year-to-year oscillation in average values is absent for the disaggregated data.

We considered including in models a categorical variable defining the West-East split but decided against this strategy because the interpretation of this variable would remain elusive. Would a significant effect indicate spatial differences, temporal differences, or both? Or neither, since space and time could each be unimportant but, when considered concurrently (as a confounded term) give the impression of an effect. Similarly, a non-significant result could not be trusted for the same reasons; space and time are exactly confounded, and so important spatial and temporal effects may 'cancel one another out' and appear unimportant. Separate analyses of each region naturally segregates this time-space confounding issue, so that is the approach we have adopted here.

[3]Before anyone gets too excited that declining fetus size may be an indicator of malnutrition, we would point out that this may be an indication of a change in the temporal migration pattern of females, an indication of changing conditions in the Ross Sea (where a majority of the female samples are taken), or simply an artifact of the spatio-temporal variation in data collection under the JARPA design.

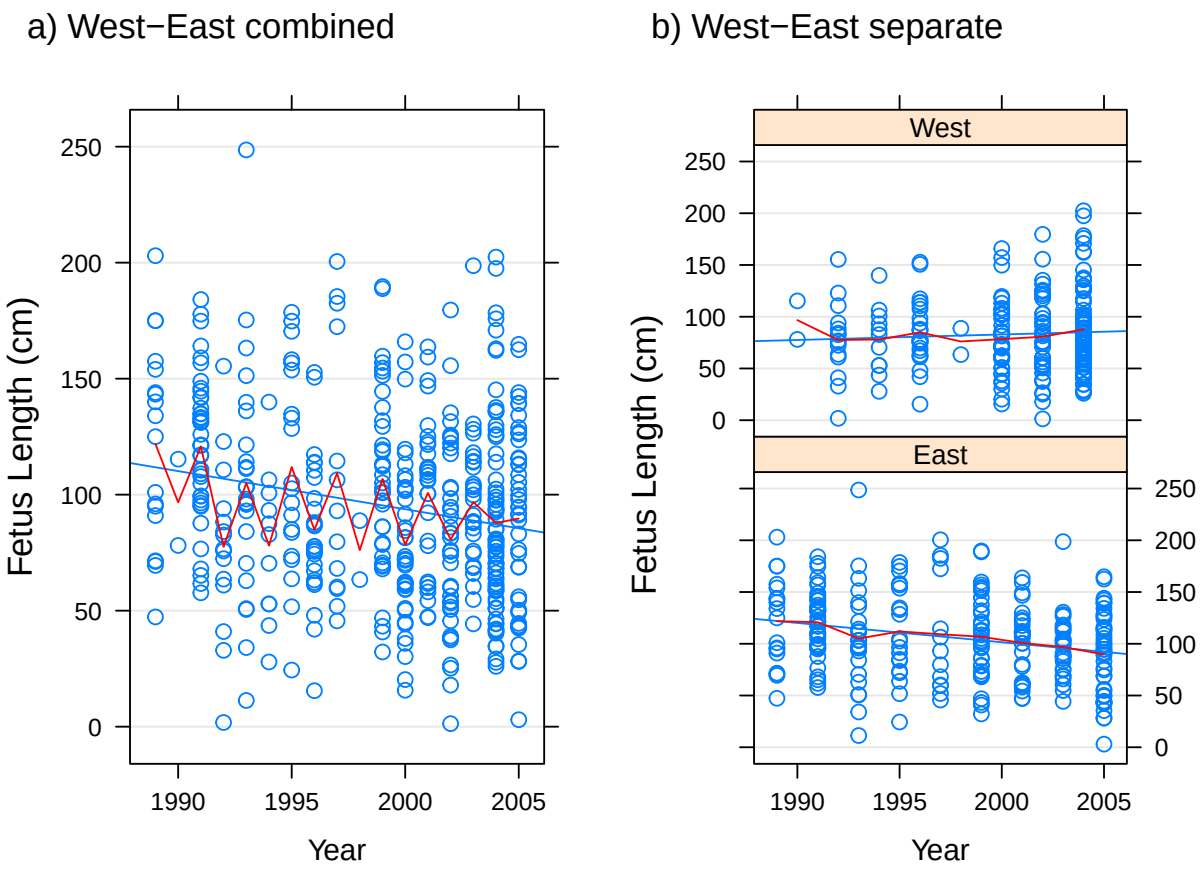

Figure 2: Plot of fetus length against year taken for high diatom females (scores 2, 3 and 4) captured in February each year, showing lines indicating a linear fit (blue) and joining the averages from each time point (red) for: a) all data from West and East combined; and, b) separately for West and East.

2.3.2 Split on sex

We decided to conduct separate analyses for each sex in McKinlay et al. (2017) because virtually all female animals captured during JARPA are supporting fetuses at various stages of development. All other things being equal, one might expect the spatial distribution, behaviour, metabolic demands and effects on condition to be appreciably different between female and male animals. These differences could be accommodated in a single (quite complicated) model if one were to include all interactions between sex and other terms. However, this makes for complicated models — complicated to fit and complicated to interpret — and we note that neither analysis team has so far attempted to fit model parameterisations that are fully separable on sex. If important higher-order interactions between sex and other covariates cannot be accommodated due to insufficient sample sizes or difficulties associated with model complexity, then the potential exists for lower-order effects to be averaged over the sexes and estimation will, in turn, be impacted by imbalance in the sample sizes for males and females. We believe this has the potential to provide misleading results.

For a covariate as important as sex, we contend that its effect should be fully separable in models. If this is difficult to achieve in a single model, then separate models achieve this goal in a transparent and defensible way. We do not try to "borrow weight" from male samples to make inferences about females, or vice versa. So while in principle we agree that in many contexts a single analysis is to be preferred over several separate analyses, this is not a universal truth and in this instance we do not consider that position to be supported by the data, or the goals of the analysis. Furthermore, if the effects of sex were being adequately captured in a single model approach (Cunen et al., 2017; Konishi et al., 2014; Konishi and Walløe, 2015), then at a minimum the same approximate mean structure should be obtained from separate models for each sex. Different results from each approach may indicate that the single model has not adequately captured all sex-related effects, and that an inappropriate averaging over the sexes may be occurring. Thus, analyses of each sex separately can serve as a check that results presented from a single, unified model have fully captured all necessary interactions involving sex.

While we find the arguments above for splitting on sex compelling — they formed the basis for our previous decision to analyse the sexes separately — we now identify another consideration. In the LMM models employed by both analysis teams, the variable fetus length has been incorporated into models containing both sexes by setting fetus lengths to zero for males, and encoding the factor representing males and females as a binary indicator with males = 0. This modelling "trick" allows both sexes to be accommodated into a single model, even though males clearly do not have fetuses. However, in the present work we have discovered that, for females, diatom score and fetus length are correlated and effectively compete with one another to explain the same deviance (see Appendix A). The sequential addition to models (for females) of terms for fetus length and diatom score revealed that fetus length accounts for substantially more deviance than diatom score, so much so that including diatom score was unnecessary in almost all models that already contained fetus length (i.e. the marginal improvement of adding diatom score was negligible for models already containing fetus length). We believe this occurs because both variables provide a proxy measure of time spent in Antarctic waters.

Interpreting fetus length in this way makes intuitive sense assuming that female animals begin their migration to summer feeding grounds after conception. Provided this condition is met, fetus length should provide a quantitative, relative measure of time spent in Antarctica free of the apparent subjectivity associated with classifying whales into different diatom score categories. While there will undoubtedly be some variation around conception and migration times, and this variation is unknown, the models of the JARPA data themselves reveal that fetus length performs far better than diatom score as a predictor of `BWt` and `BT11`, and that the marginal improvement of incorporating diatom score is negligible. For female models, diatom score is shown to be unimportant when including a term for fetus length in all cases but one. Consequently, diatom score is largely unnecessary for female models, and fetus length is inapplicable for male models. Therefore it would be strange and unconventional to force both sexes into a single model.

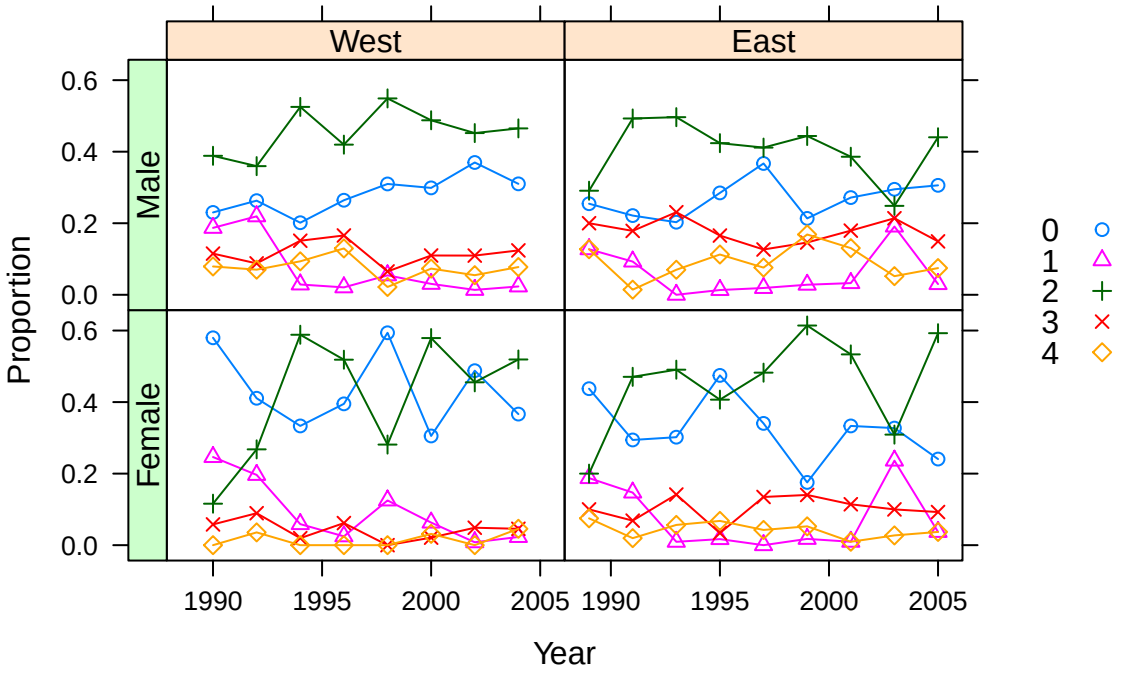

Figure 3: Proportion of samples in each diatom class (0-4) in each year, separately for each sampling region (West, East) and sex.

2.3.3 Reconsidering diatom score

We now consider our treatment of diatom score in analyses. Diatom score is reported in the data as a 5-level ordered categorical variable with levels 0 (little to no diatom coverage) to 4 (high diatom coverage). The variable has been treated as a continuous variable in some analyses (Cunen et al., 2017; Konishi and Walløe, 2015), however we have argued that this is an inappropriate use of the variable in models (McKinlay et al., 2017). This variable should ideally be considered in models as an ordered categorical variable, however this is also problematic due to the small sample sizes available for three of the five categories.

Examining diatom score data reveal there are clear differences in the proportion of animals categorised into different diatom score classes (Figure 3). For males, in any given year typically 40-50% of animals are scored as category 2, around 20-30% of animals are scored as 0, with other categories usually accounting for less — often much less — than 15% of the total. This pattern is similar for females but with some pronounced switches in the relative frequency of score 0 and 2 animals reported in some years. Clearly, categories 0 and 2 are reported far more frequently than adjacent categories for both males and female in both regions.

Several marked departures from these general patterns are worth mentioning, since these will potentially be impacting analysis results. In West and East regions for females, the first two years of sampling show that reporting of the relative frequency of animals with scores 1 and 2 are at odds with much of the remainder of the series (Figure 3). Similarly, years 1998 and 2003 are also unusual for females from the West and East, respectively. The pattern in relative frequency for males seems more consistent across years, however — like for females — anomalies are apparent in the first two years in the West; diatom score 2 starts at around 20% then drops to negligible levels. Also, the anomaly present for females in 2003 concerning the relative proportions of diatom class 1 and 2 is almost identically mirrored for males; this gives every appearance of a change in the recording practices for a single year.

The reasons why some categories (0, 2) are much more frequently reported than others (1, 3 and 4) are

not clear. If animals are smoothly transitioning from a low to a high diatom load as time spent feeding accumulates over the Austral summer then the assumption of continuity between the categories might be justified. In that case, we would expect more consistent frequencies in each category than are recorded. However, if the accumulation of diatoms is episodic then the transition from one score to a much higher score could occur in a relatively short time, in which case the assumption of linear continuity is untenable. Another possibility for the predominance of categories 0 and 2 in the data is that it is relatively easy for observers to judge diatom load as "not much" (0) or "quite a bit" (2), but assigning intermediate categories is more difficult and so is done less frequently[4]. This type of cognitive bias is well understood in psychometric circles and is typically explained in terms of the heuristics people employ when faced with uncertain or difficult decisions (Kahneman and Frederick, 2005; Shah and Oppenheimer, 2008).

We note that the analyses presented in Cunen et al. (2017) ignored the bimodal distribution of diatom scores. In fact, the primary work of CWH, and the work of Konishi and Walloe (2015) before them, treat diatom score as a continuous covariate, an approach with which we disagree. Clearly, if the data fall predominantly into two classes, then this structure should be included in models of condition. Furthermore, if there is a complete switch in some years in the relative proportion of landings with different diatom loads, then the effect of this switching behaviour needs to be carefully considered in models. At minimum, we believe models containing all data simultaneously would require an interaction between diatom score and year that was capable of capturing this switching in diatom classes (i.e. an interaction between diatom class and a linear trend in year would be inadequate; a more complicated form of interaction would be required). However, the low numbers of animals in diatom classes 1, 3 and 4 in most years would prove problematic for estimation.

In McKinlay et al. (2017) we therefore decided to consider separate analyses for low (scores 0,1 combined) and high (scores 2,3,4 combined) diatom load animals, reasoning that these two groupings accommodated the categorical nature of the data and provided readily interpretable groupings (discussed shortly). Plots of the relative proportion of the combined diatom classes serves to highlight some of the anomalies we have previously discussed (Figure 4). For example, the proportions of low and high diatom score males across year shows striking differences towards the beginning of the series in the West, and towards the end of the series in the East. Diatom scores for females show less consistent patterns, but still with several usual years; 80% of female animals landed in the West in 1990 were relatively newly arrived on the feeding grounds, for example. It is not clear to us whether this switching in the relative proportion of low and high diatom score animals evident in the data is due to actual population changes on the feeding grounds, inconsistencies in the way data are recorded year-to-year, or a change in the targeting behaviour of the catching vessels. *This is clearly a key question, as this variable is crucial for the interpretation of the JARPA data in relation to body condition, at least for the male samples (i.e. males do not have fetus length as an alternate indicator of time spent in Antarctica).* If these changes are truly reflective of the underlying population, then it appears to us that the JARPA program must effectively be sampling different components of the population year-to-year. This is obviously problematic for interpreting long-term changes in population characteristics. However, if these changes are due to biases, either due to changes in data recording or fleet targeting practices, then this raises different but equally challenging problems for interpreting the data.

For these analyses we modify our treatment of diatom score in models compared with our previous work. Here we prefer to use fetus length for females as the best predictor of time spent in Antarctic waters and so minimise our use of diatom score for female data (see discussion in Section 2.3.2 and Appendix A). For males, we are left with no choice but to use the 2-level categorical version of diatom score as a proxy for time spent feeding over summer. We denote this dichotomised version of diatom score as `DiatomF2` in models to indicate a factor variable with two levels, `Low` and `High`. This will necessarily reduce the predictive ability of diatom score in explaining changes in `BT11` or body weight, but we feel our decision to pool low-count categories is likely to be better than the alternative, which would be to discard data with diatom scores 1, 3 and 4. In contrast with our previous work in which we analysed diatom classes separately, in this work we fit both diatom classes (Low, High) concurrently in models for male animals.

[4]We are open to hearing argument that this apparent bimodal distribution of scored diatom load is actually representative of the sampled population rather than, say, due to observer bias, though we think that a "two waves of feeding migration" hypothesis is rather unlikely given that data indicate that diatom score 0 animals arrive on the feeding grounds throughout the summer season (McKinlay et al., 2017).

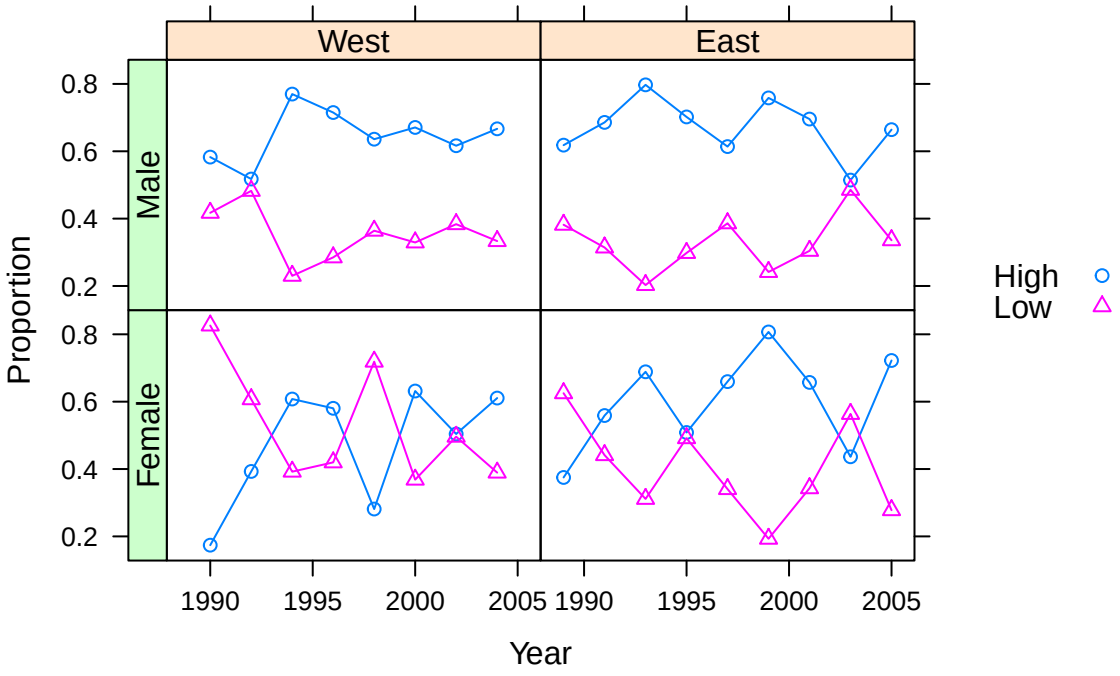

Figure 4: Proportion of landings in low (0,1) and high (2,3,4) diatom load classs in each year, separately for each sampling region (West, East) and sex.

It is also worth briefly considering when during a season animals with particular diatom scores are captured, since the data seem at odds with our intuition about when certain diatom classes should be more or less prevalent during a season. To this end, we plot sampling day (`DateNum`) against `BT11` for females and males separately, stratified by region (Figure 5a,b). These results show some patterns that are difficult to explain given our current knowledge of minke whale ecology. We note several features:

- diatom score 0 animals, for both males and females, are present throughout the within-season JARPA sampling period. These results indicate that males and females may arrive at a constant rate throughout the 100+ day the sampling season.
- for males, all other diatom classes (1,2,3,4) are also present throughout the sampling season, however for females it appears that higher diatom class animals are more likely to be present later in the sampling season.
- the propensity for animals to be scored as either class 0 or class 2 is clear for both sexes, but this propensity is perhaps is slightly more pronounced in females. We note that some caution needs to be exercised in this interpretation because males are represented by appreciably larger sample sizes than females in the data.
- caution also needs to exercised in interpreting the relationship between diatom score and BT11 since no adjustment has been made for the weight and length of animals. Nonetheless, we are surprised that low diatom class animals (0,1) can have blubber thicknesses as large as or greater than animals with high diatom scores.

If diatom scores are to be believed as representing time spent in Antarctic waters, then animals that are new to the Antarctic feeding grounds are arriving throughout the JARPA sampling season. Additionally, animals that have been present in Antarctic waters for some time (diatom scores 2,3,4) are available from the very beginning of the sampling season, particularly so for males. The degree of overlap between the classes is clearly extensive. The result that animals of all diatom classes are available throughout the JARPA sampling

season[5] casts considerable doubt on the utility of `DateNum` in explaining condition, or on the interpretation of diatom score, or both.

Despite our difficulties in understanding how diatom score and seasonal availability of animals interact, for this work we have adopted what seems to be the generally held interpretation of diatom score. That is, we assume that low scores indicate animals that have newly arrived in Antarctica, and high scores indicate those animals that have been present in Antarctic waters for some time. Based on this presumption, we conclude that results for low and high diatom load animals have distinctly different interpretations. We consider that condition for low score animals (those newly arrived in Antarctica) would indicate either changes in feeding conditions away from the summer feeding grounds, or due to the previous Antarctic season through a carry-over effect, or both. It is relevant here to recall that minke whales need to feed outside of Antarctic waters to be energetically viable (de la Mare et al., 2017; Leaper and Lavigne, 2007; Lockyer, 1981). Condition for high diatom animals would more naturally indicate summer feeding conditions, but would be somewhat dependent on arrival condition. It follows that it is the year-trend in the within-season gain in condition that will best describe a change in summer feeding conditions. In other words, it is the *difference* in condition between animals newly arrived to the summer feeding grounds and those animals who have spent considerable time feeding that is of most interest, a point we return to when we detail our modelling approach in Section 2.4.

2.3.4 Ross Sea data

During SC67A some analysts expressed concern over the decision taken in McKinlay et al. (2017) to conduct sensitivity tests of models by removing cases for animals captured within the Ross Sea. In fact, two complete sets of analyses were presented in our previous work, one including data from the Ross sea, the other excluding that data. Our rationale for presenting separate analyses was that the Ross Sea ecosystem is known to be appreciably different from other areas in East Antarctica, with a much heavier reliance by baleen whales on silverfish and ice krill (*Euphausia crystallorophias*) than in off-shore areas where Antarctic krill (*E. superba*) dominates minke whale diet (Pinkerton and Bradford-Grieve, 2014; Tamura and Konishi, 2014). These differences have obvious implications for the hypothesis put forward by Konishi et al. (2014) suggesting that recovery of some baleen whale stocks in recent years has led to the depletion of krill stocks, which in turn has impacted the condition of minke whales throughout the period of JARPA. If baleen whales feeding in the Ross Sea are subject to different ecosystem dynamics, then it would seem reasonable to assess changes in minke condition separately for the Ross sea catches.

Analysing the Ross Sea data separately from other areas sampled under JARPA is still our preferred position, however this is problematic given the sample sizes available. We have advocated that separate analyses should be undertaken for males and females, and for the West and East sampling regions. In Table 3 we provided the per-year sample sizes available in the East region (which contains the Ross Sea), cross-classified by sex and diatom score (Low, High); we note that some of those sample sizes are already very low. The problem with considering the Ross Sea separately is readily apparent if data from the East region are further stratified to show catches taken outside and within the Ross Sea (Tables 6 and 7, respectively). These data show two distinctive features: more than half of all females animals sampled in the East are taken from the Ross Sea, and annual catches of male animals within the Ross Sea are dramatically lower than catches taken outside. These characteristics of the data have so far not been accommodated in analyses by either CWH or MDW. Furthermore, these features cannot be easily or reasonably captured in models due to the low sample sizes apparent in many years.

In our analyses of the East region we do use all the data, including those from the Ross Sea, and so results are underpinned by an assumption that the ecological mechanisms affecting minke whale condition are operating equally within the different ecosystems within and outside the Ross Sea. Sensitivity tests that drop (or analyse separately) Ross Sea samples could help to inform whether this was a reasonable assumption, but lower sample sizes would mean models would be unable to assess the same model dimension afforded by

[5]This seems to be the general case, noting that the absence of females in diatom class 3 and 4 from the beginning of the JARPA sampling season may simply be due to reduced sample sizes for female animals.

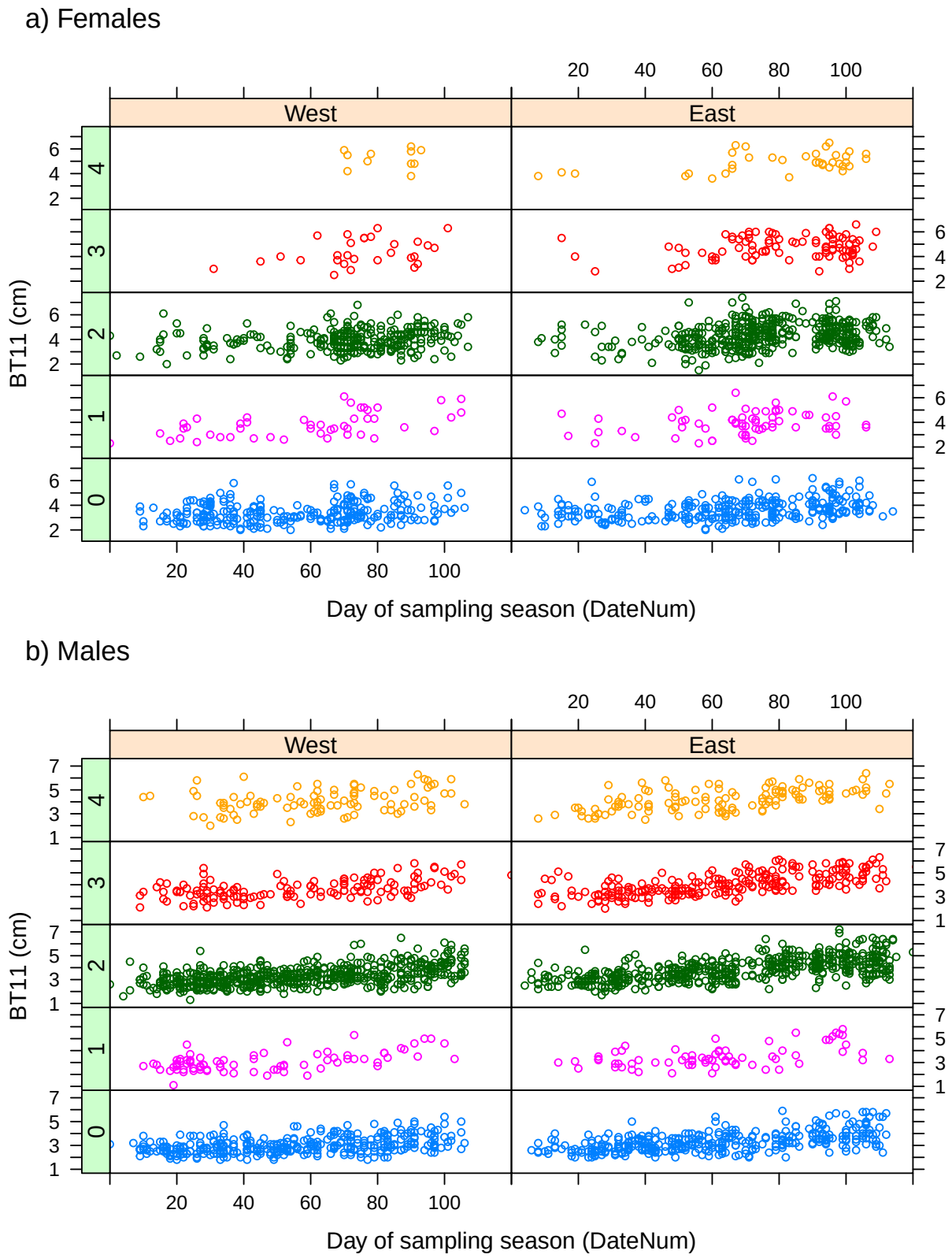

Figure 5: Plot of BT11 against date of capture (DateNum) for animals of different diatom classes (0-4, rows) in the East and West, all years combined, separately for a) female and b) male animals.

Table 6: Sample sizes by sex and diatom load (Low, High) for the East sampling region, by year, excluding the Ross Sea data.

	1989	1991	1993	1995	1997	1999	2001	2003	2005	Sum
Low.F	9	27	25	11	32	11	19	27	17	178
High.F	1	24	38	6	44	46	21	22	38	240
Low.M	13	44	24	38	54	43	55	73	38	382
High.M	15	89	91	92	82	135	107	81	66	758

Table 7: Sample sizes by sex and diatom load (Low, High), by year, for the Ross Sea (Stratum 'VES').

	1989	1991	1993	1995	1997	2001	2003	2005	Sum
Low.F	41	18	8	18	16	17	35	28	181
High.F	29	33	35	24	49	48	26	79	323
Low.M	8	0	5	7	7	1	11	7	46
High.M	19	7	23	14	15	21	8	23	130

analysing all data from the East together. To our minds, there are simply no good solutions beyond the obvious, which is to use a sampling design optimised for the questions being asked.

2.3.5 Summary: data subsetting

At face value the JARPA sampling design is similar to other designs for sightings surveys submitted for consideration by the Scientific Committee. However, there is a key difference in how the JARPA data are being used. Most sampling designs are concerned with determining an annual snap-shot of abundance, whereas the JARPA data, in this instance, are being used to try to disentangle the within-season accumulation of energy from any long-term trend in condition. For this purpose, the JARPA design is a particularly poor sampling scheme, with small realised sample sizes and both partial (within region) and total (between region) confounding between space and time.

In this work we present separate analyses for each sex (male, female) and region (West, East). For purposes of comparison, we also present results from a combined analysis of the West and East regions (although we do not recommend this analysis be used except for comparison). The cost of this approach is reduced sample sizes available for individual analyses, but the benefit is in obtaining more reliable inferences about any long-term trend in body condition. Sample size per analysis is of secondary concern to us because our goals are divorced from classical tests of linear trends. We do not see merit in conducting a pooled analysis to increase sample sizes, in order to allow more complex (yet still sub-optimal) linear models to be fitted, at the expense of obtaining potentially confounded results. We view the alternative we take — separate analyses on subsets — as less than satisfying, but as a sound approach in the circumstances.

2.4 Model development

We adopt the same additive modelling framework described in McKinlay et al. (2017), as implemented in the R package `mgcv` (Wood, 2017, 2011). It is worth briefly mentioning the longevity, stability and technical standing of this software; it has been a recommended package in R since 2001, it is supported by a second edition of Wood's acclaimed book (Wood, 2017), and underpins several other statistical estimation packages (e.g. the `dsm` package used in the design-based abundance estimation workshop conducted at SC-67a) (Hedley and Bravington, 2014; Miller et al., 2013). Concerning the current context, Marra and Wood (2011) discuss variable selection through shrinkage methods and demonstrate they perform better than competitors in terms of predictive ability, and are competitive in terms of variable selection performance. We note that Williams et al. (2013) also utilise `mgcv` for investigations into the ecological processes impacting blubber thickness in North Altlantic fin whales, while Augustin (2013) use penalised regression splines for a complex catch rate

standardisation that shares many similarities with the current analysis. While it is beyond the scope of the present work to provide more than passing commentary on the technical aspects of shrinkage smoothing as implemented in `mgcv`, it is important to understand why reduced rank smoothers have proved advantageous in modelling the JARPA data, and why we have made certain technical decisions. In summary:

- Generalised additive models, as implemented in `mgcv`, are simply (more complicated) generalised linear models (GLMs; McCullagh and Nelder, 1989), so many of the same principles apply. In our context, it proved sufficient to assume the response was Gaussian and errors were normally distributed with equal variance, for analyses of both body weight and blubber thickness.
- In additive models the response is additive in each term on the scale of the linear predictor, so traditionally assessing interactions between covariates in GAMs has proved difficult (Hastie and Tibshirani, 1990). The `mgcv` software offers two approaches for addressing interactions. The first fits full tensor-product smooths in two (or more) variables; these can be conceptualised as surfaces that simultaneously represent main effects and interactions between covariates in a single model term. An alternative, and our preferred approach, is to split full tensor products into marginal effects in each covariate, and a separate interaction-only component. In this way, we can explicitly test (through conventional methods, or through shrinkage approaches) for the requirement of an interaction over and above main effects. We have adopted this latter approach in our analyses.
- A key feature of additive models is that they are able to effectively capture non-linear behaviour. We thought at the outset that this would potentially be important for the JARPA data since exploratory work indicated that years near the beginning or end of certain series were sometimes higher or lower than a majority of other data (McKinlay et al., 2017). Such data are potentially influential since they occupy high leverage positions in the design space. Non-linear methods can capture these kinds of influential points without necessarily impacting the overall fit. In contrast, when faced with such influential points a linear technique may impose a linear trend on a series that are, for all practical purposes, stationary.
- Shrinkage methods for variable selection allow the data to speak for themselves, both in terms of the requirement for variables and the degree of non-linearity required in variables retained. There is also a considerable advantage in that selection doesn't follow a path-dependent route through the model space. Instead, we start with the maximum plausible model we are willing to consider, then let the technique reduce that complexity via a process where all variables are considered concurrently.
- Because GAMs under the `mgcv` framework have an underlying parametric form, it is possible to use the matrix which yields the values of the linear predictor when post-multiplied by the parameter vector (the so-called `lpmatrix` matrix in `mgcv`) to obtain non-standard quantities of interest. In our case, we use the `lpmatrix` along with simulation from the posterior distribution of the parameters to estimate mean and point credible intervals, and the difference between these intervals for whales captured early and late in the season in order to assess improvement in condition. Further, in these calculations we use a smoothing parameter uncertainty corrected covariance matrix to explicitly account for the fact that the degree of smoothing required had to be estimated.

In terms of formulating our models, we concentrate on describing the difference between our previous additive models and the approach we adopt here. Notably, we:

- consider a spatial effect in longitude by fitting a penalised regression spline to account for longitudinal effects in each of the West and East regions. We additionally allow separate effects in this longitudinal term for data occurring `near` to, and `far` from, the ice edge (categories of the dichotomous `Ice` variable).
- reduce the total number of basis functions considered per term, allowing more complicated model structures to be considered (e.g. spatial terms). We adopted this approach since : i) many higher order tensor product smooths presented in McKinlay et al. (2017) were unimportant; ii) diagnostics provide an indication about whether a sufficient number of basis functions have been used; and, iii) there were sufficient degrees of freedom to capture all important effects identified in our previous work.
- allow shrinkage to do the work of variable selection from a single wide model. We previously simplified models by removing terms that were judged unimportant after shrinkage (i.e. effective degrees of freedom estimated as close to zero), principally for reasons of clarity when presenting results. However, on reflection this was perhaps unwise and constitutes a case of mixing paradigms. Wood's `mgcv` package

does the selection through penalisation of the estimated smooth effects; to then remove terms manually is redundant and could potentially adversely affect variance estimates.
- consider whale age, and interactions between age and body length and year, in models of body weight and blubber thickness.
- use untransformed body weight, rather than log10(BWt), as the response in models of body weight. In McKinlay et al. (2017) we were concerned with estimating the power exponent of the length-weight relationship, and so preferred to linearise the relationship by considering body weight on a log10 scale. For ease of interpretation, here we present models and results on the natural scale of body weight. Usually one might expect an increasing mean-variance relationship when dealing with models of length and weight, but in this case the restricted range of lengths available in the JARPA data meant that for models of body weight a Gaussian response distribution with assumed equal variance was adequate.

For all models, we include smooth main effects for body length, age, year, day within season, and a smooth longitudinal effect for each of the ice strata, `near` and `far`. For females, we include a smooth main effect in fetus length. Despite diatom score proving largely unnecessary for females (Appendix A), we nonetheless include a simple main effect in diatom score (low, high) to accommodate the single data partition that gave any indication of model improvement under our assessment (Female, `BT11` in the East). In most instances, however, this term was estimated as unimportant.

In determining the degrees of freedom allowed to each smooth term prior to shrinkage, we allow 8 basis functions for all main effects and for the margins of tensor product interactions, with two exceptions: the year term is necessarily constrained to 7 basis functions due to the total number of years in the West sampling region, and we allowed 20 basis functions for any effect in longitude in an effort to capture what could potentially be complex behaviour over the distances encompassed in the West and East sampling regions. Diagnostics were used to check our allocation of basis dimensions per term, and in a few cases we were required to slightly extend the search space.

Main effects models are reported for both males and females, though simply for the purposes of comparison with models that include interactions. In all cases it is the interaction models that are used for prediction. Our interaction models include tensor product interactions between `YearNum` and `DateNum`, between `YearNum` and `Age`, between `YearNum` and `BLm`, and between `Age` and `BLm` for both sexes, and additionally for females interactions between `YearNum` and `FetusLength`, and between `DateNum` and `FetusLength`. Models for males are allowed separate smooths in all covariates according to a dichotomised diatom variable, `DiatomF2`, with levels Low and High, while for females we restrict this to a simple main effect only.

We also present results from models that consider body weight as a predictor of `BT11` as response. While neither analysis team previously considered this approach, we have come to appreciate that including body weight as a covariate, and conditioning on body weight for predictions, gives a more useful interpretation of change in `BT11` than otherwise. For example, consider predictions of the long-term trend in blubber thickness, conditioned on fixed values of age, length, and on typical values of body weight for such an animal at the beginning and end of season. Such a model asks about the trend in blubber thickness for exactly the same size/age animal at the beginning and end of each year. In all our models, body weight turns out to be the most important predictor of `BT11` from among those covariates available to the study. Further, it is also an important controlling or mediating variable when considering the effect of body length on `BT11`. In models of `BT11` that include `BLm`, but do not include `BWt`, there appears to be no relationship between `BLm` and `BT11`. However, this apparent relationship is changed appreciably in the presence of `BWt`, showing that there is a significant decrease in `BT11` with body length, after first controlling for body weight. In other words, longer animals have proportionately thinner blubber than their shorter counterparts. This relationship is lost if body length is the only size controlling variable to enter analyses. We therefore prefer models of `BT11` to include the covariate for body weight on the basis that body weight is a strong predictor, it correctly adjusts the relationship between body length and blubber thickness, and it allows predictions to be conditioned on a fixed weight. However, for purposes of comparison we also present models of `BT11` that do not contain body weight as a covariate. In models of `BT11` for male samples we allow separate smooth effects in body weight for each diatom class (Low, High), while for models of female data we allow a main effect in `BWt` as well as an interaction between `BWt` and `FetusLength`.

2.5 Prediction

Unlike the LMM approaches presented by CWH, our models do not provide us with a single parameter estimate of a linear effect and an associated p-value to assess significance of that linear trend. Rather, we obtain estimates of semi-parametric smooth effects with estimates of error. Smooth terms do have p-values associated with them, but despite these having been shown to have good frequentist properties (Marra and Wood, 2011) they should generally be considered as approximate since they do not take into account uncertainty in estimation of the smoothing parameters (Wood, 2017). Our preferred way to assess the importance of any estimated effect is to consider model predictions, which we do in the following ways:

- For models of body weight, we obtain predictions for animals at the beginning of a season by conditioning on relevant covariates to predict for an 'average' animal that has recently arrived in Antarctica. Specifically, for males we predict for median age and median body length for low diatom class animals at the 20th percentile of `DateNum`. For females, we additionally condition on the 20th percentile of fetus length. We obtain these predictions for each year considered in models and present the mean, along with 95% credible intervals (CI) for the mean function and prediction intervals (PI) for point predictions. These results are of interest since they give an indication of how the starting condition of animals may have changed through time. We generally term these predictions as `Early`.

- Predictions from models of blubber thickness are obtain in a similar way to those for models of body weight, except when body weight is included as a covariate. In that case, we use the model for weight to predict early season and late season weights for median length and median age animals, on a year-to-year basis, and those values are used to condition predictions for `BT11` in models when weight is a covariate.

- By a similar approach, we obtain predictions for animals towards the end of the feeding season. We obtain predictions from models of body weight and `BT11` for high diatom load animals (for males) and for the 80th percentile of fetus length (for females), while for both sexes we predict for the 80th percentile of `DateNum`. These results are useful since they provide a time-trend in the finishing condition of animals after spending much of the summer sampling season feeding in Antarctica. We term these predictions `Late`.

- The predictions described above, while of interest in their own right, are really just the stepping stones towards the main metric of interest, which is an estimate of the *improvement* in condition due to summer feeding. This estimate can be obtained by differencing the prediction sets obtained above, each of which is based on 10,000 draws from the posterior of the estimated model parameters. This estimate of improvement is of value since it takes into account the initial starting position of animals given year-to-year variability in arrival condition, noting that this variation is likely due to factors extraneous to the Antarctic ecosystem.

We chose the 20th and 80th percentiles of `DateNum` as the boundaries for our early and late stage predictions on the basis that these points are near to the known end-points for season- and animal-related indicators of time spent in Antarctica, but at the same time are sufficiently away from the domain limits to avoid edge effects due to data sparsity. We note that simultaneously using `DateNum`, diatom score and `FetusLength` for determining predictions makes use of all available information to inform early- and late-season condition. Additionally, we must condition predictions on set values of longitude and the `Ice` variable, the latter indicating distance to the ice edge (`near` or `far`). In all cases we obtain predictions near to the ice edge, since this is where a majority of JARPA catches are taken and where the best feeding conditions might be expected to occur.

Deciding suitable values of longitude for prediction are not so straightforward. We are most interested in predictions for the "end-posts" of the season, the beginning and the end. However, the nature of the JARPA sampling design, usually one- or two-passes per region in alternate years, means that for any given longitude there will not necessarily be data collected from the beginning or end of a season, even when all years are combined. This is not a critical point, since longitudinal effects are often (but not always) judged unimportant in models, likely for the reasons just outlined. Even so, we prefer to obtain predictions at values

of the covariates where data do exist. Our solution is to plot the data according to the important covariates, including longitude, to help us decide values for prediction. We present these chosen values in the results (Section 3) but defer our exploration of the data across the covariate space to the Appendices concerned with model development (Appendices B & C).

A further consideration in relation to longitude is deciding a value to predict on for models where both West and East data are combined. We believe there is no easy and 'natural' answer to this question. If you predict for the longitude dividing the regions then there are relatively few data compared with most other areas. Additionally, you would be predicting for a point in space where two spatially and temporally separate series meet. Yet if you predict for a longitude situated away from the dividing boundary of the regions, then the result would be substantially determined by the data from only that region (and from those distinct years sampled in that region) but through the filter of a model determined by joining the spatially and temporally distinct data from West and East. This seems like a statistical black hole to us, but for the sake of argument we predict for the combined West-East data at the median longitude of all male and female catch positions, separately. This effectively provides a catch weighted median longitude for each sex, and these positions turn out to be 145°E for females and 134°E for males. But we do not recommend a combined analysis of West and East, it is only provided for comparison with other results.

2.6 Structure of results and model nomenclature

We restrict our results section to summaries of the main findings, namely the long-term trend in accumulated body weight and accumulated `BT11` in each year. For `BT11` we show the results from models with and without the addition of body weight as a covariate. For males and females separately, we provide time trends in accumulated condition for animals taken in the West, the East, and from both regions combined. We consider models of `BT11` that do not contain body weight as a covariate to be sub-optimal compared with those that do. Similarly, we consider models of the West and East regions combined to be sub-optimal compare with models of each region separately. We colour these sub-optimal alternatives, which we have provided for comparative purposes, a drab grey in graphics presented in Results and Appendices, reserving the more cheerful colour 'bisque' to distinguish our preferred alternatives. Full model results and larger graphics can be found in Appendices B and C, except for the West-East combined model for which only summaries are provided.

We have arranged model names to have a predictable structure of the form `xx.resp.Mn`. Under this nomenclature, `xx` are two letters defining sex (F or M) and region (W or E). The component `resp` takes one of two values indicating the response variable in the model, either `BWt` or `BT11`. Finally, the `Mn` component gives the model number where `n` is replaced by an integer describing successive models. An example may help:

> `FW-BWt-M3` : model constructed using data from Females from the West (`FW`), the response is Body weight (`BWt`), and this is the third model using these data (`M3`).

For each response `BWt` or `BT11`, we present three models for each West-East partition of the female data: a main effects only model (`M1`), followed by a model containing main effects and all tensor-product interactions (`M2`), and (for models of `BT11` only) finish with model `M3` that contains body weight as a covariate. The main effects model is not used for prediction, but rather is presented for the purpose of showing the degree of model improvement when tensor product interactions are included.

Models for male animals follow a slightly different progression, so that for each response `BWt` or `BT11` we present a main effects model `M1`, a more flexible main effects model `M2` that allows separate smooths for each term according to diatom class (Low or High), an interaction model `M3` that contains the tensor product interactions noted previously but here also interacting with diatom class, and finally model `M4` that is similar to `M3` but includes (for models of `BT11` only) body weight as a covariate.

2.7 Software used and reproducibility

This document was produced from the RMarkdown script `J18v1BWt.Rmd` using RStudio Version 1.1.383, the R environment for statistical computing V3.4.1 (R Core Team, 2017) and associated packages described in Appendix E. Extensive use has been made of R packages `mgcv` (Wood, 2017), `lattice` (Sarkar, 2008) and `bookdown` (Xie, 2016). A copy of the script has been provided to the IWC Secretariat and signatories to the JARPA data sharing agreement under which this work was conducted.

3 Results

In this section we present our primary results, which comprise the long-term trend in seasonal accumulation in body weight and BT11, for animals from the West, the East, and from both regions combined (i.e. with no associated covariate defining region in models). We additionally present results from models of BT11 that contain body weight as a covariate. Note that predicted accumulation of weight and blubber is dependent on early-season and late-season predictions from the same model; their difference provides the relevant result. To keep the results section as concise as possible we defer presenting early-season and late-season predictions to Appendices B and C, along with complete model summaries and larger plots of predictions for all relevant models. Although we present trends in accumulation of body weight and BT11 graphically, point estimates by year and associated estimates of uncertainty could also be provided in tabular form to facilitate their further use in ecosystem models. We conclude this section by briefly examining the length characteristics of the JARPA catches through time, since it turns out that this result is critical for understanding how a spurious negative trend in condition might arise.

3.1 Females

We first consider the long-term trend over year in accumulated body weight and accumulated blubber thickness for female animals captured in the West, the East, and both regions combined (Figure 6, a-i). Set values used for predictions are provided in Table 8, noting that accumulated BWt and BT11 for each year is derived as the difference between predictions for those responses at the 80th and 20th percentile of DateNum for each model considered (denoted Date20 and Date80 in Table 8). For models of BT11 where body weight is included as a covariate we additionally condition, for each year separately, on the predicted early- and late-season body weights for animals of median length and age (see Appendices B.1 & B.3 for individual values). We note there is some variation in the percentiles of DateNum between regions due to variability in temporal sampling coverage year-to-year.

Accumulation of body weight for female animals from the West (Figure 6a) seems to decrease over the first four years (1990-1996), then plateaus for the remaining four time points. In the East (Figure 6b), there is a slight decline to around 1997, followed by a comparatively large recovery peaking in 2001. Combining both West and East (Figure 6c) results in a fairly stable series to 1994, followed by a decline to 1997, then incline to 2000, and finally stabilising in the final sampling years at levels similar to those observed at the beginning of the series. We note the the signal from analyses of both West and East combined in some ways approximates the signal obtained from each region separately, with some evidence that the signal in the East dominates. For example, the increase at around 2000 in Figure 6c is clearly not present in 6a, indicating that the larger female sample sizes in the East dominate the combined result.

The decrease in the accumulation of BT11 in models unadjusted for body weight for females from the West (Figure 6d) provides, in our opinion, the strongest and most consistent *apparent* signal for declining condition to be obtained from the JARPA data. Here we see a decrease in BT11 of around 1 cm accumulated between day 37 and day 85 over the eight years sampled in the West (1.58 cm in 1990 to 0.61 cm in 2004). It is worth contrasting this result with the equivalent model but for adding body weight as a covariate (Figure 6g), which shows this decline to be much less pronounced, a drop of only around 0.5 cm over the period (1.41 cm in 1990 to 0.85 cm in 2004).

Table 8: Percentiles (20, 50, 80) for variables associated with female predictions for the West and East sampling regions, and from both regions combined.

Region	BLm50	Age50	Date20	Date80	Fetus20	Fetus80	Longitude
West	8.9	19	37	85	23	85	73 E
East	8.8	20	56	96	37	120	171 W
Combined	8.9	20	50	93	30	107	145 E

We have argued that body weight should be included in models of `BT11` so that the effect of body length on blubber thickness is properly accounted, and explore this point further in Section 3.3. Here we point out that neglecting to account for `BWt` in models of `BT11` can be seen in the West to almost double the estimated decrease in accumulated `BT11` (Figure 6d), compared with the same model but for the addition of `BWt` as a covariate (Figure 6g). We find this discrepancy noteworthy, and consider that past work is likely to have been affected by not including body weight as a covariate.

We can be brief in our summary of accumulated blubber thickness in the East, for models with and without body weight as a covariate; there was a shallow decline to around 1997, followed by a shallow incline to 2005 (Figure 6e,h). Models for both regions combined showed virtually no variation in predicted accumulation of `BT11` over the period of JARPA, with approximately 1.5 cm of `BT11` accumulated between the 20th and 80th percentiles of `DateNum` in each year of the program (Figure 6f,i).

We conclude by noting that confidence intervals for the West-East combined models are, due to increased sample sizes, appreciably narrower than for either the West or East region individually. We do not consider there to be any statistical justification for combining the West and East data, which are exactly confounded with respect to space and time, and analyses that do so will be underpinned by over-optimistic estimates of precision.

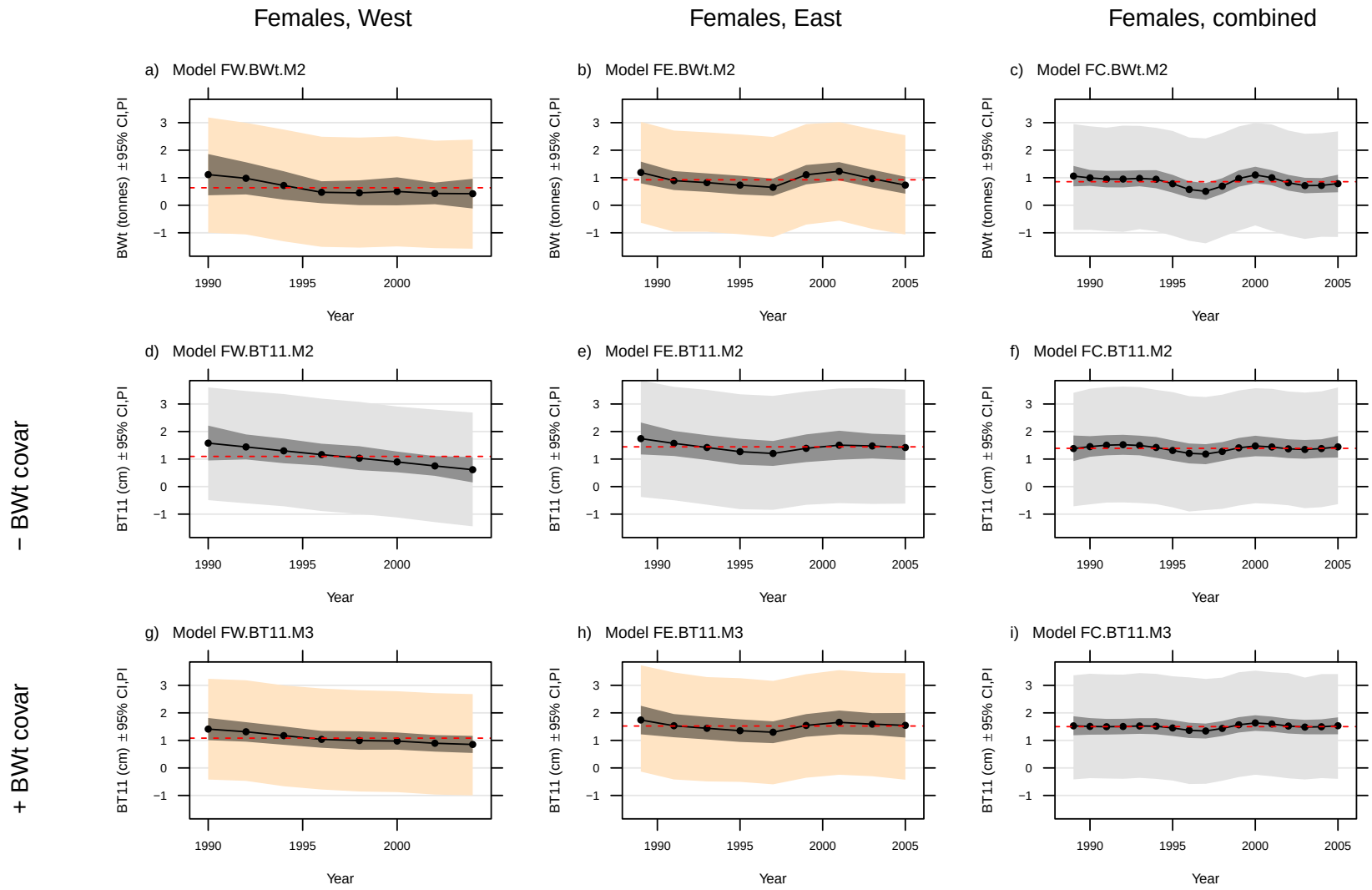

Figure 6: Long-term trend in accumulated body weight (tonnes) and blubber thickness (cm) between the 20th and 80th percentiles of DateNum for median length, median age females. Columns indicate results from the West, East, or both groups combined, while rows indicate results for accumulated body weight, accumulated BT11 unadjusted for body weight, or accumulated BT11 adjusted for body weight (top to bottom). Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided, with our preferred models a,b,g,h highlighted in colour. Dashed red lines represent overall mean(y) in each figure.

Table 9: Percentiles (20, 50, 80) for variables associated with male predictions for the West and East sampling regions, and from both regions combined.

Region	BLm50	Age50	Date20	Date80	Longitude
West	8.4	17	28	83	75E
East	8.3	18	35	95	145E
Combined	8.4	18	32	91	134E

3.2 Males

Trends in the accumulation of body weight and blubber thickness for models of the male data are provided in Figure 7, and follow the same conventions set out for female results. The set values we used for predictions are provided in Table (9) and, for early- and late-season body weights, in Appendices C.1 & C.3.

There is virtually no trend for accumulated body weight or accumulated blubber thickness (with or without including body weight as a covariate) for males in the West, though this is evident for a slight decrease in the last year or two of sampling (Figure 7a,d,g). We note that the confidence intervals for these estimated decreases show that the overall series mean (red dashed line) remains within estimated sampling variability across the series.

For males from the East there appears very little variation in estimated accumulated body weight over the sampling period (Figure 7b). Trends in accumulated `BT11` can be best described as slight; there is a slight decrease in accumulated `BT11` (when adjusted for `BT11`) in the mid-1990's, followed in the next few years by a slight increase back to around average levels (Figure 7h). We note these slight decreases are small in comparison with the size of the estimated confidence intervals.

Turning to both regions combined, the accumulation of body weight and blubber indicate very little change over the first 12 years of the program, followed by a slight downturn in the mid-2000's (Figure 7c,f,i). However, this minor decline again reached average levels by 2005.

Concerning all results presented in this section, for males and females, we draw the readers attention to the width of confidence intervals for the West-East combined models compared with separate models for each region. We note that in the combined case the CI's are appreciably smaller, which is understandable given the larger sample size available to these models. However, we see no good statistical justification for combining these data, or in relying on the false sense of confidence models based on combined data provide.

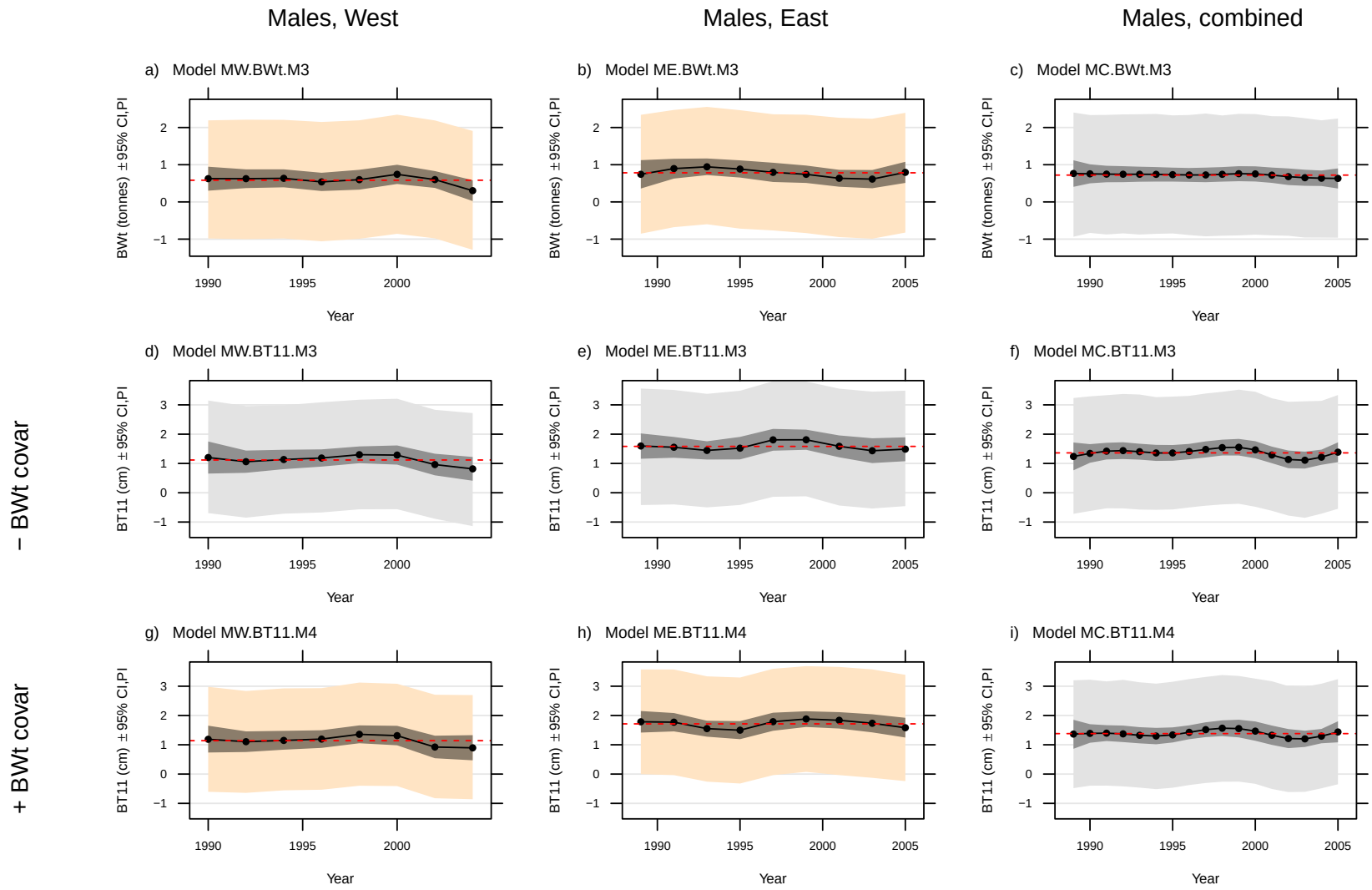

Figure 7: Long-term trend in accumulated body weight (tonnes) and blubber thickness (cm) between the 20th and 80th percentiles of DateNum for median length, median age males. Columns indicate results from the West, East, or both groups combined, while rows indicate results for accumulated body weight, accumulated BT11 unadjusted for body weight, or accumulated BT11 adjusted for body weight (top to bottom). Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided, with our preferred models a,b,g,h highlighted in colour. Dashed red lines represent overall mean(y) in each figure.

3.3 Why might other analysts be finding global negative trends in condition?

Cunen, Walloe and Hjort (2018, 2017) report significant linear declines in minke whale condition over the entire JARPA sampling region for the duration of the program, yet our work finds only a slight decline in a relatively small proportion the JARPA data (females in the West, 15% of entire dataset). The reasons for these differences are likely multifaceted, and some of these aspects we have discussed previously. These include the fact that CWH are fitting linear models when some trends are in fact non-linear, that they combine males and females in a single analysis in which sex effects are not fully separable, that they *a priori* restrict the model space to exclude any interactions with the fixed effect of year, and that they combine data from the West and East sampling regions when these data are exactly confounded with respect to space and time. If not adequately addressed in models, we believe these issues cannot help but contribute to provide poor estimates of changes in condition.

Yet there are still two key, related issues that we have not yet discussed, namely how blubber thickness changes in relation to body length, and the fact that average animal lengths in the JARPA catches have been increasing over time. Individual model results, which we develop and summarise in Appendices B and C, indicate that in addition to body length, body weight should be included in models of `BT11` as a size controlling variable. Our results show that neglecting `BWt` as a covariate in models where `BT11` is the response can have substantial impacts on estimated trends; for example, the estimated trend in blubber thickness for females in the West region is inflated almost two-fold in the absence of body weight as a covariate (Figure 6,d vs g). In this section we explore why this might be so.

We begin by comparing partial effects plots for the variable body length (`BLm`) from models of blubber thickness that exclude or include body weight (`BWt`) as a covariate (Figure 8). For models not accounting for `BWt`, the partial effect of body length is largely unrelated to blubber thickness (Figure 8a-d; see Appendix D for partial effects plots in the context of all model terms). However, introduce `BWt` as a covariate to these same models and this relationship changes dramatically; longer animals have proportionately less `BT11` than shorter animals (Figure 8e-h). This relationship is evident in all models of `BT11` we considered (i.e. for males and females in both the West and East regions).

Body weight is the most important predictor of blubber thickness available to this study, a result that is hardly surprising; the weight of an animal will usefully inform that animal's condition. The implication of the negative relationship between `BT11` and animal length, when body weight is accounted, is that longer animals have proportionately lower energy requirements than shorter ones, a consequence of the typical allometric relationship between body mass and metabolic rates (Kleiber, 1961; Leaper and Lavigne, 2007).

Importantly, we can confirm from the raw data the finding that blubber thickness proportionately decreases with body length. Demonstrating this relationship externally to our models is an important step because it shows that this effect is not dependent on the analysis decisions we have taken. To show this we plot blubber thickness as a proportion of body length, against body length, for animals of each sex in each region (Figure 9). In order to control for seasonal effects, and to show that relationships are consistent at both ends of a sampling season, we additionally separate results for early season samples with low diatom scores and late season samples with high diatom scores. Colour and symbol type are used to distinguish individual years, with linear fits provided for those years with $>= 8$ samples. We additionally provide an overall fit for all years combined (black, with 95% CI). Results for individual years can be variable, particularly for those years with small sample sizes, but overall there is a clear trend for blubber thickness to proportionately decrease with increasing body length. This effect seems consistent for male and female animals in both the West and East, and is consistent at both ends of the JARPA sampling season.

Our models of `BT11` that included body length as the only size-controlling variable failed to capture this relationship, instead showing – incorrectly – that blubber thickness remains approximately constant with increasing body length (Figure 8a-d). However, introducing body weight as a second size controlling variable allows the true effect of body length to be correctly captured in models (Figure 8e-h).

The final piece of the puzzle explaining why the relationship between body length and blubber thickness is critically important for assessing potential changes condition is provided by examining how body lengths in JARPA catches have changed over the sampling program. To assess this we plot body length against

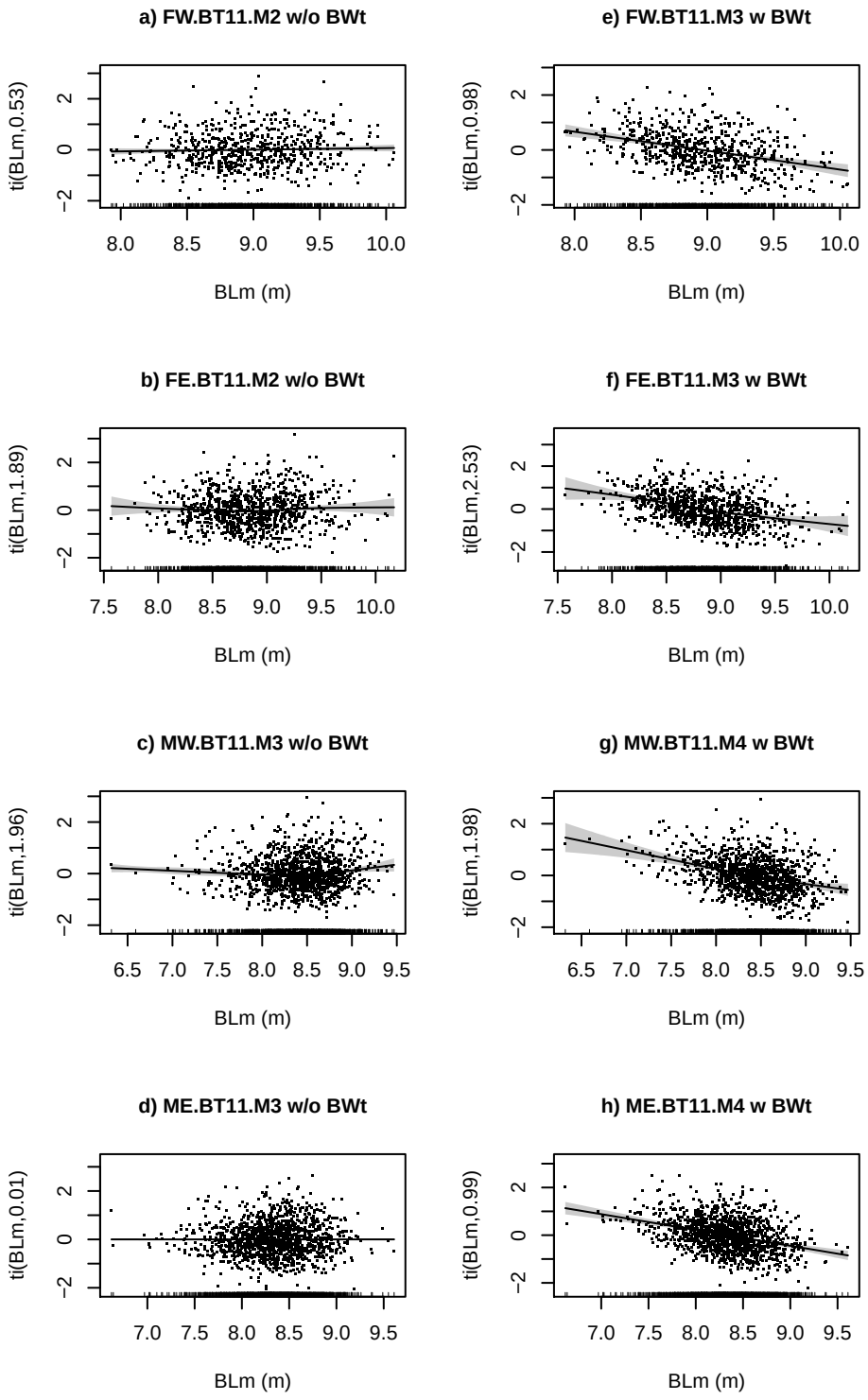

Figure 8: Partial effects plots of body length for models of BT11, with and without the mediating covariate BWt. Panels a-d show the partial effect of body length in final models used for prediction for females from the West (a) and East (b), and males from the West (c) and East (d), when models do not contain BWt as a covariate. Panels e-h show the partial effect of body length for these same models except that now BWt is included as a covariate.

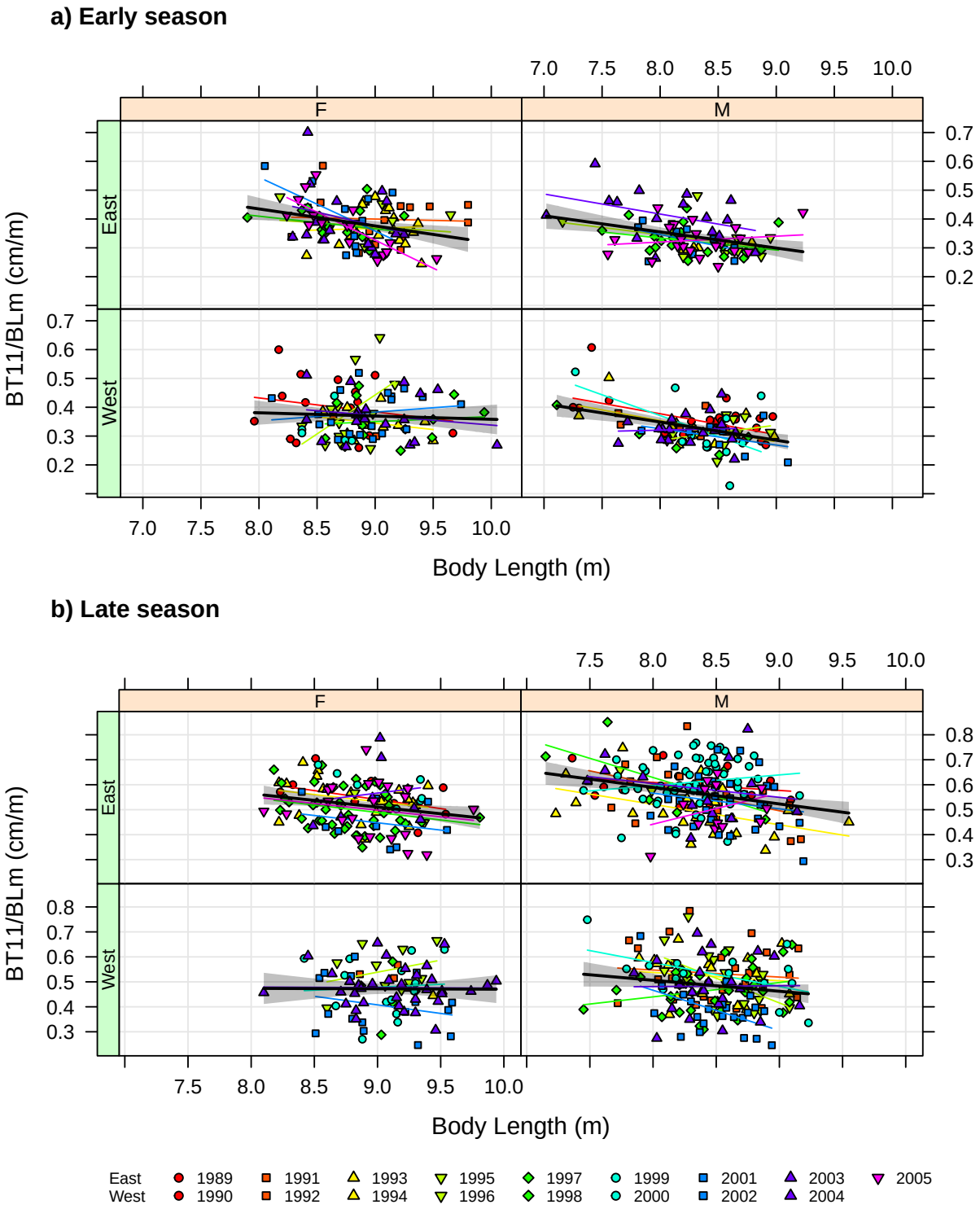

Figure 9: Plots of blubber thickness (BT11) as a proportion of body length (BLm) against body length for each sex (F, M) from each region (West, East) based on the primary measurements, showing that blubber thickness (proportionately) decreases with increasing body length, for: a) early season data, determined by choosing low diatom score animals captured before the 20th percentile of DateNum; and, b) late season data, determined by choosing high diatom score animals captured after the 80th percentile of DateNum. Simple linear fits are shown for individual years with >= 8 cases (colour), with overall fit and 95% CI for all data combined shown in black.

year for each of the sexes, stratified by month and region (West, East) (Figures 10). We fit linear trends with 95% CIs to each division of the data to give an indication, in an exploratory sense, of the main trend and variability in body length. These results show a slight but noticeable increase in the average length of animals landed over time for most components of the JARPA data, particularly in the West. We also see a consistent decrease in the average lengths of catches in December in the East region.

On the basis of results presented earlier in this section, these systematic increases or decreases in the lengths of animals in catches will potentially provide a false signal of decreasing or increasing blubber thickness in models of `BT11` that do not account for `BWt`. Exactly how changing lengths in catches will manifest as a false signal in models omitting weight is not straight forward to anticipate, since condition is not simply influenced by length, but also by how weight is changing with length. However, we can be certain of two things: longer animals have proportionately less blubber per unit length than shorter animals, and the length of animals captured during JARPA have systematically increased over the life of the program. Our results comparing models of `BT11` with and without the addition of body weight as a covariate shows that, when we do correct for body weight, the only appreciable negative trend we detected (females in the West) was effectively halved.

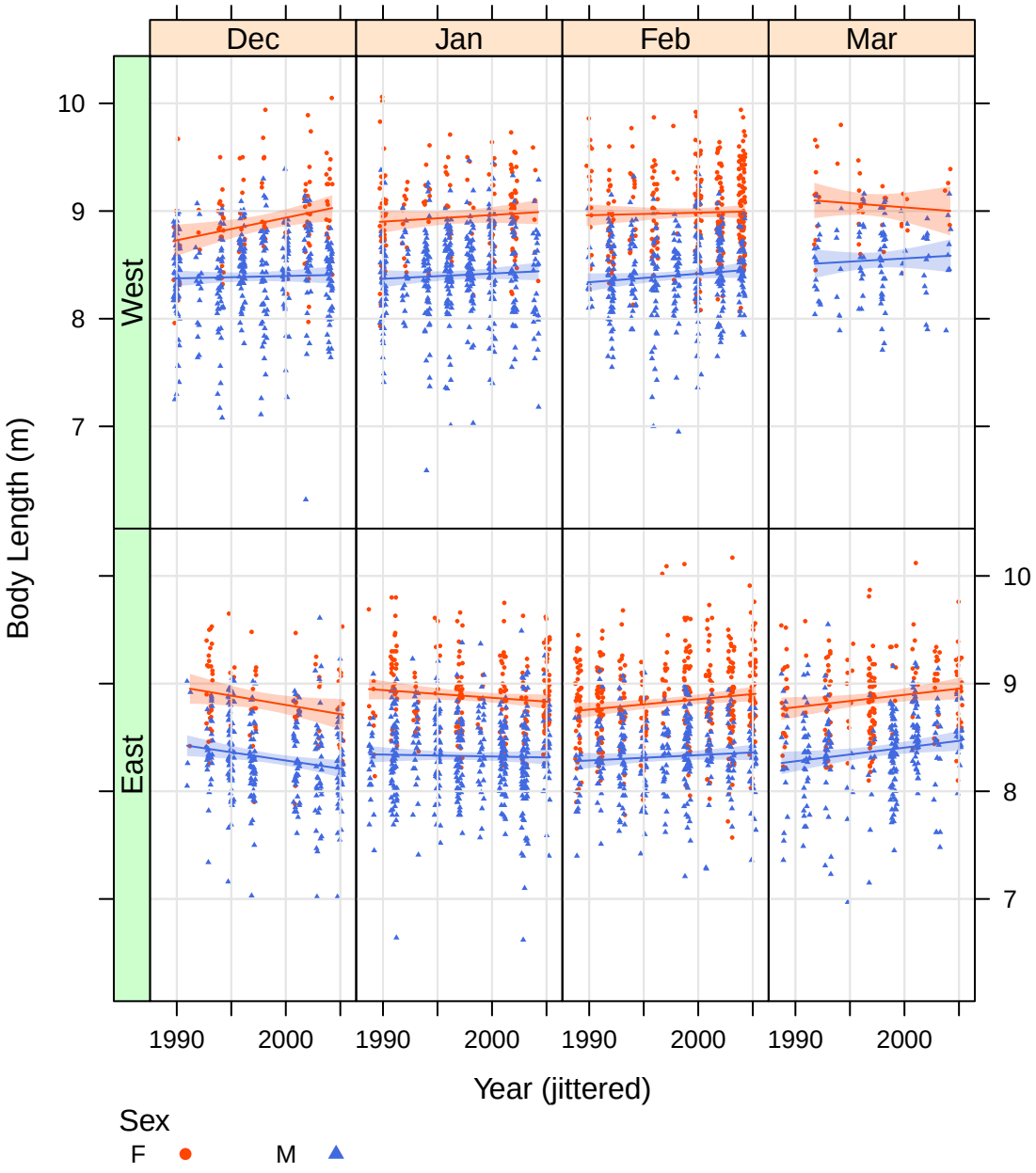

Figure 10: Trend in body length over year, by month and region (West, East), showing sex using symbols. Simple linear fits with 95% CIs are shown.

4 Discussion

The methods and analyses presented here have evolved since McKinlay et al. (2017), including addressing the issues that were raised at SC67A, but our overall conclusions have not substantially changed. We remain unconvinced by analyses that report a substantial, uniform decline in energy storage in minke whales during the JARPA program, such as those presented in Konishi et al. (2008), Konishi and Walloe (2015), or most recently by Cunen et al. (2017). Our analyses show no evidence of substantial changes in condition, positive or negative. What we have detected are *apparent* declines in body weight or blubber thickness when only the end (but not the start) of the summer season is considered, or when models of `BT11` are not corrected for body weight. But if you take arrival condition into account and base analyses on the within-season accumulation of condition, and adjust for weight in models of `BT11`, then the most we can acknowledge are some *slight* decreases or increases in `BT11` in *some* of the data. After adjusting for `BWt`, the largest decline in `BT11` was observed to be around 0.5 cm over 8 years for females in the West. However, a change in accumulated blubber of this magnitude over the entire program seems of a similar magnitude to biennial variation in some part of the data (e.g. 2000-2002 for males in the West, or 1995-1997 for males in the East, Figures 7g,h). Trends in the accumulation of body weight for males seem reasonably constant over the JARPA program, while accumulated weight for females shows slight evidence of a reduction in the seasonal accumulation of weight for 1990-1996, but not thereafter. Are these deviations important for constructing ecological models, or for the management of minke whales? We suspect not but feel that any judgement in that regard should be subject to first developing appropriate energy budgets for minke whales, something that is beyond the scope of this work.

We have been reminded throughout this analysis just how small samples sizes are across the design space, and in particular the limited capacity for the design to realise data that can inform spatial effects, or separate spatial effects from short-term and long-term temporal effects. In this and our previous work we have demonstrated that there are clear differences in the spatial distribution of animals, and in the characteristics of those animals, across the geographic extent of the JARPA program. Yet both CWH and ourselves found little evidence for spatial effects on condition. We consider it more likely that the one-pass (or at best two-pass) sampling program is simply not able to resolve spatial effects in light of the dominance of the within-season effect on condition.

We have demonstrated that, for this dataset, not all data should be analysed together. Although this is certainly an arguable proposition, since it goes against much accepted orthodoxy, part of the problem in this particular instance is that doing what is best for estimation (splitting) is at odds with what might be preferred for assessing significance (larger sample sizes). Our focus has been estimation and obtaining the best inferences from data that are ill-suited for the questions being asked of them. We do not say that some over-arching, single model cannot in principle cleanly test the hypotheses of interest (if hypothesis testing were truly the goal), but given the characteristics of the data we cannot see how such a model can be formulated without making a number of highly questionable assumptions. The West-East split in sampling seems insurmountable for a single-model approach given the data at hand. Both sexes could potentially be included in the same model, but unless the sexes remain largely separable through rich parameterisation then, again, some large assumptions must be made. But if we abandon the somewhat narrow goal of achieving statistical significance, and instead bring our attention to focus on estimation, things become clearer. In this latter approach, we conduct separate analyses on coherent subsets of data that ensure inference is as sound as it can be, in the circumstances. We also advocate basing decisions about the importance of effects on their estimated values in relation to the problem studied, rather than on statistical significance which is inextricably linked to sample size.

We anticipate some resistance to the idea of including body weight as a covariate in models of `BT11` (but look forward to being pleasantly surprised). Aside from being the strongest predictor, accounting for up to 30% of the total deviance in models of `BT11`, we find the correction this covariate makes to the effect of body length to be compelling. If, after controlling for seasonality, longer animals have proportionately less blubber than shorter ones, and this fact is only revealed by including body weight as a covariate, then we believe the case is made. The raw data clearly demonstrate this effect independently of our models. In the two instances where models of `BT11` (uncorrected for weight) showed a downward trend in the accumulation of

BT11 over summer feeding (females and males in the West), the correction for body weight greatly reduced this apparent effect. We demonstrate that the JARPA data themselves confirm why omitting body weight as a covariate can induce spurious trends, showing that in the West there has been an increasing trend in the length of animals captured over the life of the program. This increase in body length through time, without correcting for weight, cannot help but provides a false signal for changing blubber thickness.

Our analyses for male data remain unsatisfactory to us due to the substantial ambiguity associated with diatom scores. The clear differences in the distribution of scores year-to-year, and particularly the contrast between scores in the early years and later years of the JARPA program, leads us to believe that the program has sampled different components of the overall population through time, or that there have been systematic biases over years in scoring animals (data recording) or in the catching behaviour of the fleet. We remain somewhat skeptical of the results for males on this basis. For the female analyses these problems are circumvented by using fetus length as a continuous, relative measure of time spent in Antarctic waters. We believe this latter approach to be supported by the available data, however it is also underpinned by several assumptions that are, at present, untestable. These include the assumption that all females migrate to summer feeding grounds at a relatively fixed interval after conception, and that the time it takes to migrate is approximately the same for all female animals. These issues, for males and females, are the key to analyses of these data for the purposes of disentangling within-season from between-season signals in condition. Without a useful measure of time spent feeding over the summer months it is impossible to cleanly separate these signals.

Some analysts have emphasised the importance of day of season (DateNum) in determining within-season condition. However, the date an animal is captured — early or late in the season — often appears unrelated to how long it has been in Antarctic waters (if diatom scores are to be believed) (Figure 5). The data indicate that animals with low diatom load, or small fetus length in the case of females, arrive throughout the summer season. In other words, an animal can be captured toward the end of the season and be in relatively poor condition simply because it has only just arrived in Antarctica. Again, the key to determining the within-season improvement in condition is in determining how long an animal has been present during the summer feeding season, not when it is caught. We believe the JARPA data provide measurements that are poor indicators of the length of time animals have been feeding over the summer months.

We cannot conclude a modelling exercise that has concentrated on estimating non-linear relationships without mentioning the pitfalls of assuming linearity. Estimates of linear relationships can be adversely affected by high leverage, influential data near the boundaries of the design space. Much has been written about this phenomenon. We have nothing against linear relationships, it is rumored that they even occasionally occur in nature, but checking this assumption through non-linear techniques or by applying robust/resistant methods is advisable. In the case of our JARPA models, we do see some predictions with higher or lower year estimates at either end of the series; these are prime candidates to mislead linear methods.

A linear, significant trend does not necessarily guarantee us a good or useful model. CWH, in their primary work from 2017, explicitly acknowledge this point when they say:

> "A third perspective is that the best model is the one which gives the most precise estimates of the parameter of primary interest. If we agree with this perception on what the"best model" is, **we implicitly state that we are potentially willing to use a model which is not a plausible description of the system at hand** - as long as we get good and precise answers out of it. Such a view is warranted in a lot of situations (especially when we bear in mind that all models are probably quite wrong), but not always. In the context of the JARPA data, the discussion mainly focuses on whether there is a significant linear effect of year or not, and there is agreement on that precisely estimating the linear trend is a crucial point. In that light, a FIC approach with B year as the focus parameter seems warranted." (Cunen et al., 2017) (our emphasis)

These words from CWH are important in several respects. They first remind us that the FIC approach is underpinned by an assumption that the focal parameter is of prime importance, and then alert us to the fact that we must be willing to potentially make sacrifices towards that end. We are told that one of the things that might have to be sacrificed is plausibility; the model we adopt under an FIC approach may

not necessarily be a good description of the system studied, but that is acceptable, because we get precise estimates of a linear trend from this potentially implausible model. Our own results indicate that a global, linear decline in condition over the period of JARPA is not supported by the data. Our models show that certain features of the data, such as high or low values at the beginning or end of a series, are likely to cause linear techniques to find significant trends where arguably none exist. We have also shown that failing to include an important covariate, such as body weight in models of `BT11`, has the potential to misrepresent the relationship between other covariates and the response. In the case of body condition, weight is an important mediating variable on the relationship between body length and blubber thickness. When combined with non-stationary sample characteristics, such as an increase in average length in the catch over time, this type of model misspecification can induce significant but entirely spurious trends. In light of these considerations, we can only conclude that results that do assert a global, negative trend in condition are based on models that are indeed implausible descriptions of the JARPA data.

We prefer our models to be plausible descriptions of the system we're studying. Our analyses have focused on determining, as best we can, any indications in the JARPA data of changes in minke whale condition — what component of the population, the magnitude, and the shape of any signal. We have been unconcerned with assuming linear trends or with determining p-values. We understand why a focus on linear trends and p-values has developed, but it is time to move on from that focus. At SC67A, the Scientific Committee noted that in relation to trends in minke whale condition that "... when viewed in the wider context of the Committee's interest, especially from a management perspective, significance at the 5% level is probably not the most important related issue", and that for secondary use of these results in multispecies models the "... inputs desired would be annual 'standardised' estimates for blubber thickness with associated coefficients of variation" (IWC, 2017). We hope our work provides a useful starting point to that end.

5 Acknowledgements

We thank the Institute of Cetacean Research for providing the JARPA catch data so that we might independently assess claims for a decline in minke whale body condition over the period of the JARPA sampling program. We thank our colleagues Cunen, Walloe and Hjort; no-one will look at our work more closely.

Appendices

A Fetus Length vs Diatom Score

In this section we establish that, for the models and data considered in this work, fetus length is a better predictor of body weight or `BT11` than diatom score for female animals. Additionally, we show that diatom score offers little to no improvement for models that already contain a term for fetus length. We propose that fetus length is likely acting as a proxy for time spent feeding in Antarctic waters; the larger the fetus, the longer a female has been feeding, and the greater potential for within-season improvement in condition. However, the nutritional demands of a fetus obviously increase with fetus size, so the improvement due to feeding will be offset by the increasing demands of the growing fetus. While these two impacts on condition are conflated, the use of GAMs effectively captures any non-linearity that may arise due to this interaction.

The results presented in this section are not intended to represent a full analysis of the female data, but rather an assessment of the relative importance of diatom score and fetus length in explaining condition. We focus here on main effects in all relevant covariates, as well as an interaction between diatom score and all other covariates. We note that diagnostic plots are not presented for models developed in this section, though these have been checked and found satisfactory. Code for diagnostics are provided but have been "commented out" to spare the reader a much larger analysis document.

We argue in Section 2.3.1 that, to obtain reliable inferences about trends in condition, the confounding between space and time that is created by virtue of the West-East sampling regime must be respected. To this end, we have advocated separate analyses for each region. We therefore undertake an assessment of diatom score and fetus length for each region separately, for both body weight and blubber thickness (`BT11`) as responses.

A.1 West Region

Consider the association between fetus length and diatom score in each year in the West region (Figure 11). This shows the increasing, positive relationship between diatom score and fetus length. Per-year Pearson correlations are in the range 0.5-0.8, and the more appropriate, nonparametric Kendall's Tau is in the range 0.3-0.5. The over-representation of diatom scores 0 and 2 is clearly evident in all years except for years 1990 and 1992, suggesting that recording protocols for diatom score may have differed in the early years of the sampling program.

A.1.1 Body Weight

We begin by fitting main effects models to body weight as the response. As in our previous work, we include smooth terms for body length and year, but now additionally include terms for whale age and a longitudinal effect, the latter of which is allowed to vary according to ice stratum (a 2-level factor, 'near' and 'far'). Twelve basis functions are allowed for all terms except `YearNum` which is necessarily constrained to 7 basis functions, and longitude for which 20 basis functions were allowed. Models were then expanded by sequentially adding terms for fetus length and diatom score (as a 5-level factor) to assess the marginal importance of these terms. Finally, we assessed a rather more flexible model that allowed separate smooths for all main effects for each level of diatom score.

Model results for body weight are reported in Table 10. Results are discussed in terms of improvement to model R^2 values (noting that these are approximate) and differences in AIC and BIC values. The first thing to note is that all five models considered have reasonable explanatory power, accounting for 60-70% of model deviance. *BWtWest+Fetus* showed an improvement in fit over *BWtWest* by about 8%, with fetus length highly significant (<0.001). *BWtWest+Diatom* also proved a better fit compared with *BWtWest*, increasing R^2 by about 5% and revealing significant additive differences for diatom scores 2,3 and 4 over score 0. However, adding diatom score and fetus length to the same model (*BWtWest+Fetus+Diatom*) revealed

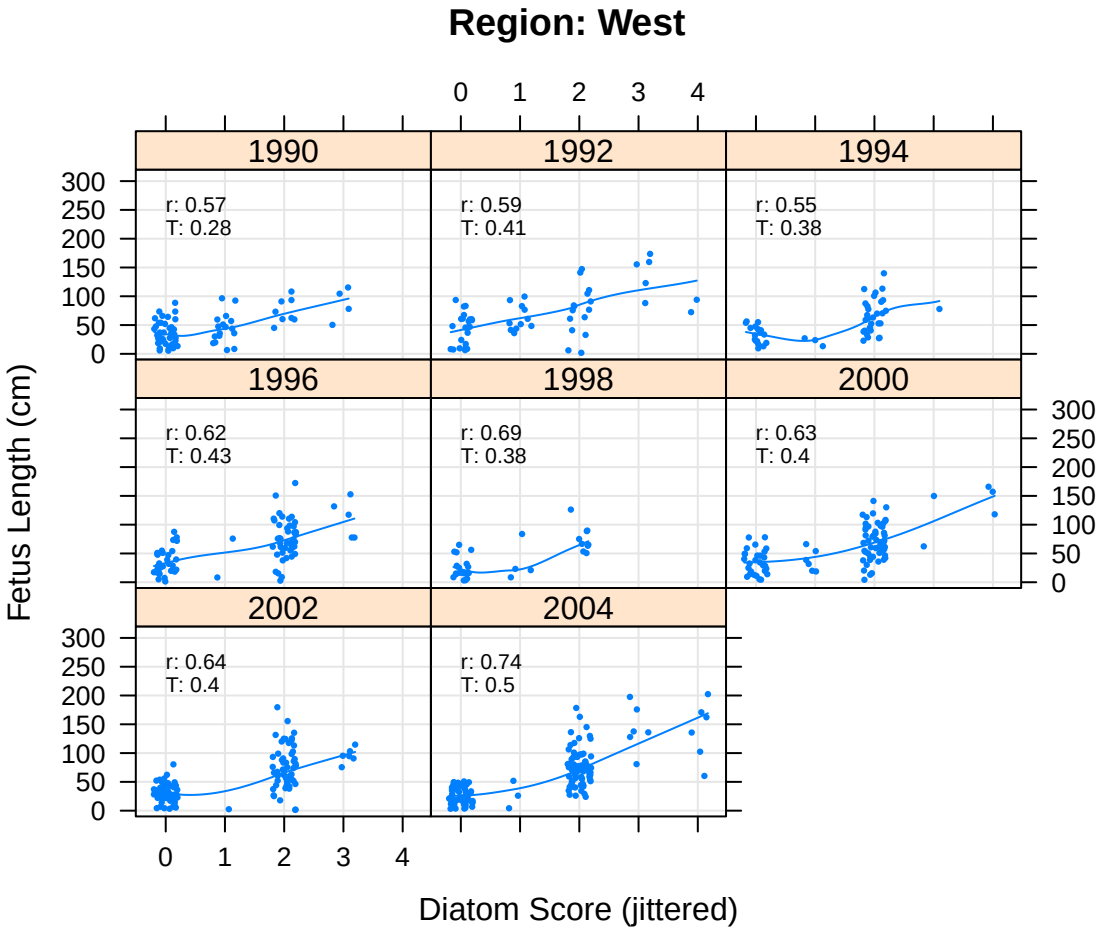

Figure 11: Relationship between fetus length and diatom score (0-4 scale) for each sample year in the West region. Pearson correlations (r) and Kendall's tau (T) are displayed with each panel, and fitted lines are loess smooths with span=2/3.

this model to be an equivalent fit to *BWtWest+Fetus* based on R^2, but an appreciably worse fit based on AIC and BIC. This result was also reflected in the approximate significance of diatom score parameter estimates in *BWtWest+Fetus+Diatom*, which were all non-significant. This assessment of the importance of diatom score assumes an additive effect due to diatom score on the smooth effects estimated for other covariates. An alternate, more flexible formulation to consider is that different levels of diatom score lead to different smooths on other covariates. We assess this possibility in model *BWtWest, by=Diatom*, in which for each covariate we fit a different smooth for each level of diatom score. Model results for individual smooths are somewhat opaque since these represent both main effects and a full interaction with diatom load as a categorical variable[6]. Results of the overall model fit are clear, however; the explanatory power of *BWtWest, by=Diatom* increases marginally (<1%) over *BWtWest+Fetus*, however both AIC and BIC are appreciably worse compared with the fit for *BWtWest+Fetus*.

We conclude that for analyses of body weight for female animals sampled in the West, there is no utility in including diatom score in models that already contain fetus length.

[6]Note that basis dimension for smooths changes between levels of diatom score due to differing sample sizes, and that, due to shrinkage, some effective degrees of freedom are almost zero.

Table 10: Summary of additive model fits to `BWt`, assessing the importance of diatom load and fetus length for females captured in the West. Models show main effects excluding fetus length and diatom score (*BWtWest*), the addition to *BWtWest* of a smooth in fetus length (*BWtWest+Fetus*) or diatom score as a 5-level categorical factor (*BWtWest+Diatom*), the addition of both diatom score and fetus length (*BWtWest+Fetus+Diatom*), and finally a model that allows separate smooths in all covariates (including fetus length) for each diatom score category (*BWtWest, by=Diatom*).

	BWtWest	BWtWest+Fetus	BWtWest+Diatom	BWtWest+Fetus+Diatom	BWtWest, by=Daitom
EDF: ti(BLm)	2.5910 (11.00)***	2.9690 (11.00)***	2.8000 (11.00)***	3.0279 (11.00)***	
EDF: ti(Age)	0.9796 (11.00)***	0.9766 (11.00)***	0.9788 (11.00)***	0.9776 (11.00)***	0.9812 (11.00)***
EDF: ti(YearNum2)	4.8951 (6.00)***	5.0981 (6.00)***	5.2608 (6.00)***	5.1798 (6.00)***	
EDF: ti(DateNum)	1.9316 (11.00)***	0.4465 (11.00)	2.1686 (11.00)***	0.6349 (11.00)	
EDF: ti(LongNum):Icefar	0.7267 (19.00)	1.3991 (19.00)*	0.7702 (19.00)*	1.2578 (19.00)*	0.9595 (19.00)*
EDF: ti(LongNum):Icenear	0.8214 (19.00)*	1.5604 (19.00)**	1.8143 (19.00)*	1.6867 (19.00)**	1.0023 (19.00)**
EDF: ti(FetusLength)		1.6665 (11.00)***		1.4500 (11.00)***	
DiatomF1			−0.1001 (0.12)	−0.1605 (0.12)	−0.1709 (0.18)
DiatomF2			0.3190 (0.07)***	0.0202 (0.08)	0.1407 (0.08)
DiatomF3			0.6888 (0.15)***	0.0472 (0.16)	0.0311 (0.26)
DiatomF4			1.0060 (0.24)***	0.2034 (0.25)	1.1910 (0.28)***
EDF: ti(BLm):DiatomF0					2.3706 (11.00)***
EDF: ti(BLm):DiatomF1					0.9829 (11.00)***
EDF: ti(BLm):DiatomF2					2.4314 (11.00)***
EDF: ti(BLm):DiatomF3					2.2627 (10.00)***
EDF: ti(BLm):DiatomF4					0.9496 (9.00)***
EDF: ti(YearNum2):DiatomF0					1.3763 (6.00)
EDF: ti(YearNum2):DiatomF1					1.2634 (6.00)**
EDF: ti(YearNum2):DiatomF2					5.2956 (6.00)***
EDF: ti(YearNum2):DiatomF3					0.2670 (6.00)
EDF: ti(YearNum2):DiatomF4					0.8376 (2.00)**
EDF: ti(DateNum):DiatomF0					0.0009 (11.00)
EDF: ti(DateNum):DiatomF1					1.2321 (11.00)*
EDF: ti(DateNum):DiatomF2					0.0002 (11.00)
EDF: ti(DateNum):DiatomF3					0.0001 (9.00)
EDF: ti(DateNum):DiatomF4					0.0003 (6.00)
EDF: ti(FetusLength):DiatomF0					1.4356 (10.00)***
EDF: ti(FetusLength):DiatomF1					1.5414 (11.00)**
EDF: ti(FetusLength):DiatomF2					1.5155 (11.00)***
EDF: ti(FetusLength):DiatomF3					0.9293 (7.00)***
EDF: ti(FetusLength):DiatomF4					1.2141 (5.00)*
AIC	1510.7137	1405.7005	1468.5961	1409.9897	1403.4544
BIC	1583.6251	1491.9883	1569.2421	1514.4014	1587.9998
Log Likelihood	−739.0029	−683.4960	−711.7232	−681.5755	−660.3339
R^2	0.6011	0.6636	0.6303	0.6634	0.6775
Num. obs.	638	638	638	638	638

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

A.1.2 Blubber Thickness

Following the same strategy as outlined for body weight, we next develop models for blubber thickness (`BT11`) for female animals sampled in the West. One difference between models for body weight and blubber thickness is that we include body weight (`BWt`) as a covariate in models of `BT11`. While models of body weight accounted for up to 60% of the variation in the data, models of `BT11`, using the same covariates, explained between 37% (`BT11West`) and 48% (`BT11West, by=Diatom`) of the variation in blubber thickness (Table 11). Model `BT11West+Fetus` revealed that fetus length was again a strong predictor, here improving R^2 by around 8%. In contrast, model `BT11West+Diatom` only improved R^2 by around 4.3%. Adding both fetus length and diatom score simultaneously (`BT11West+Fetus+Diatom`) to model `BT11West` resulted in a model R^2 of 45.7%, an improvement of <1% compared with `BT11West+Fetus`. We note that model `BT11West+Fetus+Diatom` is a marginally better fit than `BT11West+Fetus` based on AIC, but an appreciably worse based on BIC. Our final model, `BT11West, by=Diatom`, explains the highest amount of variation in `BT11` of those considered (R^2 of 48.2%), however this model is richly parameterised and so only offers marginal improvement over `BT11West+Fetus` based on AIC, and no improvement whatsoever based on BIC.

These results indicate that, for female animals in the West, diatom score is of negligible importance in models of `BT11` that already contain a term for fetus length. While a rich model that allowed separate smooths across most covariates for each diatom score category did improve model R^2 values by 3.2% compared with `BT11West+Fetus`, the more complex model was only slightly better based on AIC, and appreciably worse based on BIC.

Our overall conclusion is that, in relation to models of `BWt` and `BT11` for the West region, diatom score is for all practical purposes extraneous in the presence of a term for fetus length. Nonetheless, in acknowledgement of the slight improvement to model fits afforded by including diatom score over and above fetus length, we include a main effect of diatom score (as a 2-level categorical variable, with levels low and high) in all models for females contributing to our main results.

Table 11: Summary of additive model fits to `BT11`, assessing the importance of diatom load and fetus length for females captured in the West. Models show main effects excluding fetus length and diatom score (*BT11West*), the addition to *BT11West* of a smooth in fetus length (*BT11West+Fetus*) or diatom score as a 5-level categorical factor (*BT11West+Diatom*), the addition of both diatom score and fetus length (*BT11West+Fetus+Diatom*), and finally a model that allows separate smooths in most covariates (including fetus length) for each diatom score category (*BT11West, by=Diatom*).

	BT11West	BT11West+Fetus	BT11West+Diatom	BT11West+Fetus+Diatom	BT11West, by=Diatom
EDF: ti(BLm)	0.9871 (11.00)***	0.9753 (11.00)***	0.9854 (11.00)***	0.9778 (11.00)***	
EDF: ti(Age)	0.8433 (11.00)	0.9423 (11.00)	0.5939 (11.00)	0.8017 (11.00)	
EDF: ti(YearNum2)	0.9546 (6.00)***	1.0715 (6.00)***	0.9598 (6.00)***	0.9495 (6.00)***	
EDF: ti(DateNum)	2.7654 (11.00)***	3.5300 (11.00)***	2.2928 (11.00)***	3.1690 (11.00)***	
EDF: ti(LongNum):Icefar	0.0003 (19.00)	0.0001 (19.00)	0.0003 (19.00)	0.0001 (19.00)	0.0001 (19.00)
EDF: ti(LongNum):Icenear	0.0002 (19.00)	0.0057 (19.00)	0.5083 (19.00)	0.0058 (19.00)	0.0006 (19.00)
EDF: ti(BWt)	2.0219 (11.00)***	1.7293 (11.00)***	2.2462 (11.00)***	1.9570 (11.00)***	1.9157 (11.00)***
EDF: ti(FetusLength)		2.2193 (11.00)***		2.6394 (11.00)***	
DiatomF1			0.2596 (0.11)*	0.1860 (0.11)	0.2269 (0.12)
DiatomF2			0.3239 (0.06)***	0.0574 (0.07)	0.0419 (0.08)
DiatomF3			0.4572 (0.14)**	−0.0144 (0.15)	−0.0386 (0.25)
DiatomF4			1.2330 (0.22)***	0.7097 (0.23)**	1.2465 (0.23)***
EDF: ti(BLm):DiatomF0					0.9643 (11.00)***
EDF: ti(BLm):DiatomF1					0.9237 (10.00)***
EDF: ti(BLm):DiatomF2					0.9651 (11.00)***
EDF: ti(BLm):DiatomF3					2.3036 (10.00)**
EDF: ti(BLm):DiatomF4					0.7920 (7.00)*
EDF: ti(Age):DiatomF0					1.0048 (11.00)
EDF: ti(Age):DiatomF1					0.0004 (10.00)
EDF: ti(Age):DiatomF2					0.1996 (11.00)
EDF: ti(Age):DiatomF3					0.0000 (10.00)
EDF: ti(Age):DiatomF4					0.0000 (10.00)
EDF: ti(YearNum2):DiatomF0					1.9195 (6.00)***
EDF: ti(YearNum2):DiatomF1					0.9134 (6.00)
EDF: ti(YearNum2):DiatomF2					1.1942 (6.00)
EDF: ti(YearNum2):DiatomF3					2.7251 (6.00)*
EDF: ti(YearNum2):DiatomF4					0.0001 (2.00)
EDF: ti(DateNum):DiatomF0					4.3234 (11.00)***
EDF: ti(DateNum):DiatomF1					0.9484 (11.00)***
EDF: ti(DateNum):DiatomF2					0.7913 (11.00)
EDF: ti(DateNum):DiatomF3					0.7219 (10.00)*
EDF: ti(DateNum):DiatomF4					0.0000 (6.00)
EDF: ti(FetusLength):DiatomF0					0.9309 (10.00)***
EDF: ti(FetusLength):DiatomF1					0.0001 (11.00)
EDF: ti(FetusLength):DiatomF2					3.0275 (11.00)***
EDF: ti(FetusLength):DiatomF3					0.7304 (6.00)*
EDF: ti(FetusLength):DiatomF4					0.5954 (5.00)
AIC	1391.5374	1311.1882	1350.8905	1306.4636	1303.3517
BIC	1442.4602	1379.2118	1420.6753	1392.3214	1489.5949
Log Likelihood	−684.3468	−640.3365	−659.7926	−633.9740	−609.9017
R^2	0.3712	0.4497	0.4140	0.4570	0.4820
Num. obs.	638	638	638	638	638

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

A.2 East Region

Like in the West, plots of fetus length against diatom score in each year in the East show a positive, increasing relationship (Figure 12), with Pearson correlations in the range 0.5-0.75 and nonparametric Kendall's Tau in the range 0.3-0.5. Again, the over-representation of diatom scores 0 and 2 is clearly evident in most years.

A.2.1 Body Weight

Results from models of body weight for females sampled in the East are summarised in Table 12. These results are consistent with those for `BWt` from the West, in that they show there is no advantage to be gained by including diatom score in models that already contain fetus length.

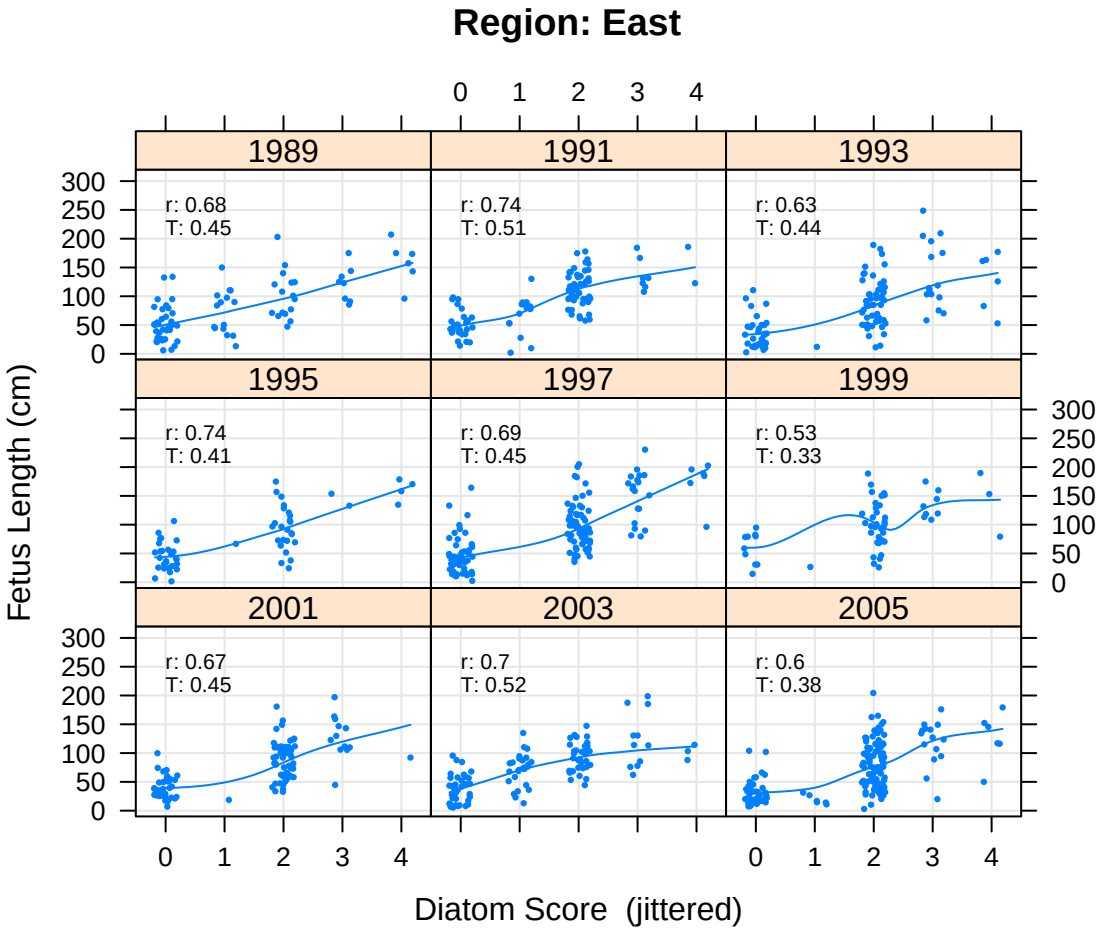

Figure 12: Relationship between fetus length and diatom score (0-4 scale) for each sample year in the East region. Pearson correlations (r) and Kendall's tau (T) are displayed with each panel, and fitted lines are loess smooths with span=2/3.

Table 12: Summary of additive model fits to `BWt`, assessing the importance of diatom load and fetus length for females captured in the East. Models show main effects excluding fetus length and diatom score (*BWtEast*), the addition to *BWtEast* of a smooth in fetus length (*BWtEast+Fetus*) or diatom score as a 5-level categorical factor (*BWtEast+Diatom*), the addition of both diatom score and fetus length (*BWtEast+Fetus+Diatom*), and finally a model that allows separate smooths in most covariates (including fetus length) for each diatom score category (*BWtEast, by=Diatom*).

	BWtEast	BWtEast+Fetus	BWtEast+Diatom	BWtEast+Fetus+Diatom	BWtEast, by=Daitom
EDF: ti(BLm)	1.0084 (11.00)***	1.2784 (11.00)***	0.9988 (11.00)***	1.0211 (11.00)***	
EDF: ti(Age)	0.9834 (11.00)***	0.9768 (11.00)***	1.0777 (11.00)***	0.9761 (11.00)***	0.9779 (11.00)***
EDF: ti(YearNum2)	4.1225 (6.00)***	4.0429 (6.00)***	3.8465 (6.00)***	3.9230 (6.00)***	
EDF: ti(DateNum)	0.9912 (11.00)***	0.9343 (11.00)***	0.9878 (11.00)***	0.9409 (11.00)***	
EDF: ti(LongNum):Icefar	0.0755 (19.00)	0.7063 (18.00)	0.0006 (19.00)	0.6912 (17.00)	0.9033 (14.00)**
EDF: ti(LongNum):Icenear	4.1394 (19.00)***	3.9698 (19.00)***	4.5086 (19.00)***	4.0747 (19.00)***	4.2570 (19.00)***
EDF: ti(FetusLength)		3.3301 (11.00)***		3.4077 (11.00)***	
DiatomF1			0.0825 (0.10)	−0.0359 (0.09)	0.1020 (0.13)
DiatomF2			0.3447 (0.05)***	0.0710 (0.06)	0.1073 (0.07)
DiatomF3			0.5765 (0.08)***	0.0516 (0.10)	0.3487 (0.14)*
DiatomF4			0.8956 (0.12)***	0.2973 (0.13)*	0.6683 (0.18)***
EDF: ti(BLm):DiatomF0					0.9953 (11.00)***
EDF: ti(BLm):DiatomF1					0.9907 (11.00)***
EDF: ti(BLm):DiatomF2					2.4870 (11.00)***
EDF: ti(BLm):DiatomF3					0.9929 (11.00)***
EDF: ti(BLm):DiatomF4					1.4779 (11.00)***
EDF: ti(YearNum2):DiatomF0					1.2097 (6.00)
EDF: ti(YearNum2):DiatomF1					0.0000 (6.00)
EDF: ti(YearNum2):DiatomF2					2.7317 (6.00)*
EDF: ti(YearNum2):DiatomF3					3.9636 (6.00)***
EDF: ti(YearNum2):DiatomF4					0.0002 (6.00)
EDF: ti(DateNum):DiatomF0					0.8300 (11.00)
EDF: ti(DateNum):DiatomF1					0.0546 (11.00)
EDF: ti(DateNum):DiatomF2					0.9239 (11.00)***
EDF: ti(DateNum):DiatomF3					0.4324 (11.00)
EDF: ti(DateNum):DiatomF4					1.6619 (11.00)
EDF: ti(FetusLength):DiatomF0					2.6583 (11.00)***
EDF: ti(FetusLength):DiatomF1					2.0081 (11.00)***
EDF: ti(FetusLength):DiatomF2					0.9859 (11.00)***
EDF: ti(FetusLength):DiatomF3					3.4858 (10.00)**
EDF: ti(FetusLength):DiatomF4					1.4750 (10.00)*
AIC	2035.9793	1861.1568	1954.2260	1863.0161	1864.3438
BIC	2109.1597	1959.1791	2048.1977	1978.7876	2106.5664
Log Likelihood	−1002.8276	−910.2694	−957.6433	−907.5216	−881.9864
R^2	0.5881	0.6616	0.6248	0.6621	0.6729
Num. obs.	922	922	922	922	922

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

A.2.2 Blubber Thickness

Results from models of blubber thickness (`BT11`) for females sampled in the East are summarised in Table 13. A model including terms for both fetus length and diatom score (`BT11East+Fetus+Diatom`) offered negligible improvement over the model containing only fetus length (`BT11East+Fetus+Diatom`). The most complicated model, which has separate smooths across most covariates for each level of diatom score, showed virtually identical values of AIC but a poorer fit based on BIC. These results are consistent with those for `BT11` from the West, in that they show there is no advantage to be gained by including diatom score in models that already contain fetus length.

Table 13: Summary of additive model fits to `BT11`, assessing the importance of diatom load and fetus length for females captured in the East. Models show main effects excluding fetus length and diatom score (*BT11East*), the addition to *BT11East* of a smooth in fetus length (*BT11East+Fetus*) or diatom score as a 5-level categorical factor (*BT11East+Diatom*), the addition of both diatom score and fetus length (*BT11East+Fetus+Diatom*), and finally a model that allows separate smooths in most covariates (including fetus length) for each diatom score category (*BT11East, by=Diatom*).

	BT11East	BT11East+Fetus	BT11East+Diatom	BT11East+Fetus+Diatom	BT11East, by=Diatom
EDF: ti(BLm)	0.9937 (11.00)***	0.9927 (11.00)***	0.9903 (11.00)***	0.9880 (11.00)***	
EDF: ti(Age)	1.5457 (11.00)**	2.1694 (11.00)***	1.5488 (11.00)**	2.0495 (11.00)***	
EDF: ti(YearNum2)	3.7373 (6.00)***	3.7472 (6.00)***	4.0994 (6.00)***	3.8941 (6.00)***	
EDF: ti(DateNum)	4.5466 (11.00)***	4.1089 (11.00)***	4.4164 (11.00)***	4.1459 (11.00)***	
EDF: ti(LongNum):Icefar	0.8482 (15.00)**	1.3850 (19.00)*	0.8667 (16.00)**	1.3615 (19.00)*	1.6382 (19.00)**
EDF: ti(LongNum):Icenear	0.8820 (16.00)**	0.8799 (18.00)**	0.8975 (19.00)**	0.8842 (17.00)**	0.9113 (15.00)***
EDF: ti(BWt)	0.9976 (11.00)***	0.9953 (11.00)***	0.9967 (11.00)***	0.9952 (11.00)***	0.9952 (11.00)***
EDF: ti(FetusLength)		3.2932 (11.00)***		3.0806 (11.00)***	
DiatomF1			0.1846 (0.10)	0.0595 (0.10)	0.1244 (0.12)
DiatomF2			0.4081 (0.06)***	0.1376 (0.07)*	0.1913 (0.08)*
DiatomF3			0.6016 (0.09)***	0.1886 (0.10)	0.3178 (0.14)*
DiatomF4			0.5888 (0.13)***	0.1713 (0.14)	0.5780 (0.14)***
EDF: ti(BLm):DiatomF0					0.9702 (11.00)***
EDF: ti(BLm):DiatomF1					0.8312 (10.00)*
EDF: ti(BLm):DiatomF2					5.2352 (11.00)***
EDF: ti(BLm):DiatomF3					0.9478 (11.00)***
EDF: ti(BLm):DiatomF4					0.9053 (11.00)**
EDF: ti(Age):DiatomF0					1.1894 (11.00)*
EDF: ti(Age):DiatomF1					1.1086 (9.00)
EDF: ti(Age):DiatomF2					1.5818 (11.00)**
EDF: ti(Age):DiatomF3					0.0001 (11.00)
EDF: ti(Age):DiatomF4					0.7447 (11.00)
EDF: ti(YearNum2):DiatomF0					0.9069 (6.00)**
EDF: ti(YearNum2):DiatomF1					0.0002 (6.00)
EDF: ti(YearNum2):DiatomF2					2.9197 (6.00)***
EDF: ti(YearNum2):DiatomF3					0.3205 (6.00)
EDF: ti(YearNum2):DiatomF4					0.8648 (6.00)
EDF: ti(DateNum):DiatomF0					0.9328 (11.00)***
EDF: ti(DateNum):DiatomF1					0.9054 (11.00)
EDF: ti(DateNum):DiatomF2					3.4455 (11.00)***
EDF: ti(DateNum):DiatomF3					3.2356 (11.00)**
EDF: ti(DateNum):DiatomF4					0.7449 (9.00)*
EDF: ti(FetusLength):DiatomF0					0.8928 (11.00)**
EDF: ti(FetusLength):DiatomF1					1.7382 (11.00)***
EDF: ti(FetusLength):DiatomF2					2.3108 (11.00)***
EDF: ti(FetusLength):DiatomF3					1.6022 (10.00)*
EDF: ti(FetusLength):DiatomF4					0.0000 (9.00)
AIC	2090.2732	1974.3757	2028.9085	1977.0992	1974.6618
BIC	2179.2221	2090.3014	2137.3768	2110.7910	2236.9723
Log Likelihood	−1026.7075	−963.1695	−991.9810	−960.8504	−932.9835
R^2	0.4191	0.4916	0.4587	0.4920	0.5107
Num. obs.	922	922	922	922	922

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

B Female models

B.1 West - Total weight

There are 638 female samples from the West. Models using body weight as the response are summarised in Table 14, and these results show that interactions are only marginally important (`M1` vs. `M2`) except for a modest interaction between `YearNum` and `DateNum`.

We obtain predictions for set values of the covariates included in model `FW.BWt.M2`. Deciding the set values of covariates over which to form predictions is critical, and to help us we plot length against age, conditioned on quintiles of DateNum (rows) and longitude (columns) (Figure 13). We additionally colour points according to quintiles of fetus length. As we have to predict over many covariates, the graph provides reassurance that we are not attempting to obtain predictions from an area of the model space where there are few or no data. In our case, we obtain predictions over year for median length, median age animals which were either: a) captured near the beginning of the season and were carrying a small fetus; or, b) captured near the end of the season and were carrying a large fetus.

To obtain predictions for animals that are relatively newly arrived to the Antarctic, we choose the 20th percentiles of DateNum and fetus length for prediction. To obtain predictions for animals that have been feeding for most of the summer months, we choose the 80th percentile of these same variables. These percentiles were chosen to be close to the ends of the distributions of the variables in question, but sufficiently away from the boundaries to reduce the likelihood of edge effects when obtaining predictions. The values of the various percentiles used for prediction are provided in Table 8 in Section 3.1. In order to help compare results between West and East, we had hoped to fix one set of values and use them for prediction across both regions. While this was possible, it turned out to be not very sensible given the large differences in values of the percentiles of `DateNum` between the regions. Our preference was therefore to allow the values for covariates used in prediction to vary between the regions, as detailed in Table 8.

We also need to condition predictions on a fixed value of longitude, something that is not so straightforward for the JARPA data given that not all areas are sampled over all dates. While the effect of longitude on models is often not substantial, and so the exact longitudes chosen for prediction are not critical, we nonetheless prefer to obtain predictions from parts of the model space that have been estimated with reasonable amounts of data. Our plot of the covariates can help us to choose a longitude that has reasonable sampling coverage for the key dates of interest to us, the 20th and 80th percentiles of DateNum. On this basis, we choose longitude 73°E as having a sufficient density of females carrying small and large fetuses at the beginning and ends of the season (Figure 13).

The predicted long-term trend in body weight for female animals that are relatively newly arrived in the West shows fairly consistent arrival weights except for two years, 1998 and 2000, when weights appear depressed (Figure 14). We would caution against ascribing too much importance to these reductions in mean weight given the estimated sampling variability. To us, these results show that there is little variation around the mean of the series (red dashed line) year-to-year, with the possible exception of 2000. We note that the 95% credible interval for the mean prediction contains the overall series mean throughout the entire period.

Turning to late season predictions for females from the West, Figure 15 indicates a pronounced decline between 1992 and 2000. The overall mean is almost contained within the 95% credible interval for mean predictions, with only 1992 and 2000 falling slightly outside. In considering how important this apparent decline might be, we note that the shape of predictions from later in the season is quite similar to shape of predictions from earlier in the season (Figure 14); it is simply a little more pronounced. Taking the difference between late and early season predictions helps to interpret these trends.

The year trend for the within-season accumulation of body weight shows no virtually trend for females captured in the West (Figure 16). The most we are willing to infer from this result is that there may have been better than usual feeding conditions in 1990/1992.

Residual diagnostics for these models are presented in Appendix D.1 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not included

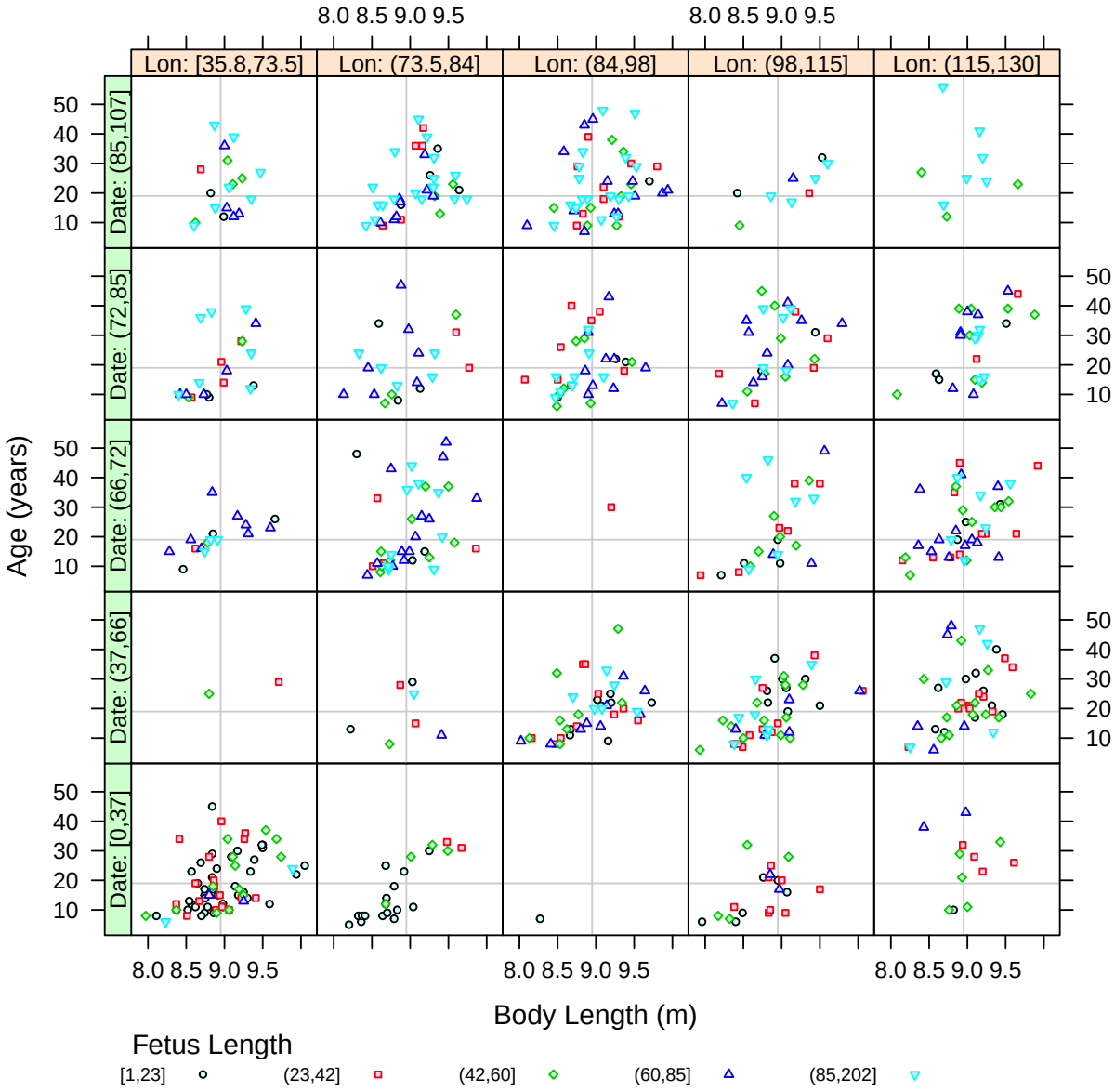

Figure 13: Spread of data across the design space for females from the West, showing scatterplot of length against age, conditioned on quintiles of DateNum and longitude, and showing quintiles of fetus length using symbol type. Grey reference lines show median length and age.

Table 14: Summary of GAM fits for response BWt for females from the West region.

	FW.BWt.M1	FW.BWt.M2
Icenear	−0.0775 (0.10)	−0.0464 (0.10)
DiatomF2High	0.0469 (0.07)	0.0824 (0.08)
EDF: ti(BLm)	2.8984 (7.00)***	2.8592 (7.00)***
EDF: ti(Age)	0.9763 (7.00)***	0.9783 (7.00)***
EDF: ti(YearNum2)	5.1136 (6.00)***	4.8233 (6.00)***
EDF: ti(DateNum)	0.4665 (7.00)	0.8145 (7.00)
EDF: ti(LongNum):Icefar	1.3371 (19.00)	0.7055 (19.00)
EDF: ti(LongNum):Icenear	1.6591 (19.00)**	0.9537 (19.00)**
EDF: ti(FetusLength)	1.5955 (7.00)***	1.8492 (7.00)***
EDF: ti(YearNum2,DateNum)		7.4043 (42.00)**
EDF: ti(YearNum2,FetusLength)		3.6561 (42.00)
EDF: ti(YearNum2,Age)		0.0006 (42.00)
EDF: ti(DateNum,FetusLength)		0.0004 (49.00)
EDF: ti(Age,BLm)		0.0019 (49.00)
EDF: ti(YearNum2,BLm)		0.0015 (49.00)
AIC	1408.5967	1408.6939
BIC	1503.3888	1570.4086
Log Likelihood	−683.0366	−668.0745
R^2	0.6631	0.6732
Num. obs.	638	638

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

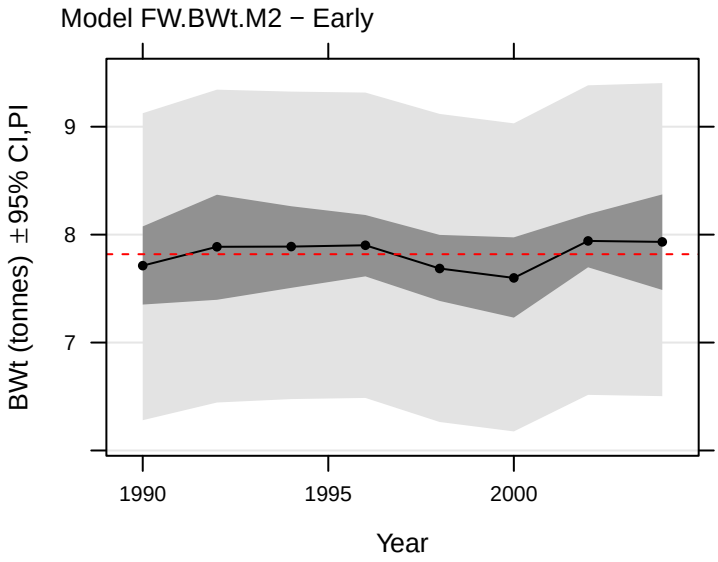

Figure 14: Predicted long-term trend in BWt from model FW.BWt.M2 predictions for median length, median age females with small fetuses (20th percentile on fetus length), that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the West. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

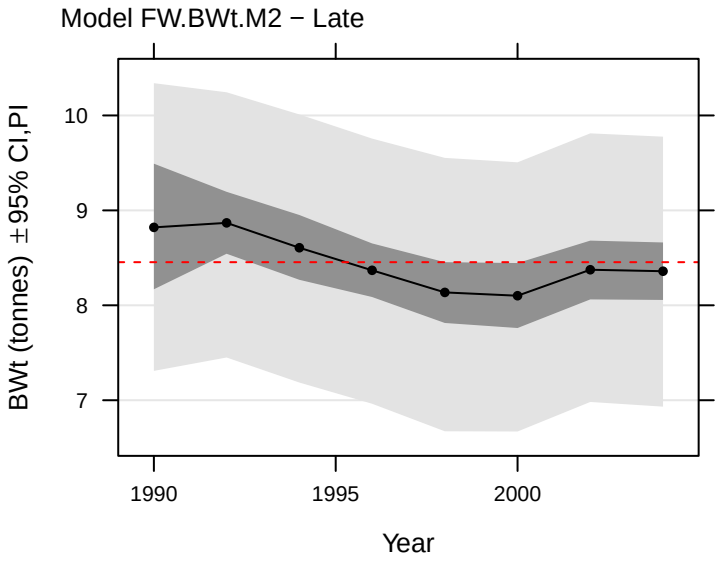

Figure 15: Predicted long-term trend in BWt from model FW.BWt.M2 predictions for median length, median age females with large fetuses (80th percentile on fetus length), that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the West. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

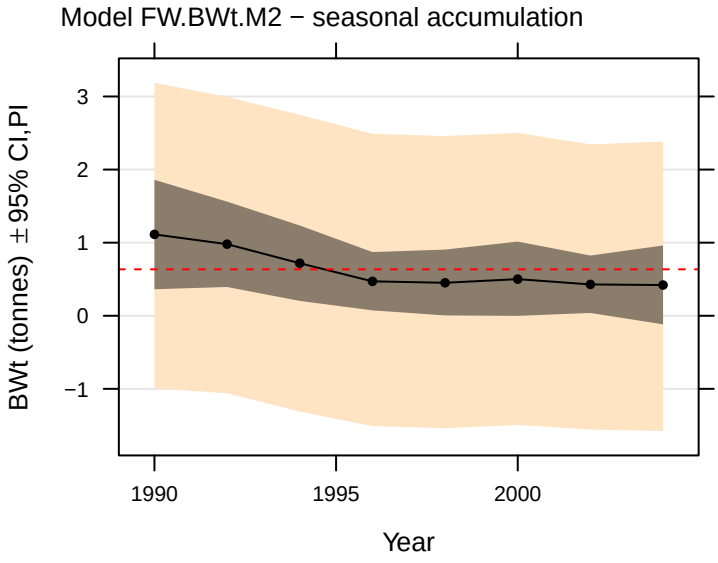

Figure 16: Predicted long-term trend in the yearly accumulation of BWt between the 20th and 80th percentiles of DateNum from model FW.BWt.M2 for median length, median age females captured near the ice edge in the West. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics proved satisfactory.

Table 15: Summary of GAM fits for response BT11 for females from the West region.

	FW.BT11.M1	FW.BT11.M2	FW.BT11.M3
Icenear	0.1355 (0.10)	0.1915 (0.10)	0.1986 (0.09)*
DiatomF2High	0.0673 (0.07)	0.0866 (0.07)	0.0476 (0.07)
EDF: ti(BLm)	0.6700 (7.00)	0.5276 (7.00)	0.9772 (7.00)***
EDF: ti(Age)	1.7292 (7.00)*	1.5900 (7.00)*	1.0714 (7.00)
EDF: ti(YearNum2)	2.0724 (6.00)***	3.9347 (6.00)***	4.7115 (6.00)***
EDF: ti(DateNum)	2.2096 (7.00)***	3.0009 (7.00)***	2.5705 (7.00)***
EDF: ti(LongNum):Icefar	0.7565 (18.00)*	0.0016 (19.00)	0.0004 (19.00)
EDF: ti(LongNum):Icenear	1.6459 (19.00)	0.0002 (19.00)	0.0002 (19.00)
EDF: ti(FetusLength)	0.9934 (7.00)***	0.9911 (7.00)***	0.9845 (7.00)***
EDF: ti(YearNum2,DateNum)		6.0825 (42.00)***	0.7670 (42.00)*
EDF: ti(YearNum2,FetusLength)		0.0003 (42.00)	0.0005 (42.00)
EDF: ti(YearNum2,Age)		0.0001 (42.00)	0.0024 (42.00)
EDF: ti(DateNum,FetusLength)		1.0069 (49.00)	3.8218 (49.00)*
EDF: ti(Age,BLm)		0.0004 (49.00)	0.0009 (49.00)
EDF: ti(YearNum2,BLm)		0.0002 (42.00)	0.0005 (42.00)
EDF: ti(BWt)			2.1141 (7.00)***
EDF: ti(BWt,FetusLength)			0.0003 (49.00)
AIC	1420.2312	1410.8583	1299.9911
BIC	1495.5869	1529.4831	1414.7197
Log Likelihood	−693.2134	−678.8217	−624.2620
R^2	0.3488	0.3704	0.4695
Num. obs.	638	638	638

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

B.2 West - Blubber thickness

Additive models with blubber thickness as the response are summarised in Table 15. These show that models of `BT11` without body weight as a covariate account for around 37% of the variation in the data (`M2`), while the addition of body weight lifts this value substantially to 47% (`M3`). As we found in models of body weight, shrinkage has effectively eliminated spatial terms from models of `BT11`.

We obtain predictions for set values of the covariates for models `FW.BT11.M2` (without `BWt` as covariate) and `FW.BT11.M3` (with `BWt` as a covariate), using the values detailed in Table 8. Our assessment of the covariate space undertaken in relation to body weight models applies here, so we are satisfied that there are sufficient data around our set values of the covariates used for prediction. Set yearly values of early- and late-season body weight, used for obtaining predictions from model `FW.BT11.M3`, are obtained from our body weight model for these data, namely `FW.BWt.M2`. We again condition on the same longitude as determined for models of body weight, and obtain predictions near to the ice edge. Predictions for animals that are relatively newly arrived to the Antarctic are obtained by using the 20th percentiles of DateNum and fetus length for prediction, while predictions for animals that have been feeding for some time are obtained by using the 80th percentile of these same variables.

The predicted long-term trend in BT11 for female animals that are relatively newly arrived in the West shows little change over time (Figure 17), irrespective of whether body weight is included in models. These results indicate that animals are arriving in about the same condition, 3-3.5 cm average `BT11`, at each biennial sampling point.

Late season predictions for females from the West, without using body weight as a covariate, indicate a slight and almost monotonic decrease over the period of sampling (Figure 18a). Additionally conditioning

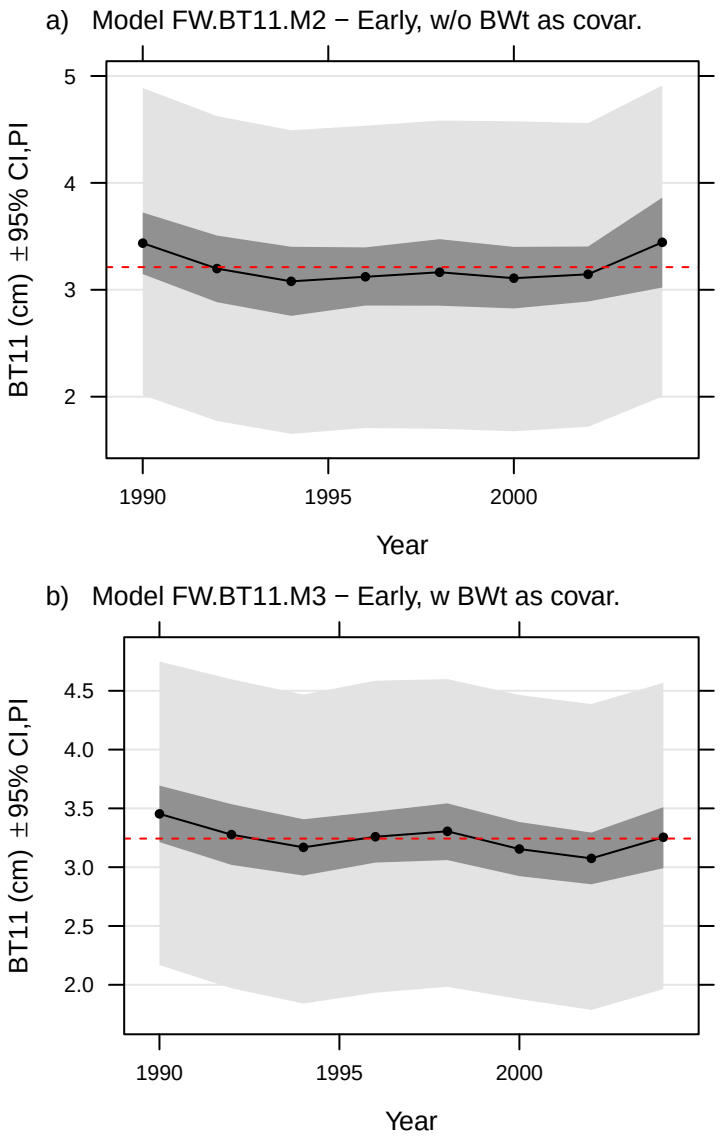

Figure 17: Predicted long-term trend in BT11 from: a) model FW.BT11.M2 predictions for median length, median age females with small fetuses (20th percentile on fetus length), that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the West; and, b) model FW.BT11.M3 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

predictions on body weight again shows an overall negative trend (Figure 18b).

The seasonal trend in `BT11` accumulation between the 20th and 80th percentiles of `DateNum` reveals a pronounced linear decline, irrespective of the inclusion of BWt as a covariate in models (Figure 19). We note that the overall mean of the predictions (red dotted line) is wholly contained within the 95% CI, however this should not be interpreted as non-significant in any traditional sense; it merely gives an indication of the variability in relation to the overall mean. What is important here is not significance, which is largely irrelevant when considering predictions, but rather the effect size and how precisely it is estimated. Here we see accumulated growth in blubber thickness between day 37 and 85 drop from around 1.58 cm in 1990 to 0.61 cm in 2004 (without conditioning on weight), or from 1.41 cm to around 0.85 cm if body weight is included as a covariate. Clearly, the decision about whether to include body weight in models of blubber thickness is critical to our understanding of trends in condition.

When considering this result it is important to recognise that the total estimated decline in accumulated `BT11` is not appreciably different in magnitude from the biennial variation in mean accumulation evident in other parts of the data. We note with interest the comparison between this result for accumulated `BT11` and the predictions of the seasonal accumulation of body weight, where the latter shows no appreciable trend other than some evidence for two higher than average years at the beginning of the series (Figure 16).

Standard diagnostics for these models are presented in Appendix D.2 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not included for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics were satisfactory except for some slight deviation in distributional assumptions evident in Q-Q plots (Appendix D.2, Figures 48 & 48). However, we note that only a handful of points in the tails of the distribution fell outside the Q-Q plot reference bands, and, given the sample sizes involved and the low number of points not conforming to our assumed Gaussian error distribution, we confirm these diagnostics to be satisfactory. As an aside, we note that this pattern is repeated across many of the models of `BT11` — some slight departures from distributional assumptions in the tails of distributions of residuals. This diagnostic aspect in models of `BT11` seems absent for models of `BWt`, in which case distributional assumptions (also Gaussian) are consistently well met.

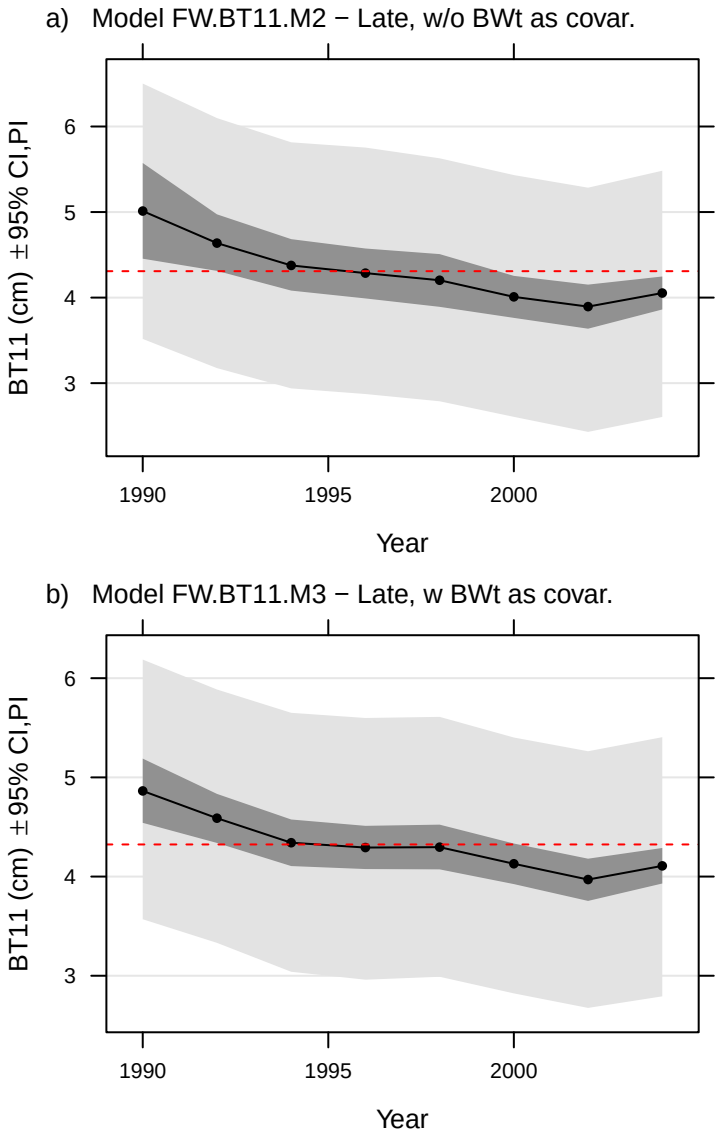

Figure 18: Predicted long-term trend in BT11 from: a) model FW.BT11.M2 predictions for median length, median age females with large fetuses (80th percentile on fetus length), that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the West; and, b) model FW.BT11.M3 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

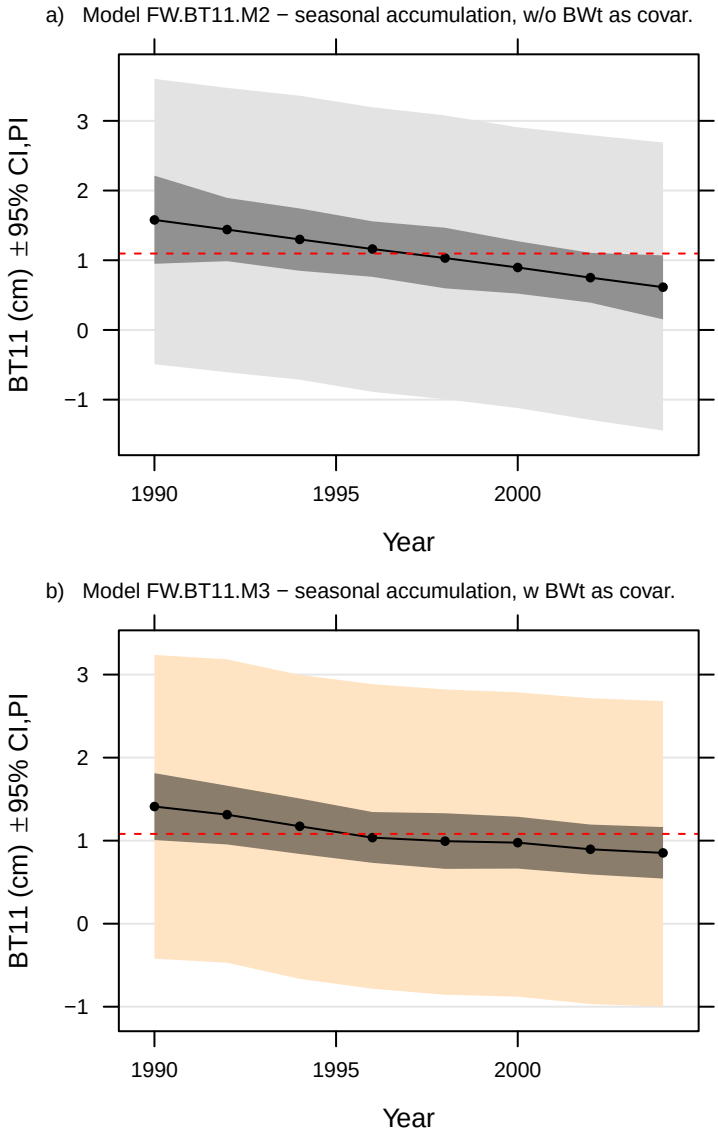

Figure 19: Predicted long-term trend in the yearly accumulation of BT11 between the 20th and 80th percentiles of DateNum from: a) model FW.BT11.M2 for median length, median age females captured near the ice edge in the West; and, b) the same calculation based on model FW.BT11.M3 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

Table 16: Summary of GAM fits for response BWt for females from the East region.

	FE.BWt.M1	FE.BWt.M2
Icenear	−0.0610 (0.08)	−0.0687 (0.08)
DiatomF2High	0.0845 (0.06)	0.0764 (0.06)
EDF: ti(BLm)	1.2265 (7.00)***	1.4664 (7.00)***
EDF: ti(Age)	0.9774 (7.00)***	0.9779 (7.00)***
EDF: ti(YearNum2)	3.9883 (6.00)***	3.7335 (6.00)***
EDF: ti(DateNum)	0.9391 (7.00)***	0.9585 (7.00)***
EDF: ti(LongNum):Icefar	0.7370 (19.00)	0.7126 (17.00)
EDF: ti(LongNum):Icenear	4.1709 (19.00)***	4.1717 (19.00)***
EDF: ti(FetusLength)	3.2177 (7.00)***	2.6049 (7.00)***
EDF: ti(YearNum2,DateNum)		0.3534 (42.00)
EDF: ti(YearNum2,FetusLength)		10.3288 (42.00)***
EDF: ti(YearNum2,Age)		0.0007 (42.00)
EDF: ti(DateNum,FetusLength)		1.0762 (49.00)*
EDF: ti(Age,BLm)		0.0109 (49.00)
EDF: ti(YearNum2,BLm)		0.2100 (42.00)
AIC	1862.0581	1846.8291
BIC	1969.4960	2032.5713
Log Likelihood	−908.7692	−884.9311
R^2	0.6619	0.6749
Num. obs.	922	922

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

B.3 East - Total weight

There are 922 female samples from the East. Models using body weight as the response are summarised in Table 16, and indicate that interactions are only marginally important over and above a main effects model (`M1` vs. `M2`).

We obtain predictions for set values of the covariates included in model `FE.BWt.M2`. As a check of the covariate values to predict over, we plot length against age, conditioned on quintiles of DateNum (rows) and longitude (columns), and additionally colour points according to quintiles of fetus length (Figure 20). We obtain predictions over year for median length, median age animals which are either: a) captured near the beginning of the season and were carrying a small fetus; or, b) captured near the end of the season and were carrying a large fetus. We also condition predictions on longitude 171°W, near to the ice, in the Ross Sea. A check of the JARPA catch position presented in de la Mare et al. (2014) confirmed catches were taken near to the ice at this longitude in most seasons.

To obtain predictions for animals that are relatively newly arrived to the Antarctic, we choose the 20th percentiles of DateNum and fetus length for prediction, while for late season predictions we choose the 80th percentile of these same variables. The values of the various percentiles used for prediction are provided in Table 8 in Section 3.1. As with our analyses of animals from the West, for these analyses we only allowed the fixed values for covariates used in prediction to be determined using data from the East.

The predicted long-term trend in body weight for female animals that are relatively newly arrived in the East shows a slight increase in the early 1990's, followed by a slight decrease around 2000 (Figure 21). We note that the overall series mean is wholly contained within the 95% CI for the trend, which if used a rough heuristic indicates that this trend is indistinguishable from the overall mean in light of sampling variability.

Late season predictions for females from the East indicates a once-off decrease in body weight of sampled animals occurred in 1997, but, aside from this aberration, average weights were reasonably consistent for the

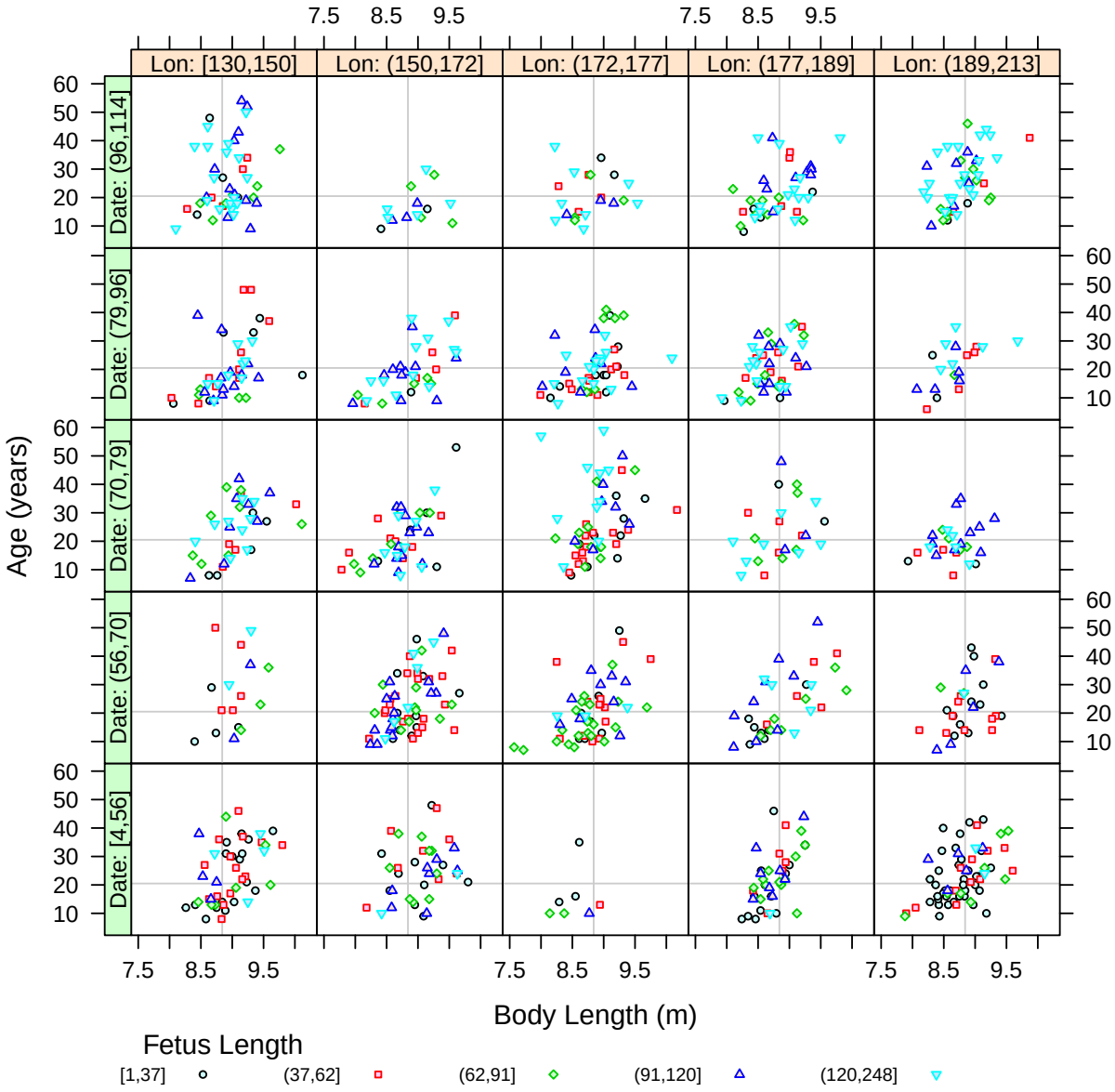

Figure 20: Spread of data across the design space for females from the East, showing scatterplot of length against age, conditioned on quintiles of DateNum and longitude, and showing quintiles of fetus length by symbol type. Grey reference lines show median length and age.

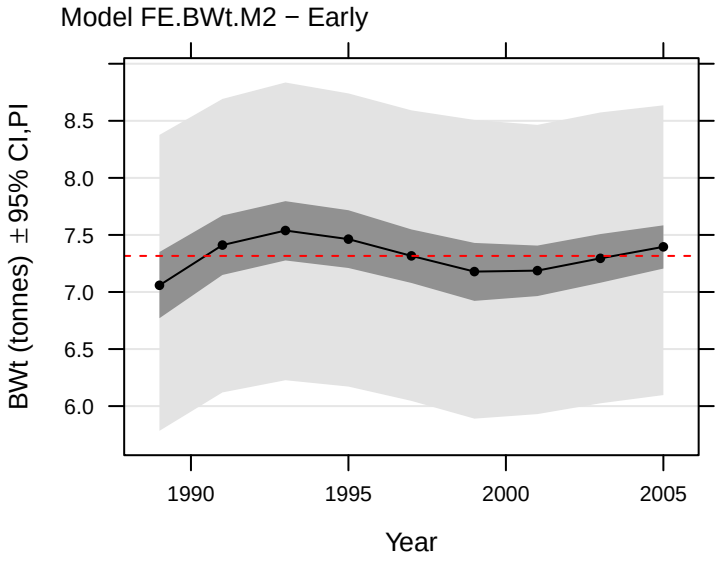

Figure 21: Predicted long-term trend in BWt from model FE.BWt.M2 predictions for median length, median age females with small fetuses (20th percentile on fetus length), that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the West. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

rest of the sampling period (Figure 22).

The year trend for the within-season accumulation of body weight shows a slight decline over the first five years of sampling, followed by a recovery in the early 2000's (Figure 23). While the consistency of the predicted pattern in each half of the sampling period could be indicative of decreasing, then improving, conditions, given the small number of time points and estimated variability this pattern could be equally due to sampling heterogeneity.

Residual diagnostics for these models are presented in Appendix D.3 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not appended for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). These checks prompted us to use a slightly larger number of basis dimensions for the smooths in longitude, but otherwise all diagnostics looked satisfactory.

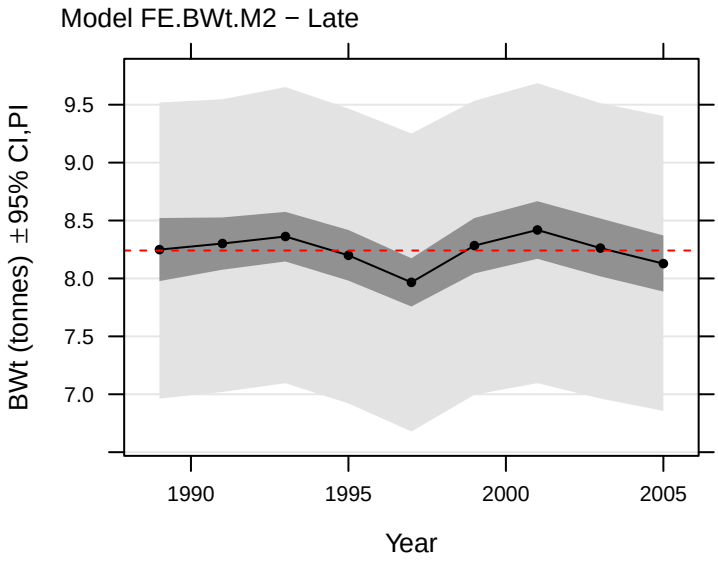

Figure 22: Predicted long-term trend in BWt from model FE.BWt.M2 predictions for median length, median age females with large fetuses (80th percentile on fetus length), that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the East. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

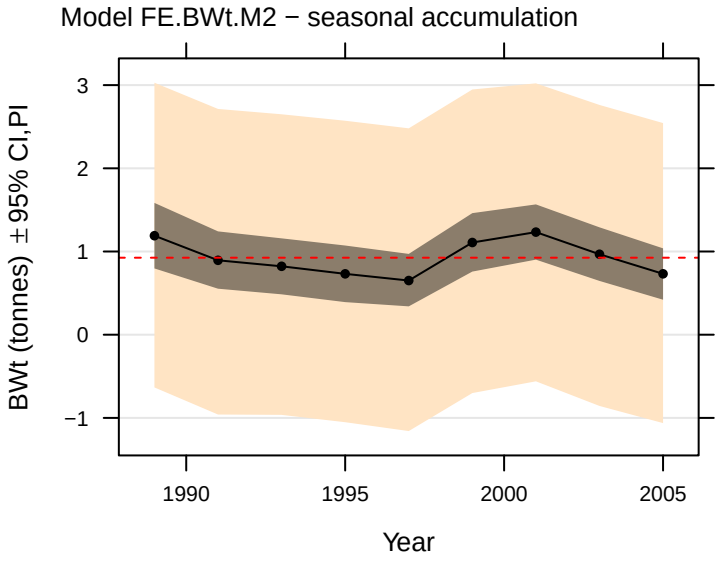

Figure 23: Predicted long-term trend in the accumulation of BWt between the 20th and 80th percentiles of DateNum from model FE.BWt.M2 for median length, median age females captured near to the ice edge in the East. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

Table 17: Summary of GAM fits for response BT11 for females from the East region.

	FE.BT11.M1	FE.BT11.M2	FE.BT11.M3
Icenear	−0.0437 (0.09)	−0.0485 (0.09)	−0.0495 (0.08)
DiatomF2High	0.1706 (0.07)*	0.1628 (0.07)*	0.1299 (0.06)*
EDF: ti(BLm)	1.2475 (7.00)	3.0740 (7.00)*	2.8877 (7.00)***
EDF: ti(Age)	1.9601 (7.00)	0.0002 (7.00)	1.8397 (7.00)***
EDF: ti(YearNum2)	4.2185 (6.00)***	4.0812 (6.00)***	3.8702 (6.00)***
EDF: ti(DateNum)	3.6690 (7.00)***	3.4011 (7.00)***	3.6830 (7.00)***
EDF: ti(LongNum):Icefar	0.8561 (19.00)	0.5899 (19.00)	1.4471 (19.00)*
EDF: ti(LongNum):Icenear	0.0043 (19.00)	0.0011 (19.00)	0.9129 (13.00)***
EDF: ti(FetusLength)	3.5694 (7.00)***	2.6496 (7.00)***	2.8669 (7.00)***
EDF: ti(YearNum2,DateNum)		2.1357 (42.00)	1.6361 (42.00)
EDF: ti(YearNum2,FetusLength)		3.6960 (42.00)**	1.4791 (42.00)
EDF: ti(YearNum2,Age)		3.5726 (42.00)**	3.5751 (42.00)**
EDF: ti(DateNum,FetusLength)		1.7179 (49.00)*	2.7493 (49.00)
EDF: ti(Age,BLm)		7.8910 (49.00)*	3.0054 (49.00)
EDF: ti(YearNum2,BLm)		11.4659 (42.00)***	10.0456 (42.00)***
EDF: ti(BWt)			0.9951 (7.00)***
EDF: ti(BWt,FetusLength)			0.0001 (49.00)
AIC	2153.5721	2129.6251	1956.9938
BIC	2268.9942	2445.2887	2249.2052
Log Likelihood	−1052.8720	−999.4110	−917.9543
R^2	0.3825	0.4320	0.5258
Num. obs.	922	922	922

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

B.4 East - Blubber thickness

Additive models of blubber thickness (`BT11`) for female animals from the East show that several tensor product interactions are important, both for models excluding `BWt` (`M2`) and including `BWt` (`M3`) as a covariate (Table 17). Body weight is again identified as one of the most important predictors of `BT11`. Longitudinal effects are evident only when models control for body weight, in which case partial effects plots indicate an slight increasing effect on `BT11` from West to East (Appendix D.4, Figure 56).

We obtain predictions for set values of the covariates for models `FE.BT11.M2` (without `BWt` as covariate) and `FE.BT11.M3` (with `BWt` as a covariate), using the values detailed in Section 3.1, Table 8. We again predict for median Longitude 174°E, which equates to an area within the Ross Sea, and obtain early and late season predictions by conditioning on the 20th and 80th percentiles of DateNum and fetus length. Set values for conditioning predictions on weight are again determined for early- and late-season time points using our model for body weight, `FE.BWt.M2`.

The predicted long-term trend in BT11 for female animals that are relatively newly arrived in the East shows virtually no change over time in our model without `BWt` as a covariate, but with slight evidence (loss of 0.5 cm `BT11`, 1992-2005) of a downward trend in models accounting for `BWt` (Figure 24a,b).

Late season predictions for females from the East, without using body weight as a covariate, indicate a slight decrease to around 1997, followed by an equally slight recovery in the mid-2000's (Figure 25a). Additionally conditioning predictions on body weight increased the overall mean of the series slightly, but resulted in little change to the general pattern (Figure 25b). We also note that this trend reflects closely the trend in body weight for this same group of animals (Figure 22).

The seasonal trend in the accumulation of `BT11` between the 20th and 80th percentiles of `DateNum` weakly follows the pattern seen in earl- and late-season predictions, but ultimately reveals virtually no trend in light of sampling variation, irrespective of whether predictions are conditioned on body weight or not (Figure 26).

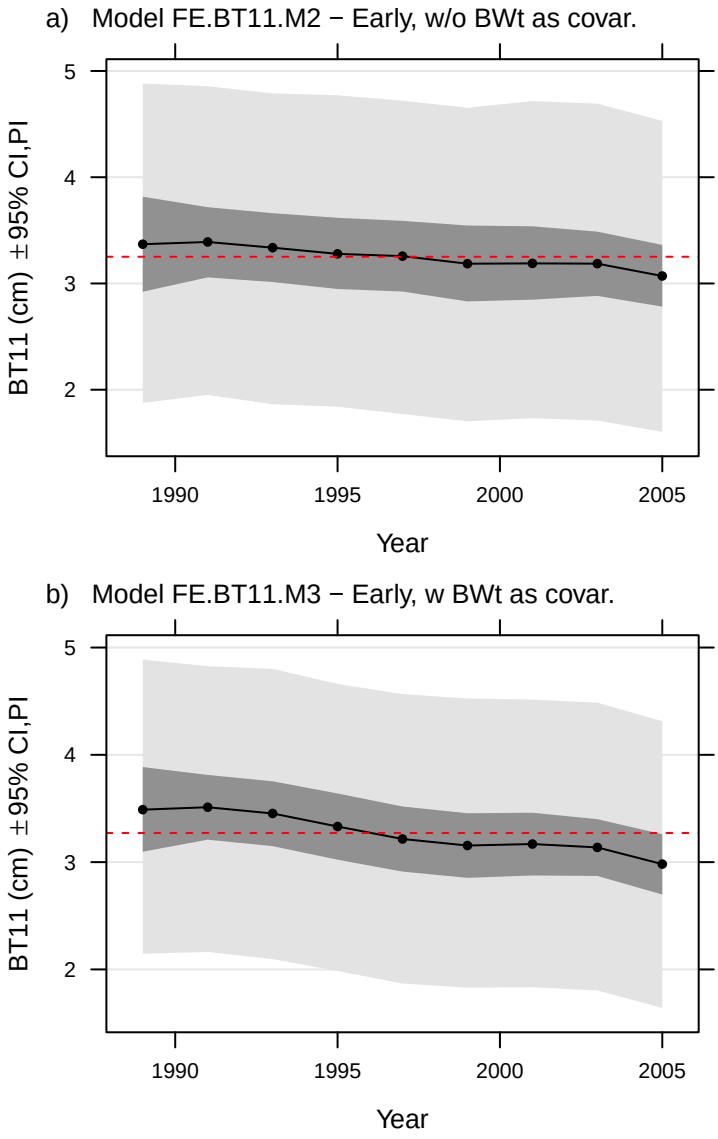

Figure 24: Predicted long-term trend in BT11 from: a) model FE.BT11.M2 predictions for median length, median age females with small fetuses (20th percentile on fetus length), that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the East; and, b) model FE.BT11.M3 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

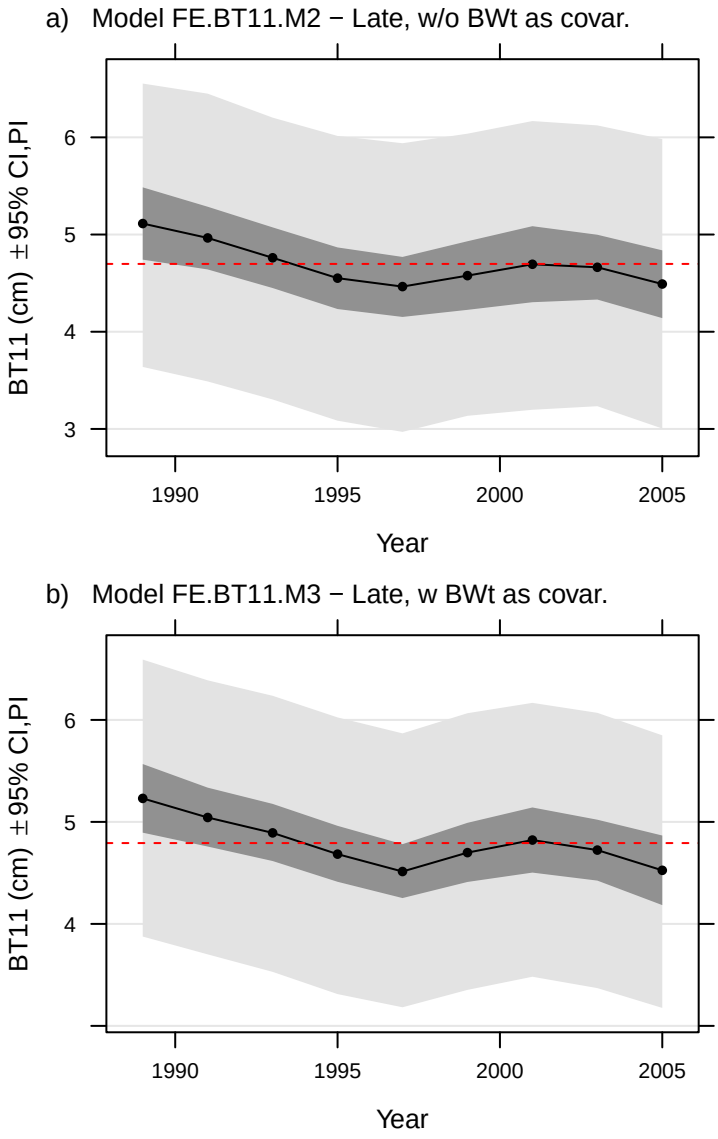

Figure 25: Predicted long-term trend in BT11 from: a) model FE.BT11.M2 predictions for median length, median age females with large fetuses (80th percentile on fetus length), that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the East; and, b) model FE.BT11.M3 predictions for the same covariate set but additionally conditioned on the 80th percentile of body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

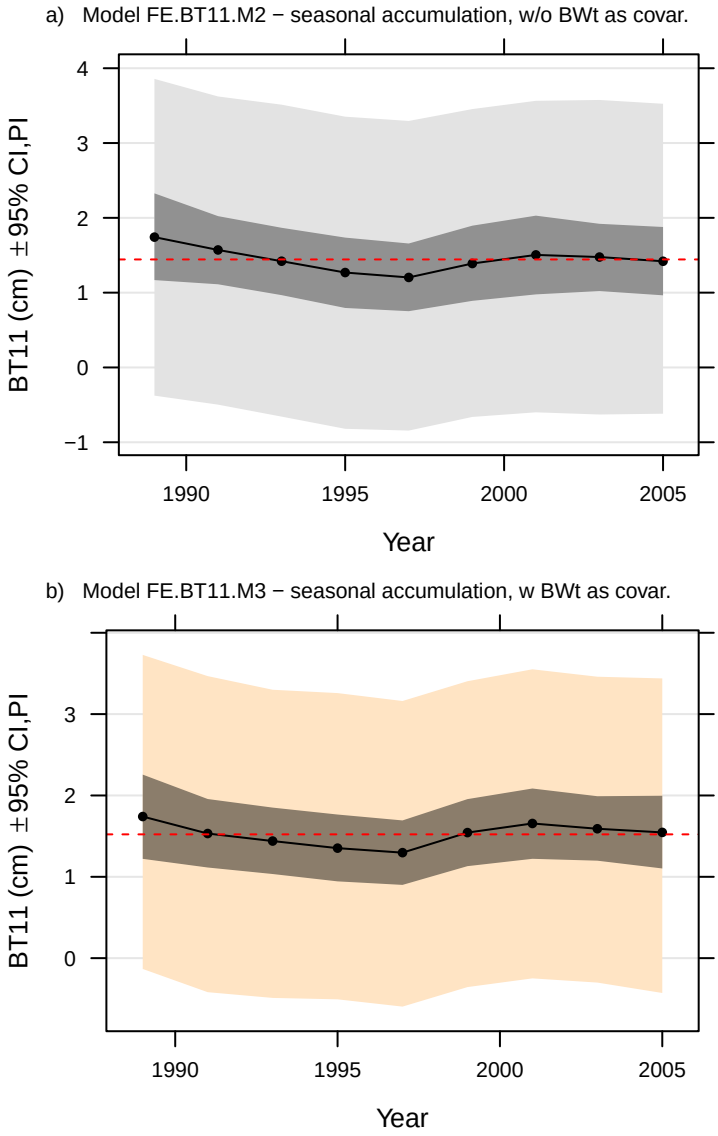

Figure 26: Predicted long-term trend in the yearly accumulation of BT11 between the 20th and 80th percentiles of DateNum from: a) model FE.BT11.M2 for median length, median age females in the East; and, b) the same calculation based on model FE.BT11.M3 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

Standard diagnostics for these models are presented in Appendix D.4 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not appended for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics looked satisfactory.

C Male models

C.1 West - Total weight

There are 1208 male samples from the West.

Models using body weight as the response are summarised in Table 18, and shows that models that allow separate smooths per diatom class (`M2`), or full tensor product interactions by diatom class (`M3`), offer only marginal improvement over a much simpler main effects model (`M1`). Nonetheless, as our goal is the best prediction possible we prefer model `M3`, noting that the impacts of unimportant higher-order terms are down-weighted or eliminated altogether under shrinkage.

We obtain predictions for set values of the covariates for model `MW.BWt.M3` using the values detailed in Section 3.2 Table 9. A plot of the data according to the important covariates confirms there are sufficient data around our fixed values of the covariates used for prediction (Figure 27). This plot also serves to demonstrate how loose the relationship is between date and diatom score; animals scored as low (0,1) and high (2,3,4) appear in all panels, irrespective of time within season or longitude. We predict for longitude 75°E (Prydz Bay), and obtain early and late season prediction by conditioning on the 20th and 80th percentiles of DateNum for Low and High diatom score groups, respectively.

The predicted long-term trend in `BWt` for male animals that are relatively newly arrived in the West shows a gradual increase in body weight in the early 1990's, followed by a gradual decrease to around the the year 2000 (Figure 28). Given the periodicity of sampling, it is difficult to view these results as anything other than a relatively stable arrival weight meandering around a long-term mean.

Late season predictions for males from the West show little variation around the overall mean, indicating that late season weights seem stable over the period of JARPA (Figure 29).

The seasonal trend in the accumulation of `BWt` in the West between the 20th and 80th percentiles of `DateNum` for low and high diatom load animals, respectively, reveals that the addition of body weight over the summer season has remained fairly constant over the period of JARPA (Figure 30). The only interesting deviation from the overall mean occurred in 2004, when accumulated weight appears to have dropped by up to 50%. However, while this unusual change may warrant further investigation, estimated uncertainty around the mean trend also indicates it may simply have arisen due to sampling variability.

Standard diagnostics for these models are presented in Appendix D.5 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not included for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics looked satisfactory.

Table 18: Summary of GAM fits for response BWt for males from the West region.

	MW.BWt.M1	MW.BWt.M2	MW.BWt.M3
Icenear	-0.0896 (0.04)*	-0.0844 (0.04)*	-0.0575 (0.04)
DiatomF2High	0.1969 (0.04)***	0.1940 (0.04)***	0.1811 (0.04)***
EDF: ti(BLm)	3.9802 (7.00)***	3.9319 (7.00)***	3.8245 (7.00)***
EDF: ti(Age)	3.5700 (7.00)***	3.5532 (7.00)***	3.3482 (7.00)***
EDF: ti(YearNum2)	3.3764 (6.00)***		
EDF: ti(DateNum)	2.1042 (7.00)***		
EDF: ti(LongNum):Icefar	1.4867 (7.00)***	1.5091 (7.00)***	1.1762 (7.00)**
EDF: ti(LongNum):Icenear	1.6134 (7.00)***	1.6774 (7.00)***	0.9657 (7.00)**
EDF: ti(YearNum2):DiatomF2Low		2.7497 (6.00)***	2.7716 (6.00)***
EDF: ti(YearNum2):DiatomF2High		2.9236 (6.00)**	0.4963 (6.00)
EDF: ti(DateNum):DiatomF2Low		0.9804 (7.00)***	0.9783 (7.00)***
EDF: ti(DateNum):DiatomF2High		2.0603 (7.00)***	2.3307 (7.00)***
EDF: ti(YearNum2,DateNum):DiatomF2Low			0.5604 (42.00)
EDF: ti(YearNum2,DateNum):DiatomF2High			6.1099 (42.00)***
EDF: ti(YearNum2,Age):DiatomF2Low			0.4217 (41.00)
EDF: ti(YearNum2,Age):DiatomF2High			1.2372 (42.00)
EDF: ti(Age,BLm)			0.4424 (47.00)
EDF: ti(YearNum2,BLm):DiatomF2Low			0.0003 (42.00)
EDF: ti(YearNum2,BLm):DiatomF2High			0.0003 (42.00)
AIC	2118.0839	2122.2358	2111.7467
BIC	2240.7609	2265.2872	2296.7652
Log Likelihood	-1034.9722	-1033.0505	-1019.5719
R^2	0.6940	0.6941	0.6995
Num. obs.	1208	1208	1208

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

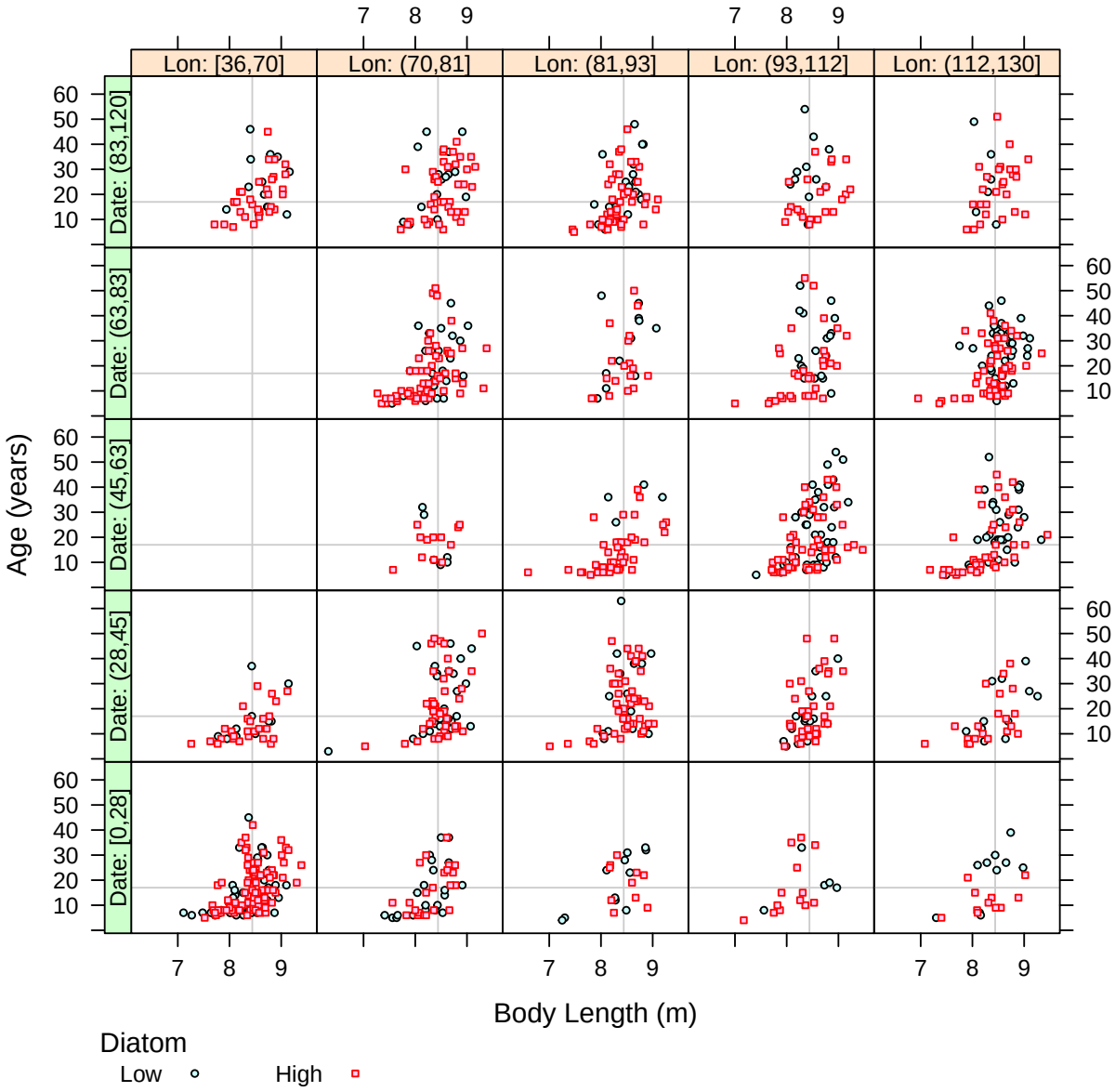

Figure 27: Spread of data across the design space for males from the West, showing scatterplot of length against age, conditioned on quintiles of DateNum and longitude, and showing diatom score using symbol type. Grey reference lines show median length and age.

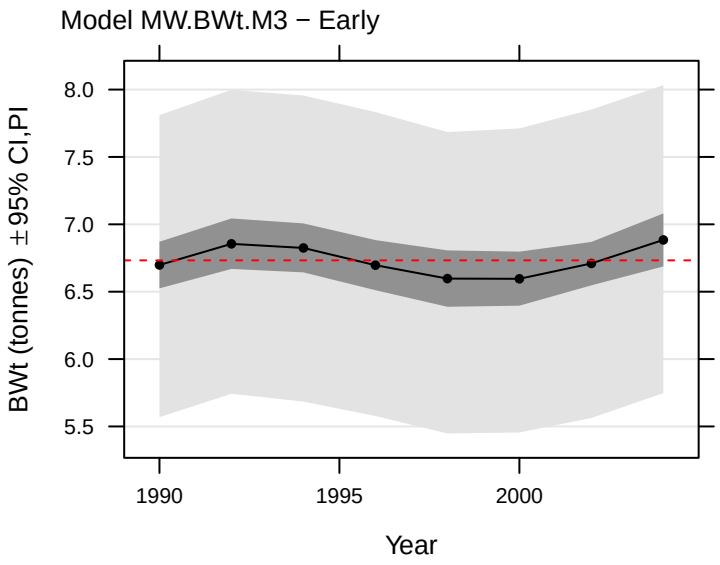

Figure 28: Predicted long-term trend in BWt from model MW.BWt.M3 for median length, median age males with low diatom loads that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the West. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

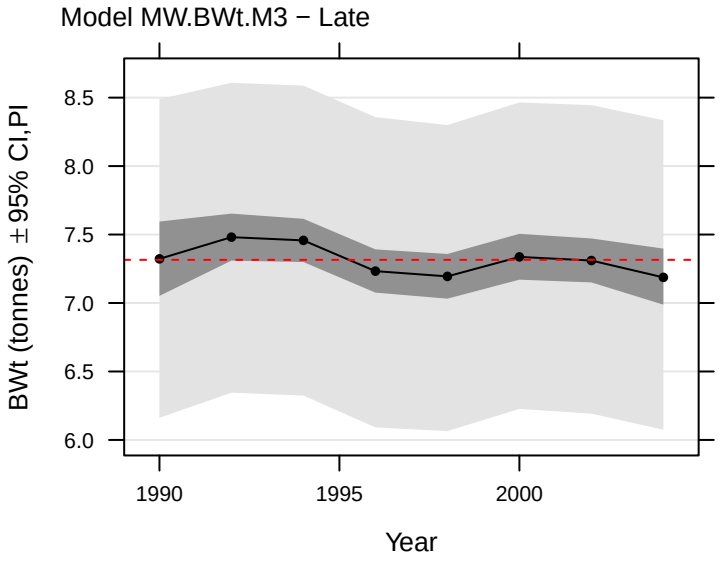

Figure 29: Predicted long-term trend in BWt from model MW.BWt.M3 for median length, median age males with high diatom loads that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the West. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

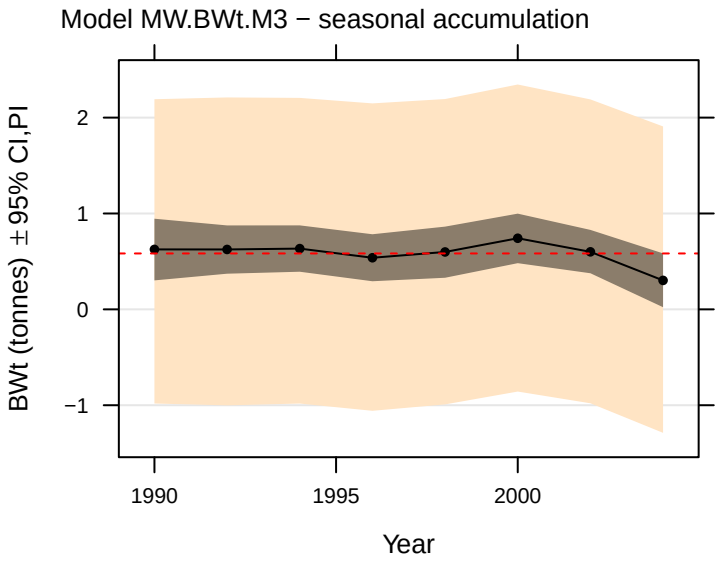

Figure 30: Predicted long-term trend in the yearly accumulation of BWt between the 20th and 80th percentiles of DateNum (28-83) from model MW.BWt.M3 for median length, median age males captured near the ice edge in the West. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

C.2 West - Blubber thickness

For models of blubber thickness there are 1208 male samples from the West. Models with `BT11` as response show that several first-order tensor-product interactions are important in explaining changes in blubber thickness, and that body weight is an important predictor (Table 19).

We obtain predictions for set values of the covariates for models `MW.BT11.M3` (without `BWt` as covariate) and `MW.BT11.M4` (with `BWt` as a covariate), using the values detailed in Section 3.2, Table 9. We predict for median longitude (86°E), and obtain early and late season prediction by conditioning on the 20th and 80th percentiles of DateNum for Low and High diatom score groups, respectively. Our model for total weight (`MW.BWt.M3`) is used to obtain early- and late-season predicted values of body weight for median length, median age animals.

The predicted long-term trend in `BT11` for male animals that are relatively newly arrived in the West shows a gradual decrease to around the year 2000, followed by a recovery in `BT11` to around mean levels by 2004 (Figure 31a,b). These results were largely consistent irrespective of whether body weight was included as a covariate in models. Total change in mean `BT11` over the sampling period was around 0.5 cm, which we take to indicate relatively stable condition.

Late season predictions for the trend in `BT11` for males from the West, irrespective of using body weight as a covariate, indicates a gradual but largely consistent decrease in `BT11` over the period of JARPA (Figure 32a,b). Over the period 1990-2002 late season predictions of `BT11` decreased by up to 0.6 cm, irrespective of whether `BWt` was included as a covariate or not in models. We would point out, however, that in the model accounting for `BWt` all but two points are quite similar to one another, indicating relative stability over most the period.

The seasonal trend in the accumulation of `BT11` takes into account both the early- and late-season predictions to estimate the improvement in `BT11` between the 20th and 80th percentiles of `DateNum` for male animals in the West (Figure 33). These results show that the improvement in condition over summer has been remarkably consistent, except perhaps for a slight downturn in 2002 and 2004. This result is consistent for models both with and without body weight as a covariate.

Table 19: Summary of GAM fits for response BT11 for males from the West region.

	MW.BT11.M1	MW.BT11.M2	MW.BT11.M3	MW.BT11.M4
Icenear	0.0614 (0.04)	0.0611 (0.04)	0.0760 (0.05)	0.0978 (0.04)*
DiatomF2High	0.3713 (0.04)***	0.3802 (0.04)***	0.3708 (0.04)***	0.2943 (0.04)***
EDF: ti(BLm)	1.8139 (7.00)**	1.7903 (7.00)**	2.0313 (7.00)***	2.0732 (7.00)***
EDF: ti(Age)	0.0008 (7.00)	0.0002 (7.00)	0.0002 (7.00)	2.8866 (7.00)***
EDF: ti(YearNum2)	4.9088 (6.00)***			
EDF: ti(DateNum)	2.7034 (7.00)***			
EDF: ti(LongNum):Icefar	0.0012 (7.00)	0.0048 (7.00)	0.0002 (7.00)	0.0004 (7.00)
EDF: ti(LongNum):Icenear	1.9337 (7.00)*	1.9511 (7.00)*	2.0423 (7.00)**	2.2073 (7.00)**
EDF: ti(YearNum2):DiatomF2Low		4.2201 (6.00)***	4.2288 (6.00)***	3.0814 (6.00)***
EDF: ti(YearNum2):DiatomF2High		3.2071 (6.00)***	2.8738 (6.00)***	2.8086 (6.00)***
EDF: ti(DateNum):DiatomF2Low		1.9189 (7.00)***	1.9118 (7.00)***	1.6468 (7.00)***
EDF: ti(DateNum):DiatomF2High		2.3586 (7.00)***	2.0081 (7.00)***	0.9946 (7.00)***
EDF: ti(YearNum2,DateNum):DiatomF2Low			0.1894 (42.00)	1.2541 (42.00)
EDF: ti(YearNum2,DateNum):DiatomF2High			11.9059 (42.00)***	10.9151 (42.00)***
EDF: ti(YearNum2,Age):DiatomF2Low			0.3292 (42.00)	0.3019 (42.00)
EDF: ti(YearNum2,Age):DiatomF2High			0.9893 (42.00)	0.0016 (42.00)
EDF: ti(Age,BLm)			1.6687 (47.00)	1.1235 (47.00)
EDF: ti(YearNum2,BLm):DiatomF2Low			0.0004 (42.00)	0.0006 (42.00)
EDF: ti(YearNum2,BLm):DiatomF2High			2.2017 (39.00)	2.2778 (42.00)*
EDF: ti(BWt):DiatomF2Low				0.9897 (7.00)***
EDF: ti(BWt):DiatomF2High				0.9947 (7.00)***
AIC	2506.4634	2516.6164	2502.1203	2328.6676
BIC	2598.0796	2635.4495	2742.9746	2573.4031
Log Likelihood	−1235.2562	−1234.9926	−1203.8035	−1116.3156
R^2	0.3229	0.3208	0.3457	0.4334
Num. obs.	1208	1208	1208	1208

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

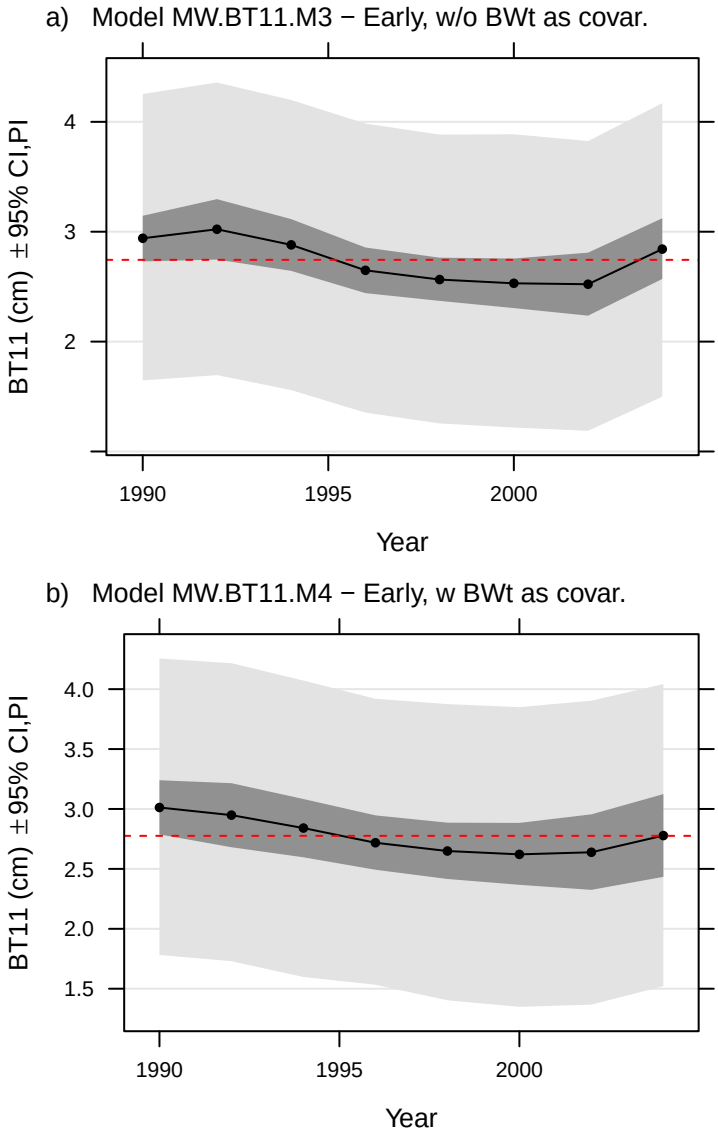

Figure 31: Predicted long-term trend in BT11 from: a) model MW.BT11.M3 predictions for median length, median age males with low diatom loads that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the West; and, b) model MW.BT11.M5 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

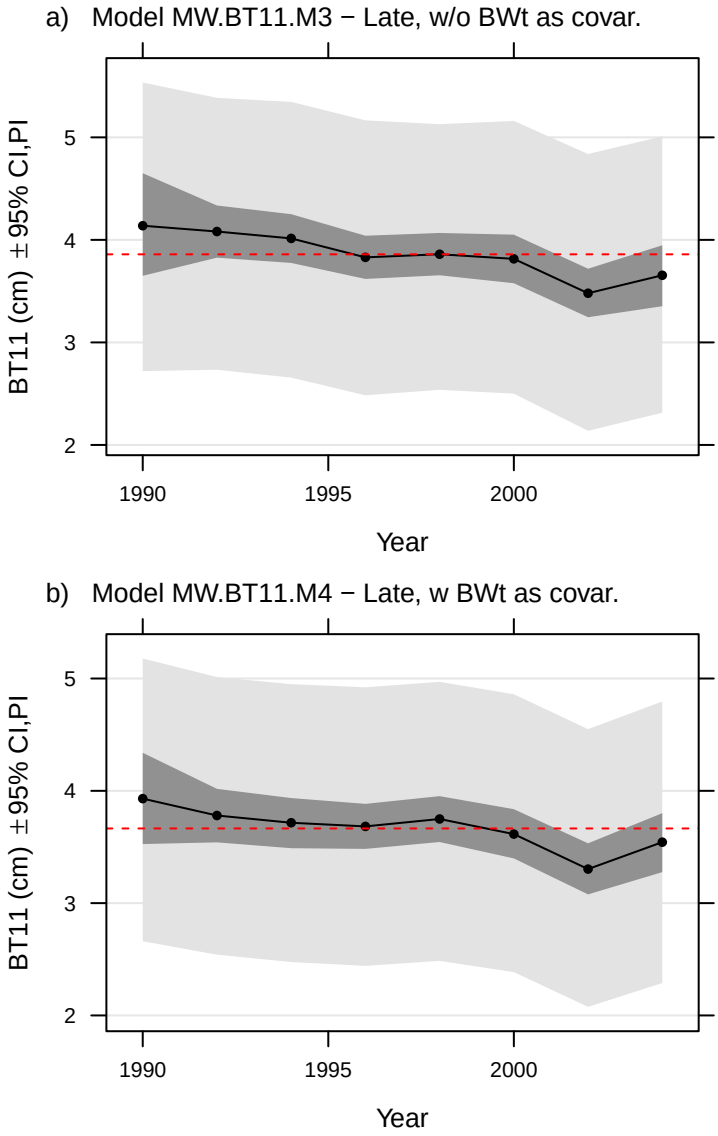

Figure 32: Predicted long-term trend in BT11 from: a) model MW.BT11.M4 predictions for median length, median age males with high diatom loads that are captured towards the end of the sampling period (80th percentile of DateNum) in the West; and, b) model MW.BT11.M5 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

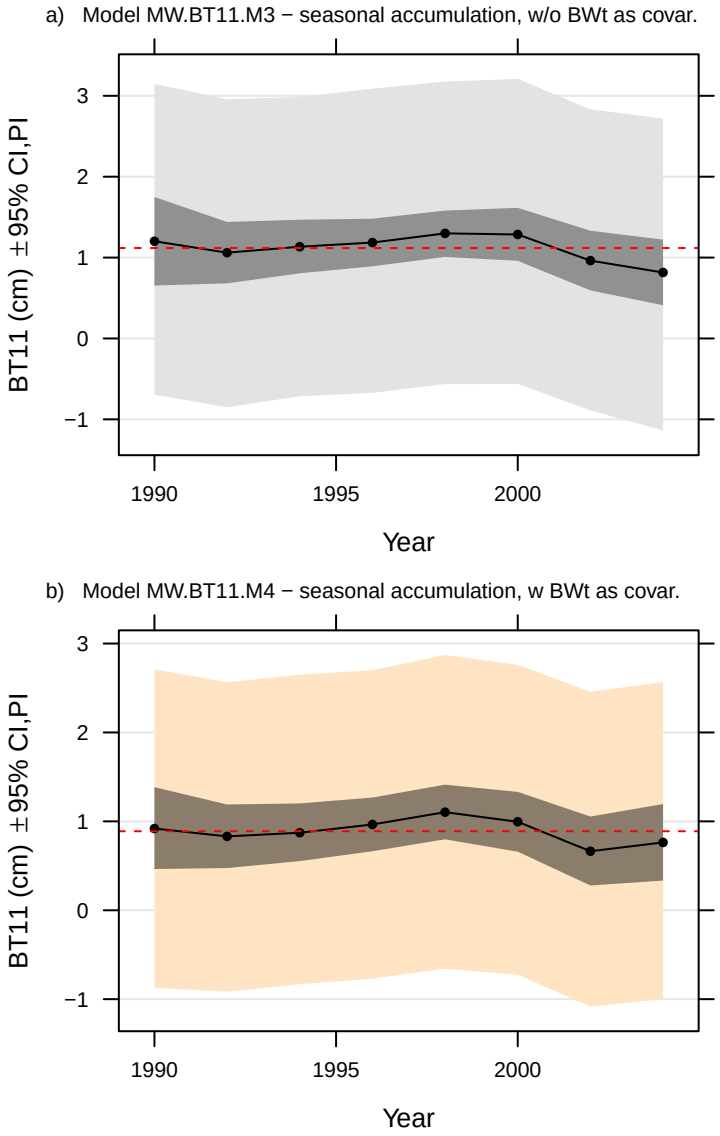

Figure 33: Predicted long-term trend in the yearly accumulation of BT11 between the 20th and 80th percentiles of DateNum from: a) model MW.BT11.M3 for median length, median age males captured near the ice edge in the West; and, b) the same calculation based on model MW.BT11.M4 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

Diagnostics for these models are presented in Appendix D.6 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not included for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics looked satisfactory, except for some slight deviations in the distributional assumptions revealed by Q-Q plots which are considered acceptable given the sample size and degree of deviation.

Table 20: Summary of GAM fits for response BWt for males from the East region.

	ME.BWt.M1	ME.BWt.M2	ME.BWt.M3
Icenear	0.0218 (0.04)	0.0356 (0.04)	0.0438 (0.04)
DiatomF2High	0.2204 (0.03)***	0.2275 (0.03)***	0.2304 (0.03)***
EDF: ti(BLm)	3.8166 (7.00)***	3.9826 (7.00)***	0.9989 (7.00)***
EDF: ti(Age)	3.9244 (7.00)***	3.8332 (7.00)***	3.8079 (7.00)***
EDF: ti(YearNum2)	3.6647 (6.00)***		
EDF: ti(DateNum)	3.0809 (7.00)***		
EDF: ti(LongNum):Icefar	0.5246 (7.00)	0.5580 (7.00)	0.7413 (7.00)*
EDF: ti(LongNum):Icenear	4.1871 (7.00)***	4.2072 (7.00)***	4.2919 (7.00)***
EDF: ti(YearNum2):DiatomF2Low		1.0342 (6.00)	0.0000 (6.00)
EDF: ti(YearNum2):DiatomF2High		3.5511 (6.00)***	3.4970 (6.00)***
EDF: ti(DateNum):DiatomF2Low		1.8990 (7.00)***	1.8076 (7.00)***
EDF: ti(DateNum):DiatomF2High		0.9969 (7.00)***	1.0541 (7.00)***
EDF: ti(YearNum2,DateNum):DiatomF2Low			1.8908 (42.00)**
EDF: ti(YearNum2,DateNum):DiatomF2High			0.1398 (42.00)
EDF: ti(YearNum2,Age):DiatomF2Low			0.0001 (42.00)
EDF: ti(YearNum2,Age):DiatomF2High			0.0001 (42.00)
EDF: ti(Age,BLm)			2.4907 (41.00)*
EDF: ti(YearNum2,BLm):DiatomF2Low			0.7332 (42.00)
EDF: ti(YearNum2,BLm):DiatomF2High			0.6079 (40.00)
AIC	2284.6949	2287.7270	2279.4753
BIC	2429.9254	2437.6733	2442.5210
Log Likelihood	−1114.3234	−1114.9295	−1108.2759
R^2	0.6937	0.6932	0.6958
Num. obs.	1316	1316	1316

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

C.3 East - Total weight

There are 1316 male samples from the East. Models with body weight as response are summarised in Table 20, and indicate the importance of several interaction terms, including an interaction between age and body length. Allowing separate smooths per diatom class in main effects and interactions (`M3`) provides little gain over a main effects model (`M1`), however it does allow small-scale differences between diatom classes to be captured and accumulatively these may be important for predictions.

We obtain predictions for model `ME.BWt.M3` at set values of the covariates, as detailed in Section 3.2, Table 9. A plot of the data according to the important covariates confirms there are sufficient data around our fixed values of the covariates for purposes of prediction (Figure 34). We again obtain predictions for low and high diatom load animals at the 20th and 80th percentiles of `DateNum`, respectively.

The predicted long-term trend in BWt for male animals that are relatively newly arrived in the East shows very little variation around the overall mean of the series (Figure 35). These results indicate essentially no change in early season body weight over the nine sampling years.

Late season predictions for males from the East show a slight increase in the early 1990's, followed by slight decrease until around 2001, and end with an upward inflection in 2005 (Figure 36). To our eyes, this appears to be slight variation around a fairly constant mean.

The trend in the seasonal accumulation of `BWt` between the 20th and 80th percentiles of `DateNum` for low and high diatom load animals, respectively, reveals a pattern very similar to that estimated for late season BWt (Figure 37). Again, we see not much variation around what appears to be a constant mean.

Standard diagnostics for these models are presented in Appendix D.7 and include model summaries, partial

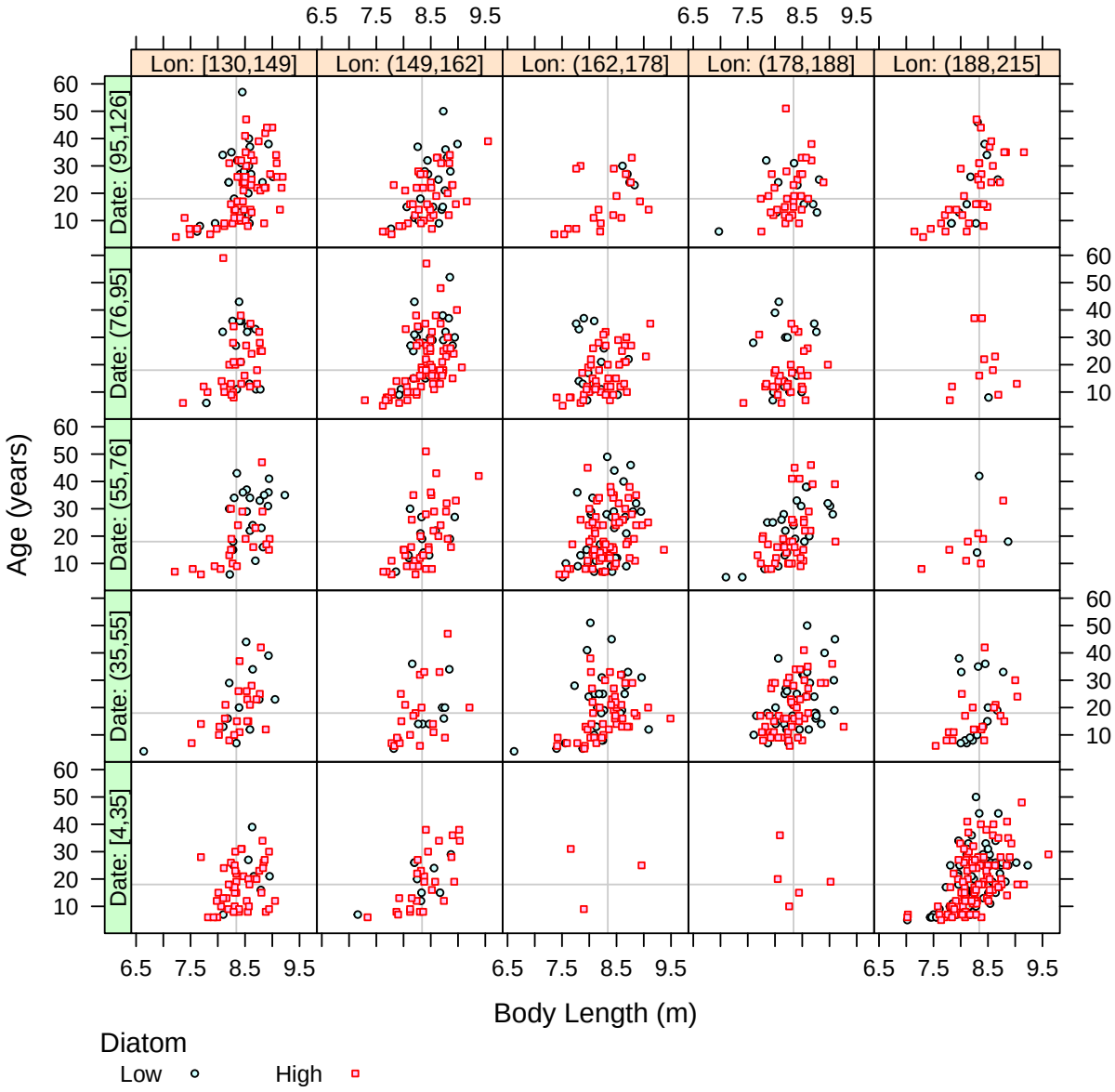

Figure 34: Spread of data across the design space for males from the East, showing scatterplot of length against age, conditioned on quintiles of DateNum and longitude, and showing diatom score using symbol type. Grey reference lines show median length and age.

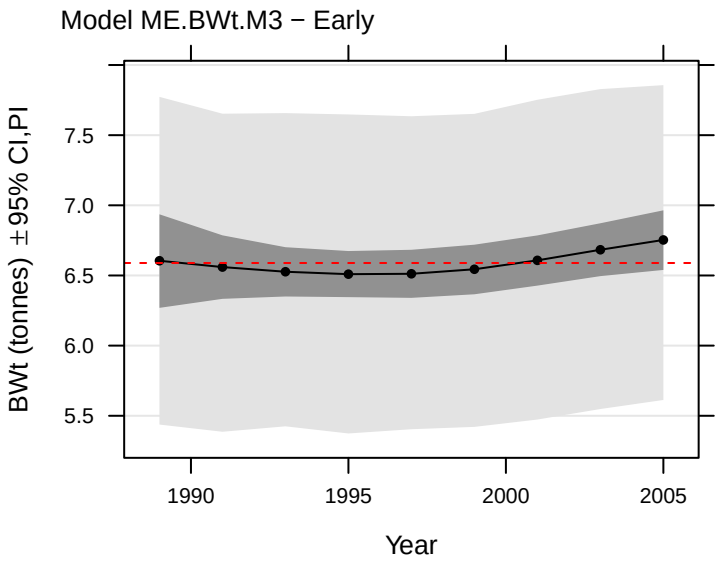

Figure 35: Predicted long-term trend in BWt from model ME.BWt.M3 for median length, median age males with low diatom loads that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the East. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

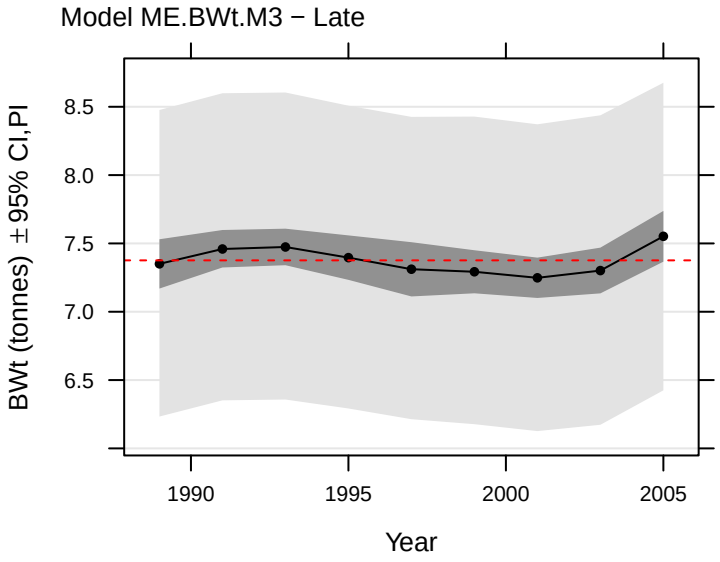

Figure 36: Predicted long-term trend in BWt from model ME.BWt.M3 for median length, median age males with high diatom loads that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the East. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

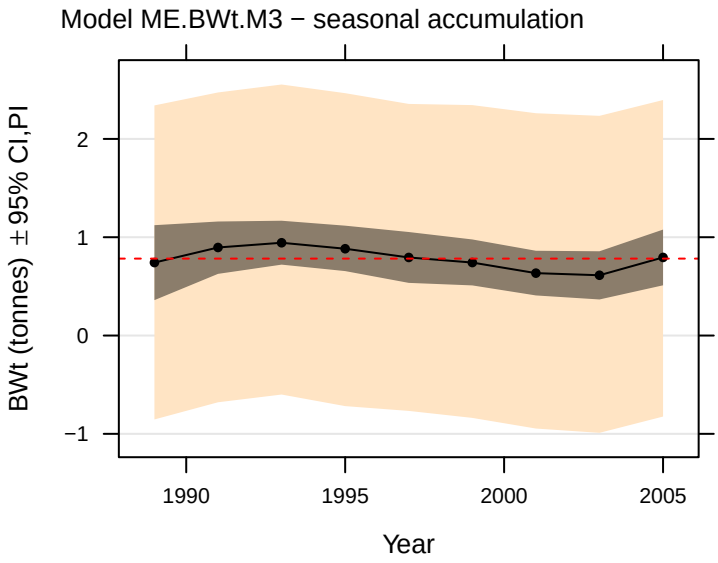

Figure 37: Predicted long-term trend in the yearly accumulation of BWt between the 20th and 80th percentiles of DateNum from model ME.BWt.M3 for median length, median age males captured near the ice edge in the East. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not appended for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). All diagnostics looked satisfactory except for some slight departures from error distribution assumptions evident in the tails of Q-Q plots. However, these departures are considered minor in light of sample sizes.

Table 21: Summary of GAM fits for response BT11 for males from the East region.

	ME.BT11.M1	ME.BT11.M2	ME.BT11.M3	ME.BT11.M4
Icenear	0.1488 (0.05)**	0.1492 (0.05)**	0.1322 (0.05)**	0.1401 (0.04)**
DiatomF2High	0.4952 (0.04)***	0.4982 (0.04)***	0.4845 (0.04)***	0.3861 (0.04)***
EDF: ti(BLm)	0.1345 (7.00)	0.0014 (7.00)	0.3682 (7.00)	0.9873 (7.00)***
EDF: ti(Age)	1.3464 (7.00)*	1.3271 (7.00)**	1.3356 (7.00)**	0.8335 (7.00)*
EDF: ti(YearNum2)	2.8802 (6.00)**			
EDF: ti(DateNum)	4.0643 (7.00)***			
EDF: ti(LongNum):Icefar	1.8753 (7.00)**	1.6297 (7.00)**	1.8731 (7.00)**	1.7583 (7.00)***
EDF: ti(LongNum):Icenear	3.5167 (7.00)***	3.3106 (7.00)***	2.9377 (7.00)***	2.1231 (7.00)***
EDF: ti(YearNum2):DiatomF2Low		0.0003 (6.00)	0.0011 (6.00)	0.3802 (6.00)
EDF: ti(YearNum2):DiatomF2High		2.5397 (6.00)*	2.4079 (6.00)	3.6849 (6.00)***
EDF: ti(DateNum):DiatomF2Low		2.5209 (7.00)***	2.5258 (7.00)***	1.1302 (7.00)***
EDF: ti(DateNum):DiatomF2High		4.1681 (7.00)***	4.0710 (7.00)***	3.7245 (7.00)***
EDF: ti(YearNum2,DateNum):DiatomF2Low			2.4016 (42.00)*	0.0013 (42.00)
EDF: ti(YearNum2,DateNum):DiatomF2High			7.8673 (42.00)***	6.1079 (42.00)***
EDF: ti(YearNum2,Age):DiatomF2Low			0.0007 (42.00)	0.0010 (42.00)
EDF: ti(YearNum2,Age):DiatomF2High			0.0017 (42.00)	0.5611 (42.00)
EDF: ti(Age,BLm)			0.0006 (48.00)	0.0007 (48.00)
EDF: ti(YearNum2,BLm):DiatomF2Low			0.1545 (42.00)	0.7962 (42.00)
EDF: ti(YearNum2,BLm):DiatomF2High			2.4326 (42.00)*	4.6318 (42.00)*
EDF: ti(BWt):DiatomF2Low				2.4702 (7.00)***
EDF: ti(BWt):DiatomF2High				2.2050 (7.00)***
AIC	2849.6709	2846.4145	2830.3607	2645.1338
BIC	2962.6103	2969.8671	3049.5805	2880.3343
Log Likelihood	−1403.0423	−1399.3856	−1372.8791	−1277.1820
R^2	0.4494	0.4517	0.4681	0.5390
Num. obs.	1316	1316	1316	1316

***p<0.001, **p<0.01, *p<0.05. Model notation: parametric terms are reported as estimated coefficient (SE); smooth terms are prefaced with 'EDF' and are labelled as either s(term) for thin plate regression splines main effects, or, in the case of tensor product interaction decompositions, as ti(term) for main effects or ti(term1, term2) for interaction components. Effective degrees of freedom under shrinkage are shown for all smooth terms, followed in brackets by the basis dimension used.

C.4 East - Blubber thickness

For models of blubber thickness there are also 1316 male samples from the East.

Model results with `BT11` as response for male animals from the East show that interactions (`M3`) offer only minor improvement over a main effects model (`M1`, `M2`), and that body weight is again an important predictor (`M4`) (Table 21). Longitudinal effects are shown to be moderately important, and partial effects plots show a similar increase in `BT11` for both diatom classes as animals are caught West to East (74).

We obtain predictions for set values of the covariates for models `ME.BT11.M3` (without `BWt` as covariate) and `ME.BT11.M4` (with `BWt` as a covariate), using the values detailed in Table 9. As before, we obtain early and late season prediction by conditioning on the 20th and 80th percentiles of DateNum, separately for Low and High diatom load groups, for animals captured near the ice edge. We additionally condition on biennial predicted weights of median length, median age animals at early- and late-season time points from model `ME.BWt.M3`.

The predicted long-term trend in `BT11` for new arrived male animals caught near the ice edge in the East is stable over the sampling period, markedly so when `BWt` is included as a covariate (Figure 38a,b).

The predicted long-term trend in `BT11` for lat-season male animals caught near the ice edge in the East shows a slight decline in `BT11`, of the order of 2-3 mm, for years 1991 and 1993, but otherwise is reasonably stable (Figure 39a,b).

The seasonal trend in the accumulation of `BT11` between the 20th and 80th percentiles of `DateNum` for low

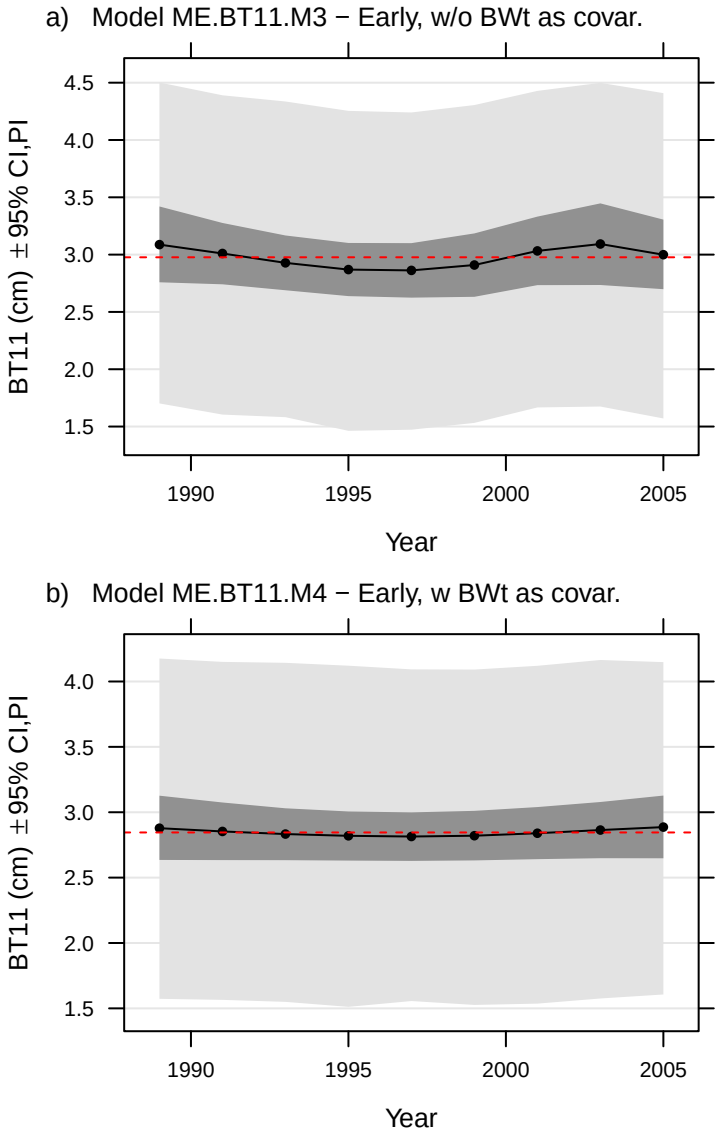

Figure 38: Predicted long-term trend in BT11 from: a) model ME.BT11.M3 predictions for median length, median age males with low diatom loads that are captured near the ice edge at the beginning of the sampling period (20th percentile of DateNum) in the East; and, b) model ME.BT11.M4 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

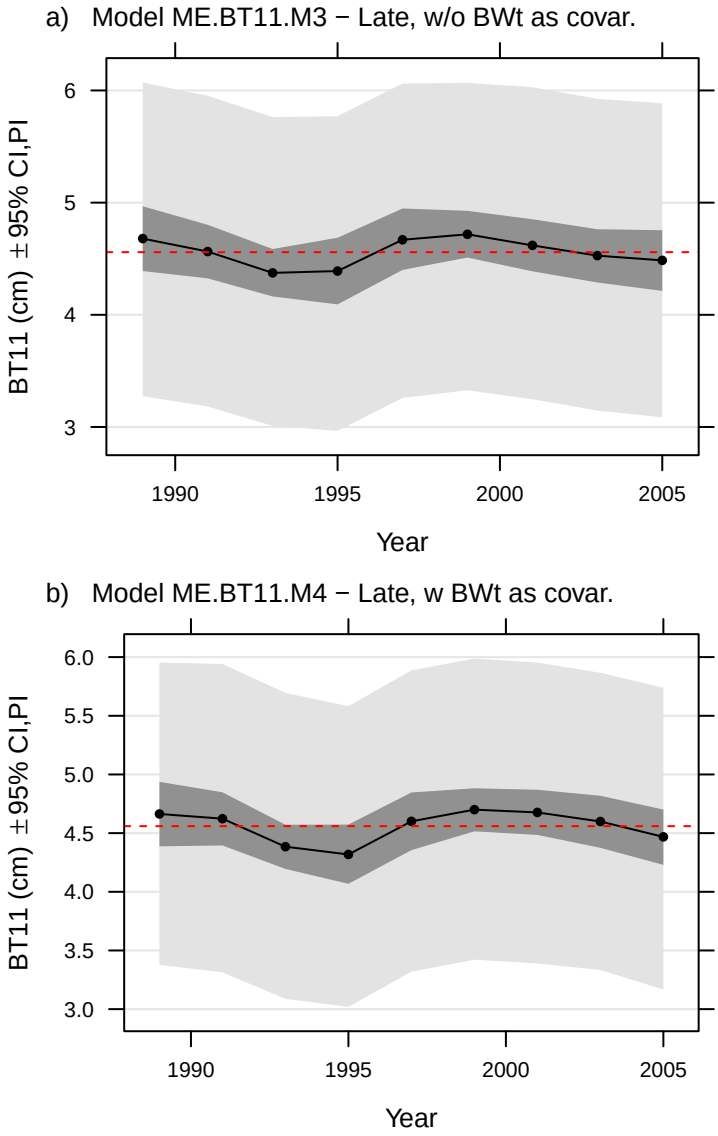

Figure 39: Predicted long-term trend in BT11 from: a) model ME.BT11.M3 predictions for median length, median age males with high diatom loads that are captured near the ice edge towards the end of the sampling period (80th percentile of DateNum) in the East; and, b) model ME.BT11.M4 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean function (dark) and point predictions (light) are provided. Dashed red lines represent overall mean(y).

and high diatom load animals, respectively, shows only minor deviations around a fairly constant mean value (Figure 40).

Standard diagnostics for these models are presented in Appendix D.8 and include model summaries, partial effects plots of all terms and residual diagnostics. Several other checks were undertaken but were not appended for reasons of brevity, including a check that the number of basis dimensions per smooth term was adequate and an assessment of concurvity between model terms (the gam analogue of colinearity). Other than some slight deviations in the tails of Q-Q plots, the diagnostics were unremarkable.

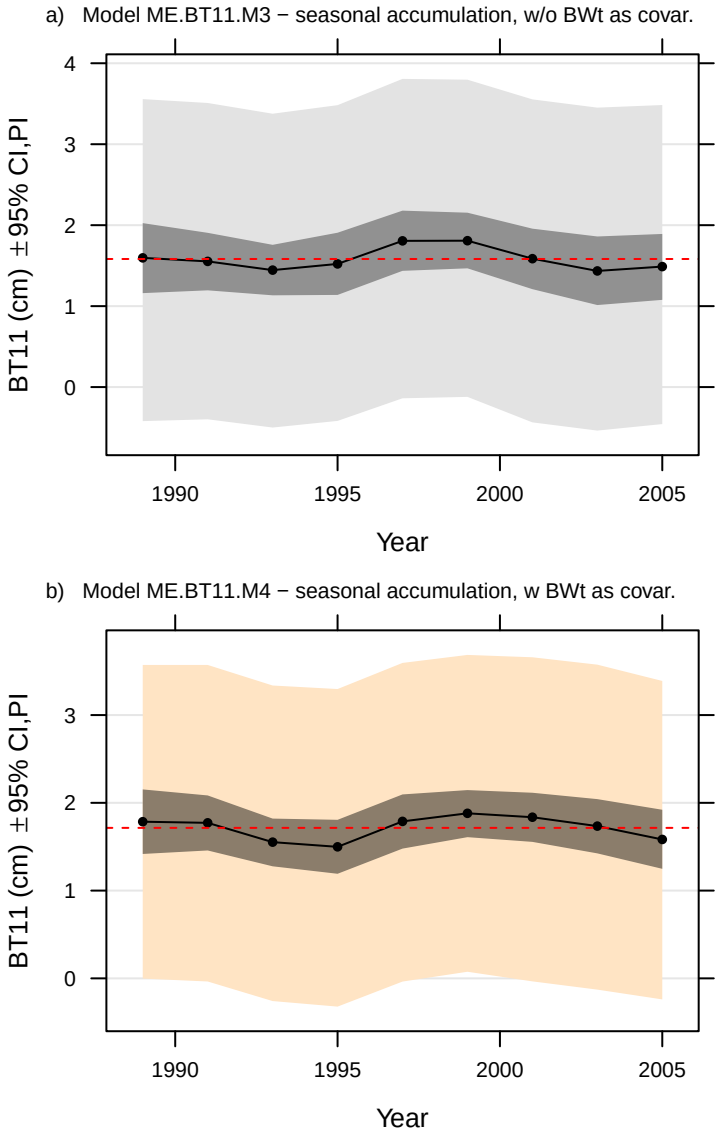

Figure 40: Predicted long-term trend in the yearly accumulation of BT11 between the 20th and 80th percentiles of DateNum from: a) model ME.BT11.M3 for median length, median age males captured near the ice edge in the East; and, b) the same calculation based on model ME.BT11.M4 predictions for the same covariate set but additionally conditioned on median body weight. Bayesian 95% credible intervals for the mean difference (dark) and point differences (light) are provided. Dashed red lines represent overall mean(y).

D Model diagnostics

D.1 Females - West - Total weight

Diagnostics are presented for gam model `FW.BWt.M2` of body weight, first presented in Section B.1.

A summary for `FW.BWt.M2` is provided in Figure 41.

Partial effects plots are provided for model `FW.BWt.M2` in Figure 42.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for the sake of brevity.

A check of standard residual diagnostics for model `FW.BWt.M2` revealed no appreciable lack-of-fit or departures from usual model assumptions (Figure 43).

Summary of model FW.BWt.M2

```
Family: gaussian
Link function: identity

Formula:
BWt ~ ti(BLm, k = FW.k1) + ti(Age, k = FW.k1) + ti(YearNum2,
    k = FW.k2) + ti(DateNum, k = FW.k1) + Ice + ti(LongNum, k = FW.k3,
    by = Ice) + ti(FetusLength, k = FW.k1) + ti(YearNum2, DateNum,
    k = c(FW.k2, FW.k1)) + ti(YearNum2, FetusLength, k = c(FW.k2,
    FW.k1)) + ti(YearNum2, Age, k = c(FW.k2, FW.k1)) + ti(DateNum,
    FetusLength, k = FW.k1) + ti(Age, BLm, k = FW.k1) + ti(YearNum2,
    BLm, k = FW.k1) + DiatomF2

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.24360    0.10086  81.731   <2e-16 ***
Icenear      -0.04643    0.10203  -0.455    0.649
DiatomF2High  0.08235    0.07565   1.089    0.277
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df      F  p-value
ti(BLm)                  2.8592341      7 93.914  < 2e-16 ***
ti(Age)                  0.9782536      7  6.318 2.47e-11 ***
ti(YearNum2)             4.8233438      6  5.445 2.60e-07 ***
ti(DateNum)              0.8144742      7  0.203  0.11046
ti(LongNum):Icefar       0.7055225     19  0.081  0.10842
ti(LongNum):Icenear      0.9537271     19  0.233  0.00502 **
ti(FetusLength)          1.8492328      7 11.618  < 2e-16 ***
ti(YearNum2,DateNum)     7.4043019     42  0.417  0.00199 **
ti(YearNum2,FetusLength) 3.6560777     42  0.155  0.06187 .
ti(YearNum2,Age)         0.0005958     42  0.000  0.71104
ti(DateNum,FetusLength)  0.0004033     49  0.000  0.69096
ti(Age,BLm)              0.0018782     49  0.000  0.38433
ti(YearNum2,BLm)         0.0015497     49  0.000  0.36485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.673   Deviance explained = 68.7%
-REML = 715.45  Scale est. = 0.49645   n = 638
```

Figure 41: Summary of GAM model FW.BWt.M2, female animals captured in the West region.

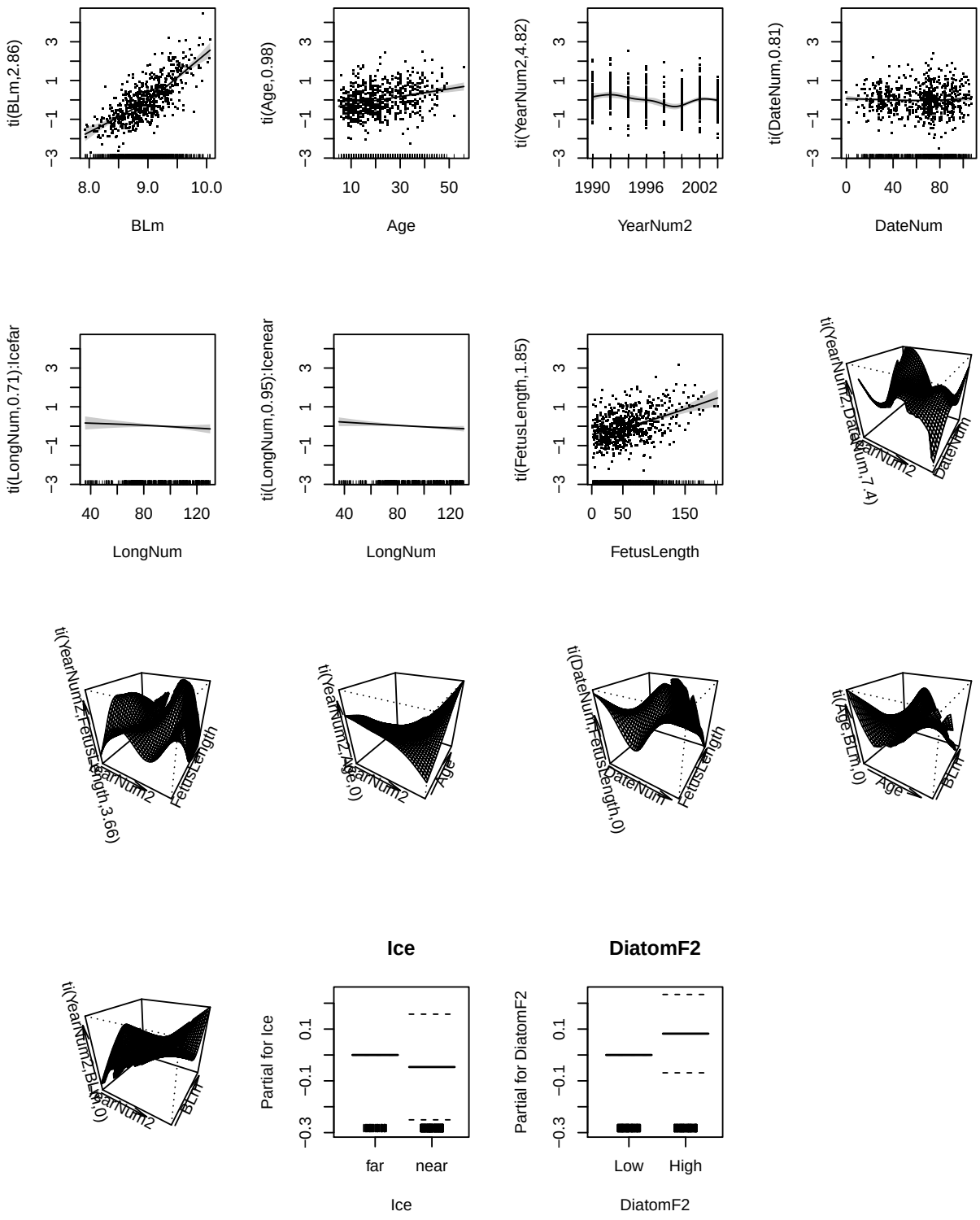

Figure 42: Partial effects plots for model FW.BWt.M2, gam fit to body weight for female animals captured in the West region.

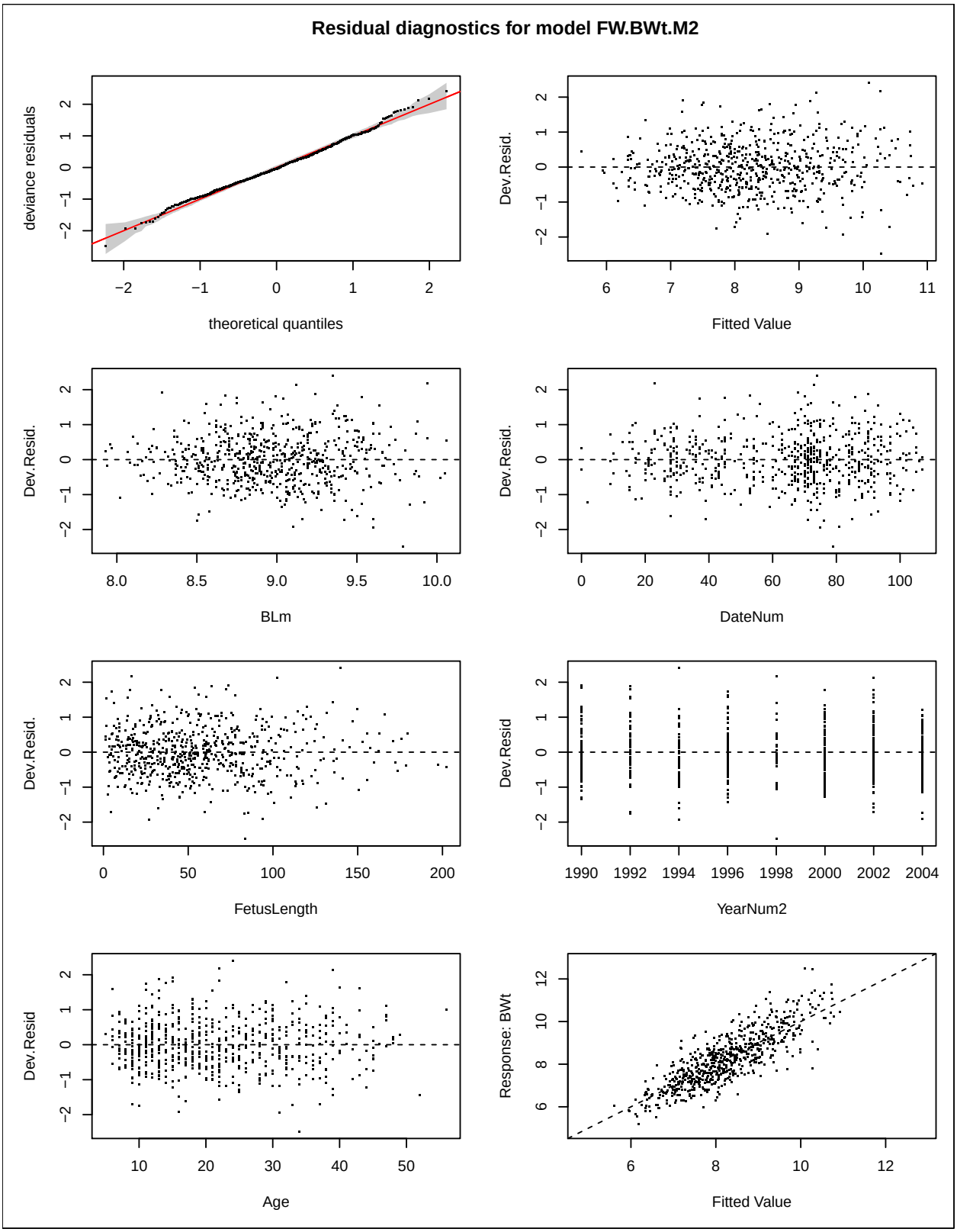

Figure 43: Residual diagnostics for model FW.BWt.M2 for female animals captured in the West region.

D.2 Females - West - Blubber thickness

Diagnostics are presented for gam models `FW.BT11.M2` and `FW.BT11.M3` of response `BT11`, first presented in Section B.2.

Model summaries for `FW.BT11.M2` and `FW.BT11.M3` are provided in Figures 44 and 45, respectively.

Partial effects plots are provided for models `FW.BT11.M2` and `FW.BT11.M3` in Figures 46 and 47.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth for models `FW.BT11.M2` and `FW.BT11.M3` proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for the sake of brevity.

A check of standard residual diagnostics for models `FW.BT11.M2` and `FW.BT11.M3` shows showed some slight deviations from distributional assumption in the tails, but no appreciable lack-of-fit or departures from usual model assumptions (Figures 48 and 49).

Summary of model FW.BT11.M2

```
Family: gaussian
Link function: identity

Formula:
BT11 ~ ti(BLm, k = FW.k1) + ti(Age, k = FW.k1) + ti(YearNum2,
    k = FW.k2) + ti(DateNum, k = FW.k1) + Ice + ti(LongNum, k = FW.k3,
    by = Ice) + ti(FetusLength, k = FW.k1) + ti(YearNum2, DateNum,
    k = c(FW.k2, FW.k1)) + ti(YearNum2, FetusLength, k = c(FW.k2,
    FW.k1)) + ti(YearNum2, Age, k = c(FW.k2, FW.k1)) + ti(DateNum,
    FetusLength, k = FW.k1) + ti(Age, BLm, k = FW.k1) + ti(YearNum2,
    BLm, k = c(FW.k2, FW.k1)) + DiatomF2

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.54850    0.09961  35.625   <2e-16 ***
Icenear       0.19145    0.10072   1.901   0.0578 .
DiatomF2High  0.08663    0.07500   1.155   0.2485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df      F  p-value
ti(BLm)                  0.5276411      7  0.160 0.134245
ti(Age)                  1.5900202      7  0.907 0.011696 *
ti(YearNum2)             3.9347466      6  6.661 1.93e-09 ***
ti(DateNum)              3.0009064      7  2.468 6.45e-05 ***
ti(LongNum):Icefar       0.0016076     19  0.000 0.317518
ti(LongNum):Icenear      0.0001796     19  0.000 0.567879
ti(FetusLength)          0.9911444      7 15.898  < 2e-16 ***
ti(YearNum2,DateNum)     6.0824739     42  0.459 0.000425 ***
ti(YearNum2,FetusLength) 0.0002928     42  0.000 0.533425
ti(YearNum2,Age)         0.0001129     42  0.000 0.998371
ti(DateNum,FetusLength)  1.0068769     49  0.050 0.073197 .
ti(Age,BLm)              0.0004004     49  0.000 0.450746
ti(YearNum2,BLm)         0.0001665     42  0.000 0.876169
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.37   Deviance explained = 38.9%
-REML = 713.77  Scale est. = 0.50771   n = 638
```

Figure 44: Summary of GAM model FW.BT11.M2, female animals captured in the West region.

```
Family: gaussian
Link function: identity
```

Summary of model FW.BT11.M3

```
Formula:
BT11 ~ ti(BLm, k = FW.k1) + ti(Age, k = FW.k1) + ti(YearNum2,
    k = FW.k2) + ti(DateNum, k = FW.k1) + Ice + ti(LongNum, k = FW.k3,
    by = Ice) + ti(FetusLength, k = FW.k1) + ti(YearNum2, DateNum,
    k = c(FW.k2, FW.k1)) + ti(YearNum2, FetusLength, k = c(FW.k2,
    FW.k1)) + ti(YearNum2, Age, k = c(FW.k2, FW.k1)) + ti(DateNum,
    FetusLength, k = FW.k1) + ti(Age, BLm, k = FW.k1) + ti(YearNum2,
    BLm, k = c(FW.k2, FW.k1)) + DiatomF2 + ti(BWt, k = FW.k1) +
    ti(BWt, FetusLength, k = FW.k1)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.56289    0.09524  37.409   <2e-16 ***
Icenear       0.19858    0.09232   2.151   0.0319 *
DiatomF2High  0.04755    0.06923   0.687   0.4924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                              edf Ref.df      F  p-value
ti(BLm)                 0.9772346      7  6.115 3.04e-11 ***
ti(Age)                 1.0714377      7  0.363 0.084606 .
ti(YearNum2)            4.7115336      6  6.988 8.31e-09 ***
ti(DateNum)             2.5704533      7  1.950 0.000368 ***
ti(LongNum):Icefar      0.0004124     19  0.000 0.538420
ti(LongNum):Icenear     0.0002103     19  0.000 0.763899
ti(FetusLength)         0.9844926      7  9.020  < 2e-16 ***
ti(YearNum2,DateNum)    0.7670260     42  0.078 0.033396 *
ti(YearNum2,FetusLength) 0.0004746    42  0.000 0.738706
ti(YearNum2,Age)        0.0024102     42  0.000 0.346723
ti(DateNum,FetusLength)  3.8218171    49  0.163 0.032590 *
ti(Age,BLm)             0.0008713     49  0.000 0.504194
ti(YearNum2,BLm)        0.0004809     42  0.000 0.943388
ti(BWt)                 2.1140572      7 18.700  < 2e-16 ***
ti(BWt,FetusLength)     0.0003268     49  0.000 0.999438
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.469   Deviance explained = 48.5%
```

Figure 45: Summary of GAM model FW.BT11.M3, female animals captured in the West region. This model is identical to model FW-BT11-M2 with the exception for the addition of BWt as an explanatory variable of blubber thickness.

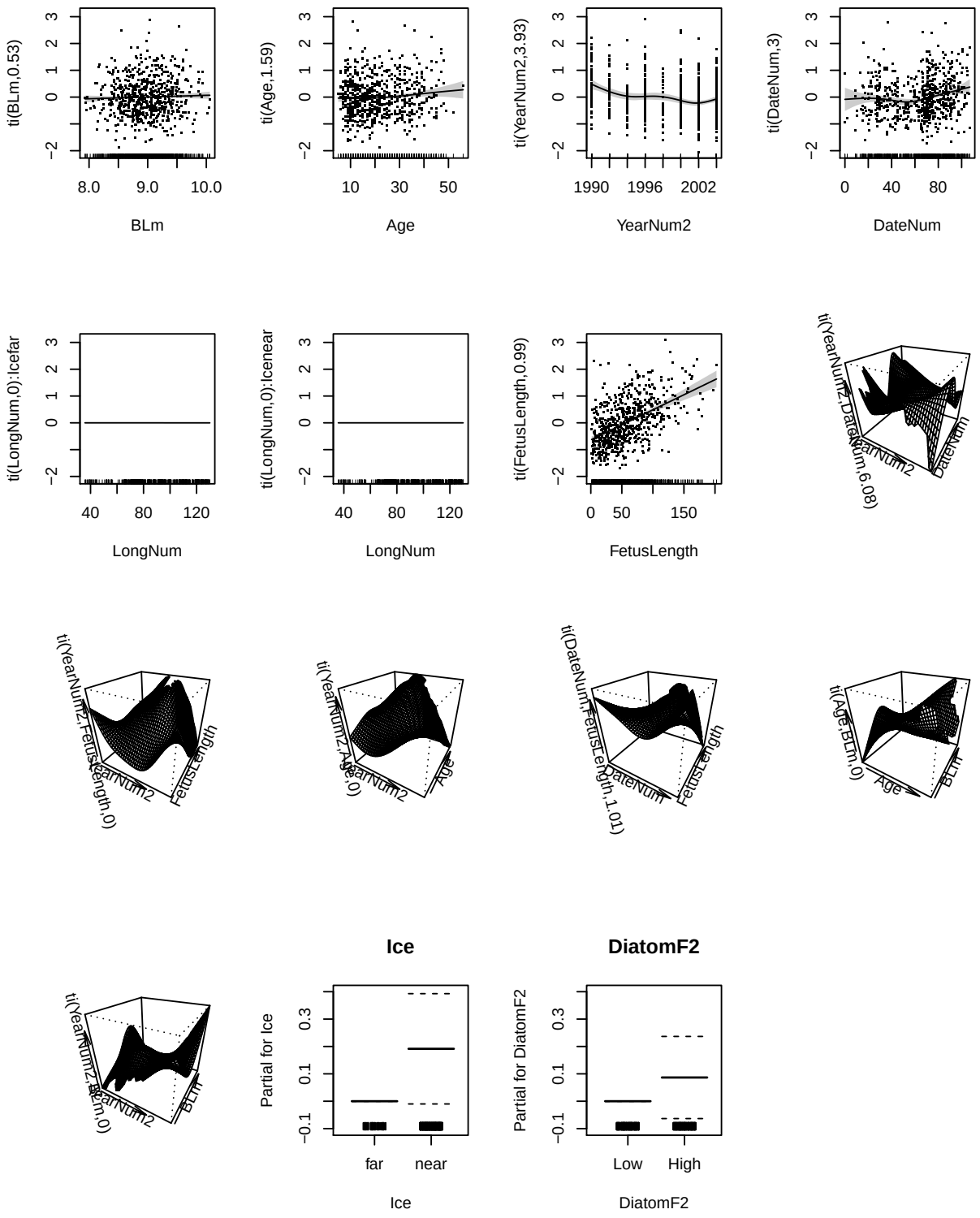

Figure 46: Partial effects plots for model FW.BT11.M2, gam fit to BT11 for female animals captured in the West region.

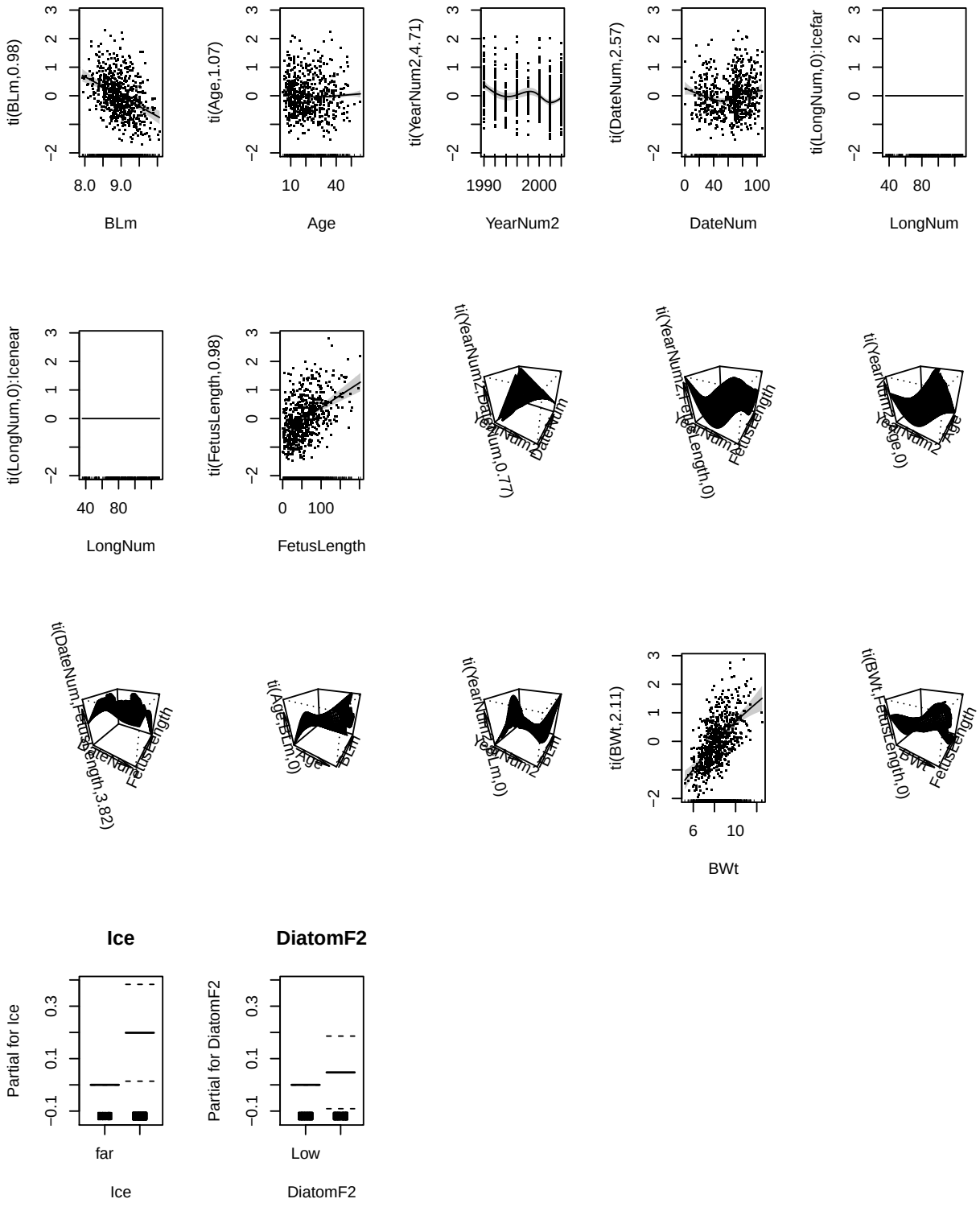

Figure 47: Partial effects plots for model FW.BT11.M3, gam fit to BT11 for female animals captured in the West region.

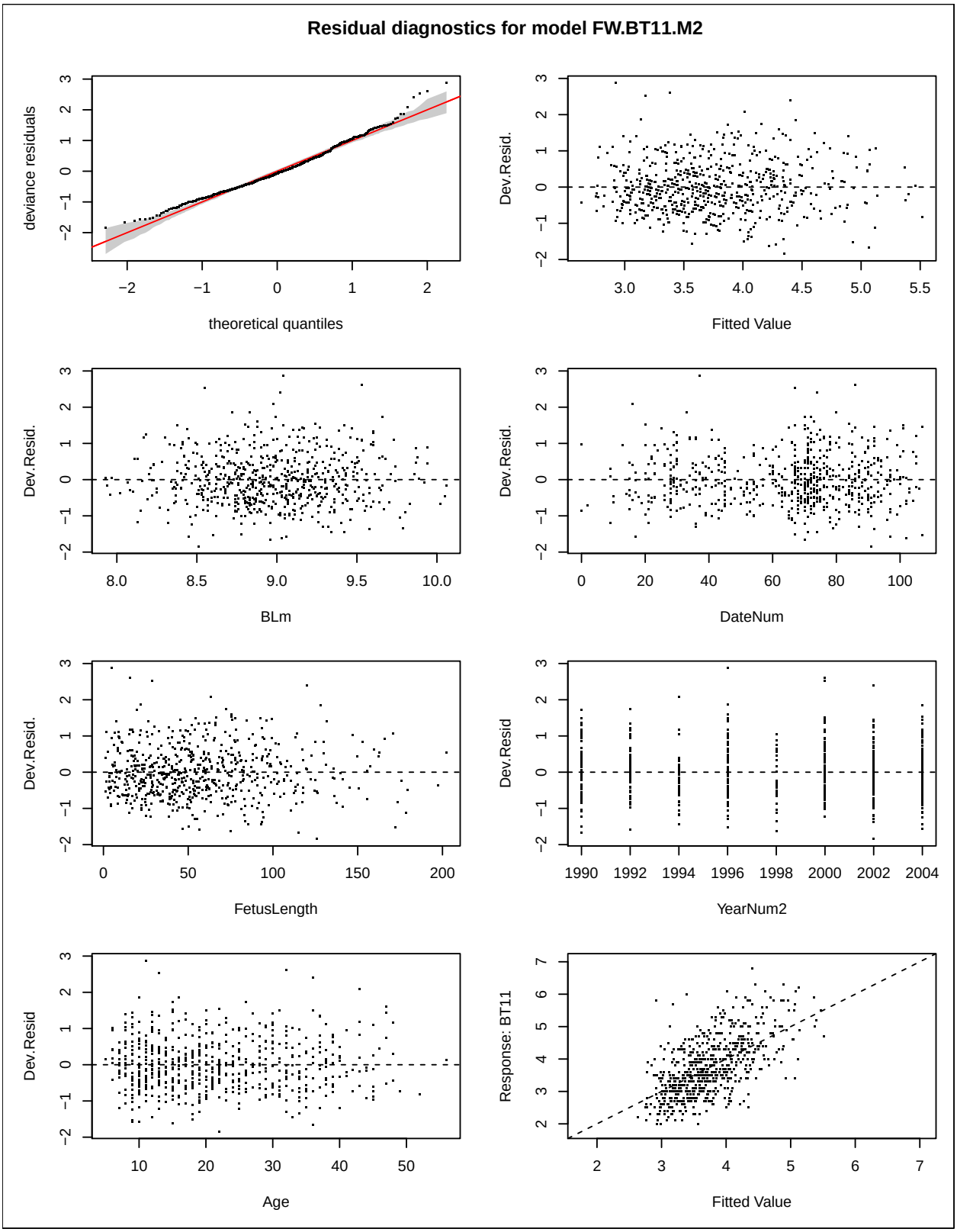

Figure 48: Residual diagnostics for model FW.BT11.M2 for female animals captured in the West region.

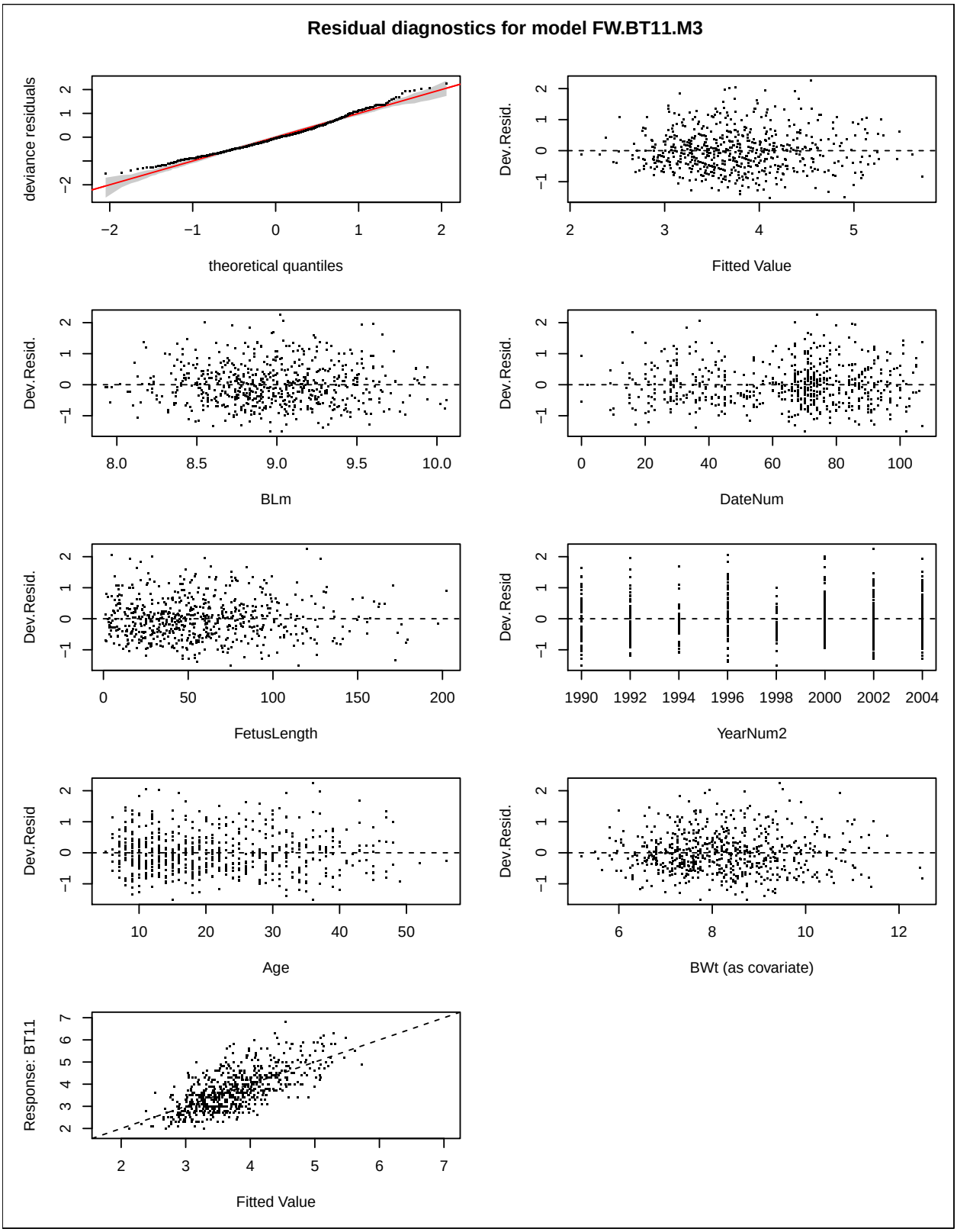

Figure 49: Residual diagnostics for model FW.BT11.M3 for female animals captured in the West region.

D.3 Females - East - Total weight

Diagnostics are presented for gam model `FE.BWt.M2` of body weight, first presented in Section B.3.

A summary for `FE.BWt.M2` is provided in Figure 50.

Partial effects plots are provided for model `FE.BWt.M2` in Figure 51.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth and a check of the concurvity between model terms (the gam equivalent of colinearity in linear models) proved satisfactory (results not shown).

A check of standard residual diagnostics for model `FE.BWt.M2` revealed no appreciable lack-of-fit or departures from usual model assumptions (Figure 52).

Summary of model FE.BWt.M2

```
Family: gaussian
Link function: identity

Formula:
BWt ~ ti(BLm, k = FE.k1) + ti(Age, k = FE.k1) + ti(YearNum2,
    k = FE.k2) + ti(DateNum, k = FE.k1) + Ice + ti(LongNum, k = FE.k3,
    by = Ice) + ti(FetusLength, k = FE.k1) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(FE.k2, FE.k1)) + ti(YearNum2, FetusLength,
    k = c(FE.k2, FE.k1)) + ti(YearNum2, Age, k = c(FE.k2, FE.k1)) +
    ti(DateNum, FetusLength, k = FE.k1) + ti(Age, BLm, k = FE.k1) +
    ti(YearNum2, BLm, k = c(FE.k2, FE.k1))

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.02103    0.07823 102.535   <2e-16 ***
Icenear      -0.06872    0.07632  -0.900    0.368
DiatomF2High  0.07641    0.05714   1.337    0.181
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df       F  p-value
ti(BLm)                  1.466e+00      7 121.385  < 2e-16 ***
ti(Age)                  9.779e-01      7   6.266 2.71e-11 ***
ti(YearNum2)             3.734e+00      6   3.567 5.09e-05 ***
ti(DateNum)              9.585e-01      7   2.877 2.32e-06 ***
ti(LongNum):Icefar       7.126e-01     17   0.146   0.0580 .
ti(LongNum):Icenear      4.172e+00     19   1.454 1.46e-06 ***
ti(FetusLength)          2.605e+00      7  19.567  < 2e-16 ***
ti(YearNum2,DateNum)     3.534e-01     42   0.010   0.2409
ti(YearNum2,FetusLength) 1.033e+01     42   0.757 5.07e-05 ***
ti(YearNum2,Age)         7.311e-04     42   0.000   0.5619
ti(DateNum,FetusLength)  1.076e+00     49   0.076   0.0293 *
ti(Age,BLm)              1.090e-02     49   0.000   0.3033
ti(YearNum2,BLm)         2.100e-01     42   0.006   0.2738
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.675   Deviance explained = 68.5%
-REML = 939.92  Scale est. = 0.41245   n = 922
```

Figure 50: Summary of GAM model FE.BWt.M2, female animals captured in the East region.

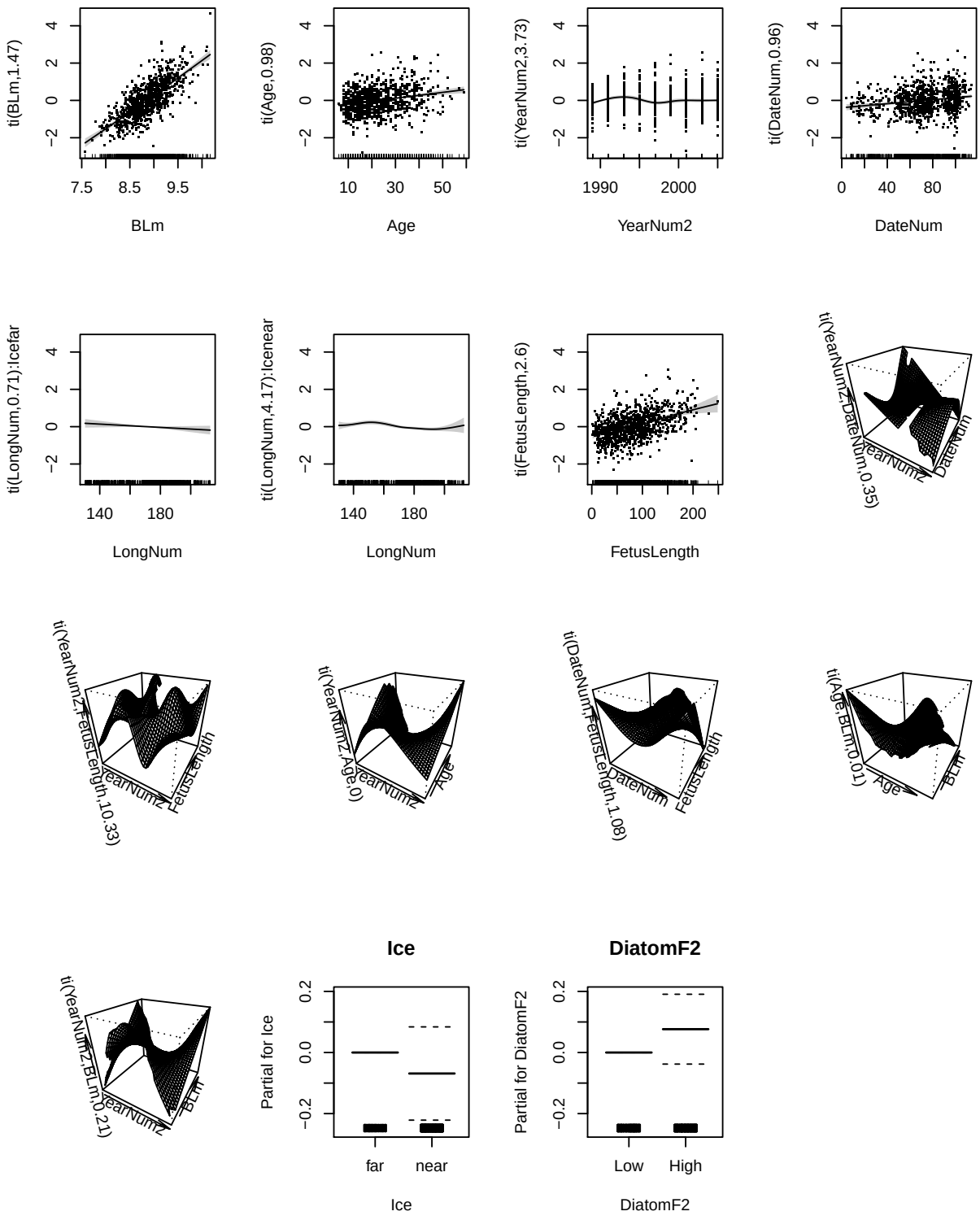

Figure 51: Partial effects plots for model FE.BWt.M2, gam fit to body weight for female animals captured in the East region.

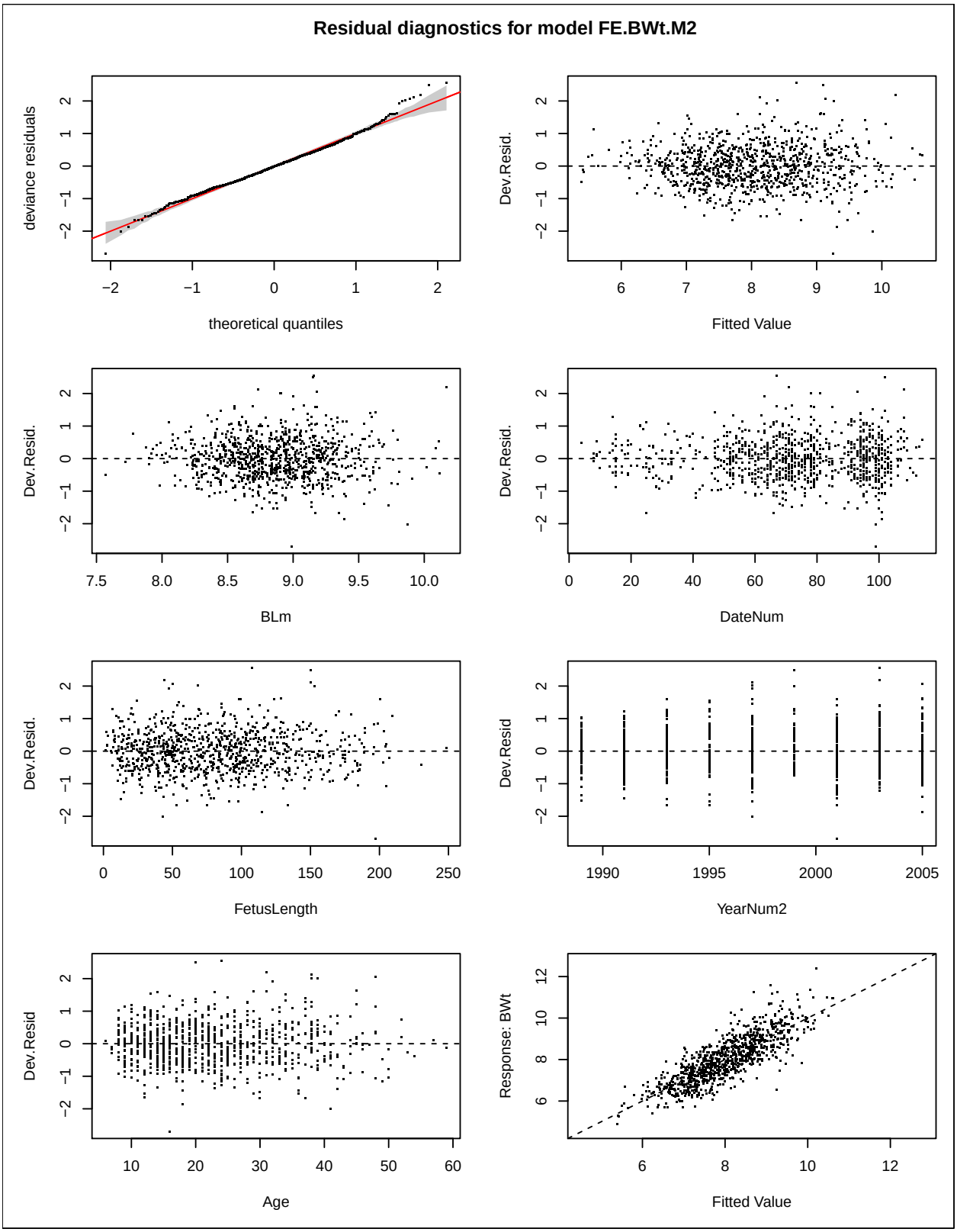

Figure 52: Residual diagnostics for model FE.BWt.M2 for female animals captured in the East region.

D.4 Females - East - Blubber thickness

Diagnostics are presented for gam models `FE.BT11.M2` and `FE.BT11.M3` of response `BT11`, first presented in Section B.4.

Model summaries for `FE.BT11.M2` and `FE.BT11.M3` are provided in Figures 53 and 54, respectively.

Partial effects plots are provided for models `FE.BT11.M2` and `FE.BT11.M3` in Figures 55 and 56.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth for models `FE.BT11.M2` and `FE.BT11.M3` proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for the sake of brevity.

A check of standard residual diagnostics for models `FE.BT11.M2` and `FE.BT11.M3` shows no appreciable lack-of-fit or departures from usual model assumptions (Figures 57 and 58).

Summary of model FE.BT11.M2

```
Family: gaussian
Link function: identity

Formula:
BT11 ~ ti(BLm, k = FE.k1) + ti(Age, k = FE.k1) + ti(YearNum2,
    k = FE.k2) + ti(DateNum, k = FE.k1) + Ice + ti(LongNum, k = FE.k3,
    by = Ice) + ti(FetusLength, k = FE.k1) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(FE.k2, FE.k1)) + ti(YearNum2, FetusLength,
    k = c(FE.k2, FE.k1)) + ti(YearNum2, Age, k = c(FE.k2, FE.k1)) +
    ti(DateNum, FetusLength, k = FE.k1) + ti(Age, BLm, k = FE.k1) +
    ti(YearNum2, BLm, k = c(FE.k2, FE.k1))

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.11077    0.09156  44.899   <2e-16 ***
Icenear      -0.04850    0.08914  -0.544   0.5865
DiatomF2High  0.16279    0.06646   2.450   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                               edf Ref.df      F  p-value
ti(BLm)                  3.074e+00      7  1.158 0.019131 *
ti(Age)                  1.611e-04      7  0.000 0.541314
ti(YearNum2)             4.081e+00      6  3.501 5.62e-05 ***
ti(DateNum)              3.401e+00      7  4.894 3.19e-09 ***
ti(LongNum):Icefar       5.899e-01     19  0.049 0.168631
ti(LongNum):Icenear      1.053e-03     19  0.000 0.438972
ti(FetusLength)          2.650e+00      7 25.013  < 2e-16 ***
ti(YearNum2,DateNum)     2.136e+00     42  0.103 0.051539 .
ti(YearNum2,FetusLength) 3.696e+00     42  0.246 0.006375 **
ti(YearNum2,Age)         3.573e+00     42  0.253 0.003250 **
ti(DateNum,FetusLength)  1.718e+00     49  0.114 0.014217 *
ti(Age,BLm)              7.891e+00     49  0.311 0.013176 *
ti(YearNum2,BLm)         1.147e+01     42  0.691 0.000276 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.432   Deviance explained = 46.1%
-REML =   1074  Scale est. = 0.53939   n = 922
```

Figure 53: Summary of GAM model FE.BT11.M2, female animals captured in the East region.

Summary of model FE.BT11.M3

```
Family: gaussian
Link function: identity

Formula:
BT11 ~ ti(BLm, k = FE.k1) + ti(Age, k = FE.k1) + ti(YearNum2,
    k = FE.k2) + ti(DateNum, k = FE.k1) + Ice + ti(LongNum, k = FE.k3,
    by = Ice) + ti(FetusLength, k = FE.k1) + ti(BWt, k = FE.k1) +
    DiatomF2 + ti(YearNum2, DateNum, k = c(FE.k2, FE.k1)) + ti(YearNum2,
    FetusLength, k = c(FE.k2, FE.k1)) + ti(YearNum2, Age, k = c(FE.k2,
    FE.k1)) + ti(DateNum, FetusLength, k = FE.k1) + ti(Age, BLm,
    k = FE.k1) + ti(YearNum2, BLm, k = c(FE.k2, FE.k1)) + ti(BWt,
    FetusLength, k = FE.k1)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.14517    0.08472  48.927   <2e-16 ***
Icenear      -0.04950    0.08214  -0.603   0.5469
DiatomF2High  0.12988    0.06060   2.143   0.0324 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                             edf Ref.df      F  p-value
ti(BLm)                 2.888e+00      7  9.871  < 2e-16 ***
ti(Age)                 1.840e+00      7  1.651 0.000415 ***
ti(YearNum2)            3.870e+00      6  3.739 1.96e-05 ***
ti(DateNum)             3.683e+00      7  4.380 3.11e-08 ***
ti(LongNum):Icefar      1.447e+00     19  0.279 0.015080 *
ti(LongNum):Icenear     9.129e-01     13  0.806 0.000215 ***
ti(FetusLength)         2.867e+00      7 11.019  < 2e-16 ***
ti(BWt)                 9.951e-01      7 29.149  < 2e-16 ***
ti(YearNum2,DateNum)    1.636e+00     42  0.051 0.183735
ti(YearNum2,FetusLength) 1.479e+00     42  0.077 0.072553 .
ti(YearNum2,Age)        3.575e+00     42  0.248 0.003851 **
ti(DateNum,FetusLength)  2.749e+00     49  0.097 0.072381 .
ti(Age,BLm)             3.005e+00     49  0.107 0.059198 .
ti(YearNum2,BLm)        1.005e+01     42  0.701 8.26e-05 ***
ti(BWt,FetusLength)     1.232e-04     49  0.000 0.664115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.526   Deviance explained = 54.8%
```

Figure 54: Summary of GAM model FE.BT11.M3, female animals captured in the East region. This model is identical to model FE-BT11-M2 with the exception for the addition of BWt as an explanatory variable of blubber thickness.

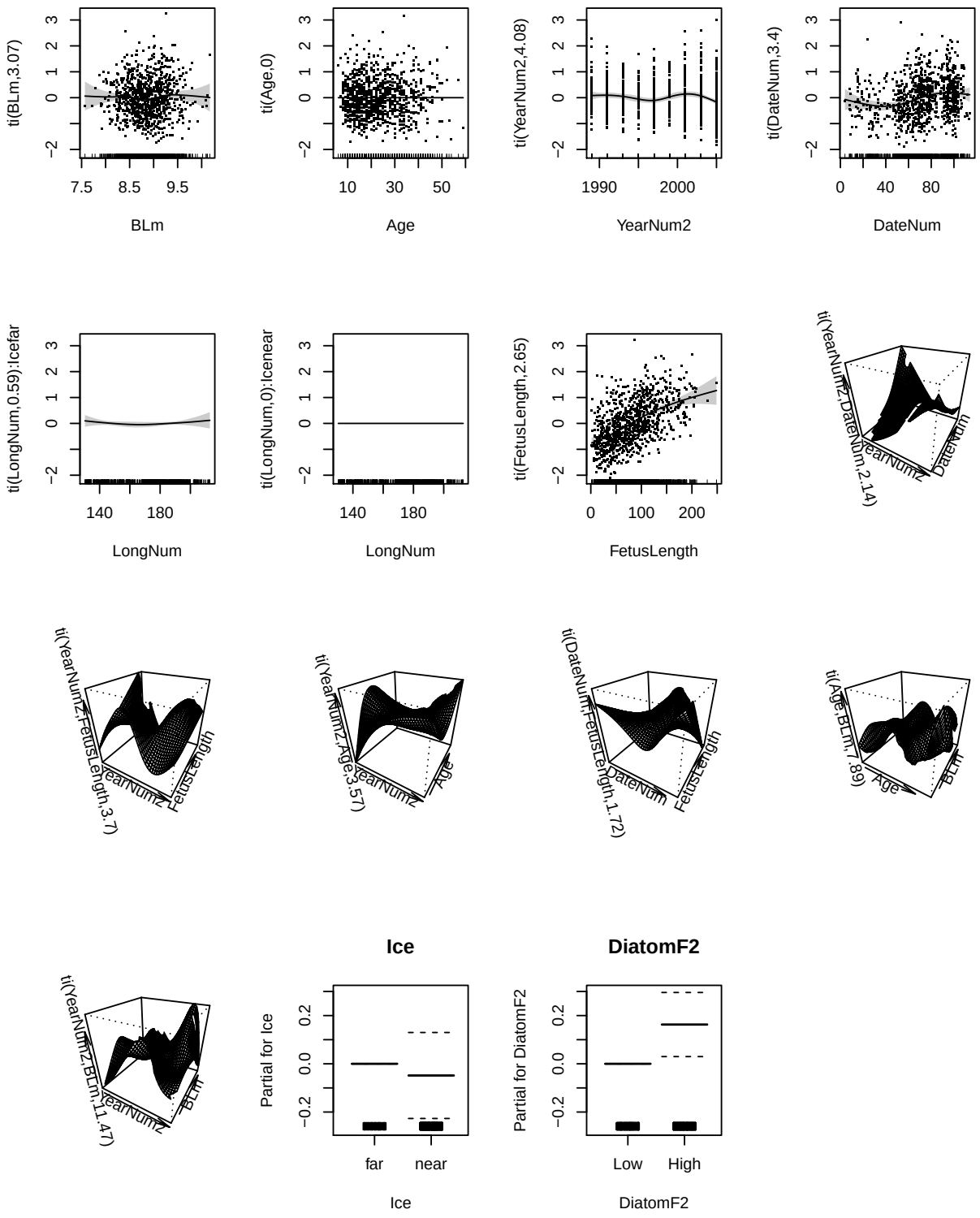

Figure 55: Partial effects plots for model FE.BT11.M2, gam fit to BT11 for female animals captured in the East region.

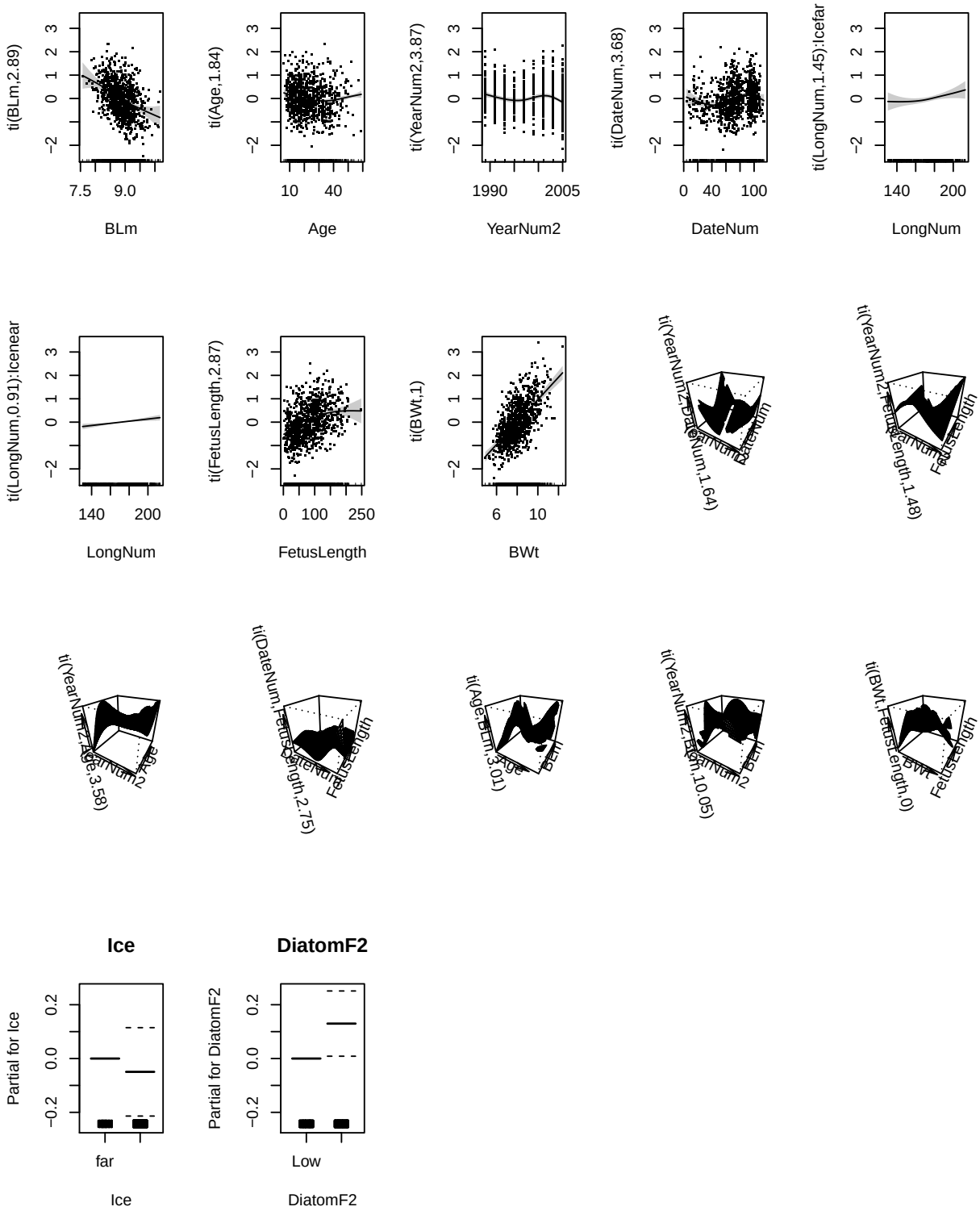

Figure 56: Partial effects plots for model FE.BT11.M3, gam fit to BT11 for female animals captured in the East region.

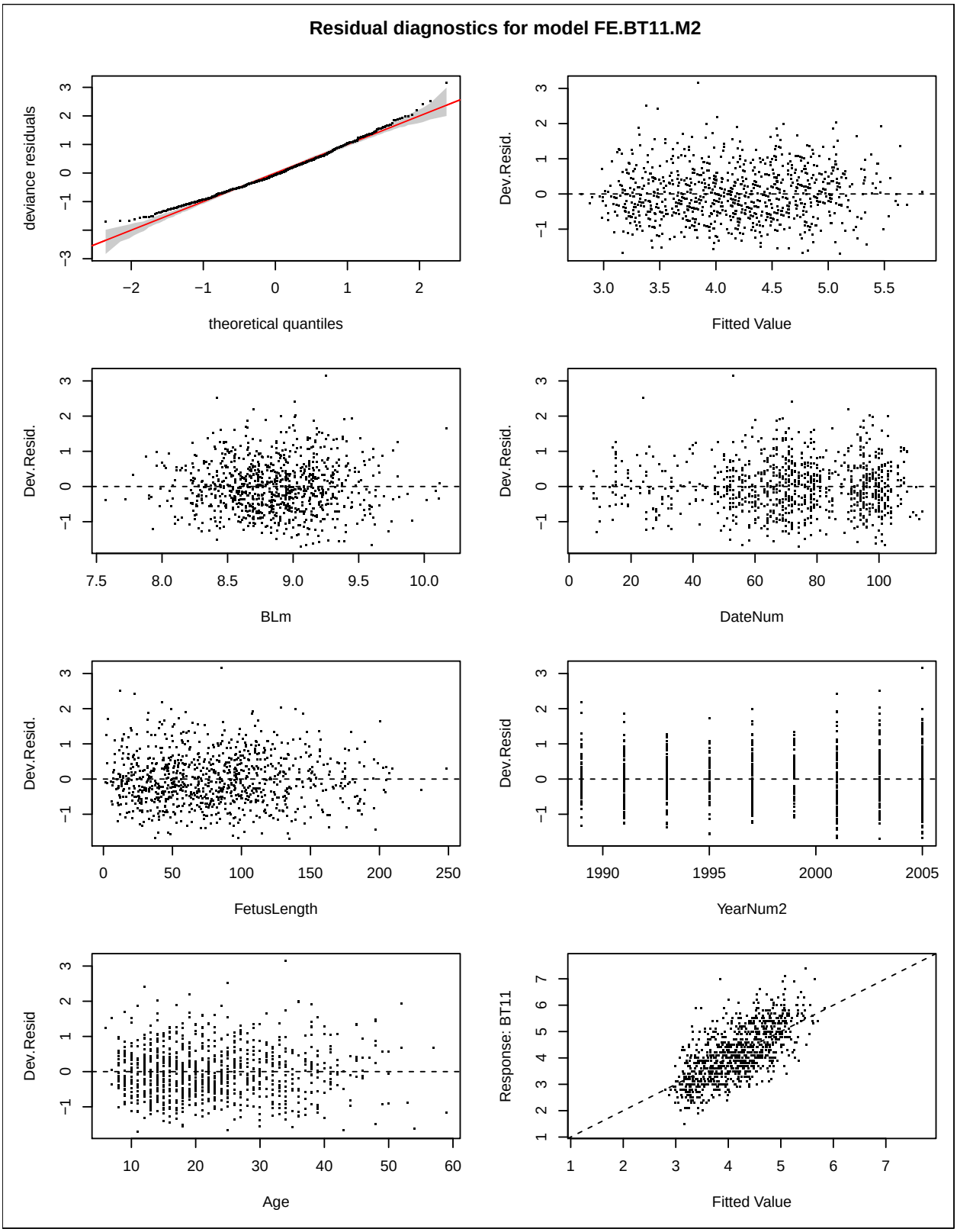

Figure 57: Residual diagnostics for model FE.BT11.M2 for female animals captured in the East region.

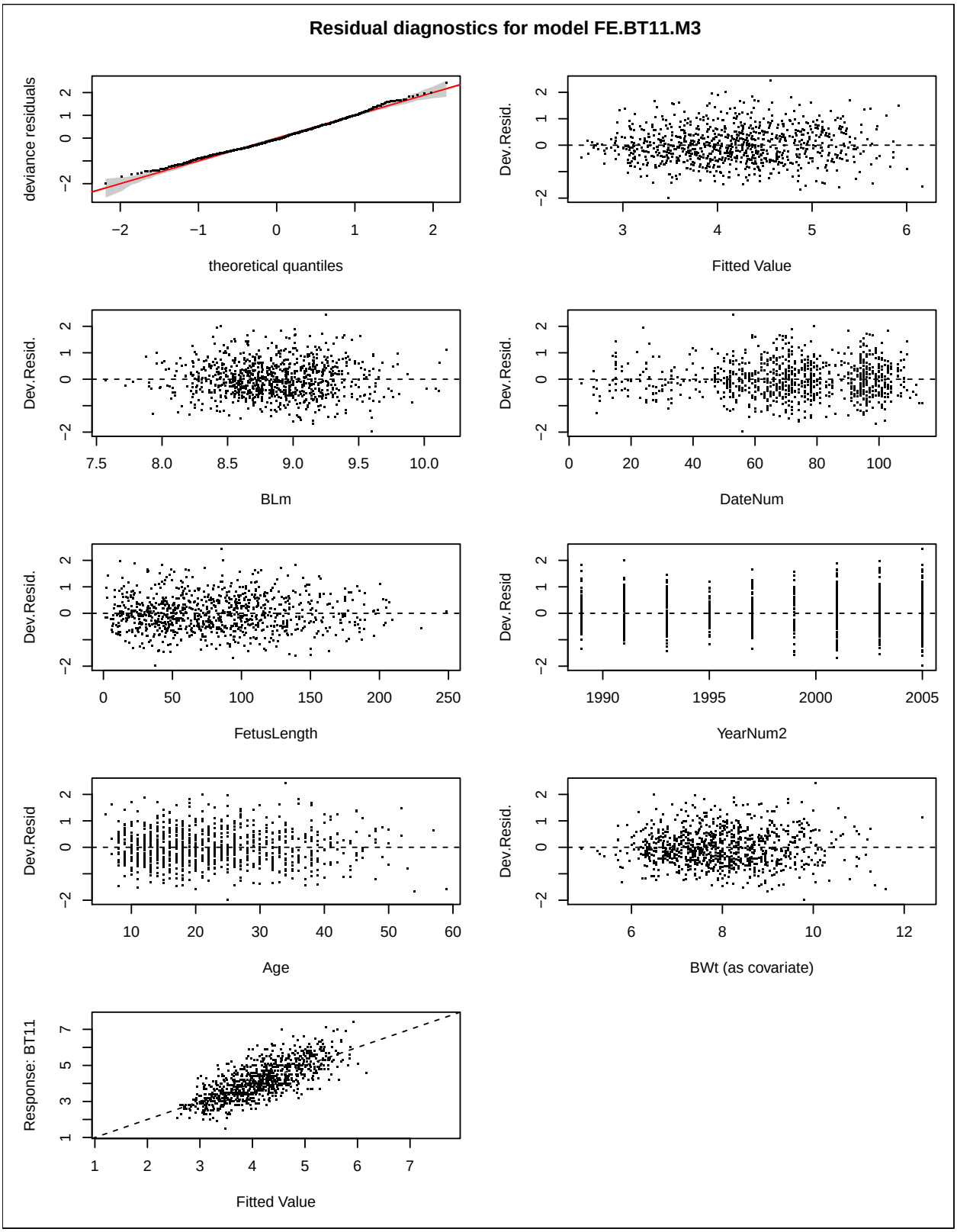

Figure 58: Residual diagnostics for model FE.BT11.M3 for female animals captured in the East region.

D.5 Males - West - Total weight

Diagnostics are presented for gam model `MW.BWt.M3` for body weight, first presented in Section C.1.

A model summary for `MW.BWt.M3` is provided in Figure 59.

Partial effects plots are provided for model `MW.BWt.M3` in Figure 60.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for brevity.

A check of standard residual diagnostics for model `MW.BWt.M3` revealed no appreciable lack-of-fit or departures from usual model assumptions (Figure 61).

Summary of model MW.BWt.M3

```
Family: gaussian
Link function: identity

Formula:
BWt ~ ti(BLm, k = MW.k1) + ti(Age, k = MW.k1) + ti(YearNum2,
    k = MW.k2, by = DiatomF2) + ti(DateNum, k = MW.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = MW.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(Age, BLm, k = MW.k1) +
    ti(YearNum2, BLm, k = c(MW.k2, MW.k1), by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.89931    0.03871 178.239  < 2e-16 ***
Icenear      -0.05749    0.03840  -1.497    0.135
DiatomF2High  0.18113    0.03643   4.971 7.63e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df       F  p-value
ti(BLm)                          3.8245106      7 155.026  < 2e-16 ***
ti(Age)                          3.3482158      7  20.492  < 2e-16 ***
ti(YearNum2):DiatomF2Low         2.7716077      6   2.773 0.000164 ***
ti(YearNum2):DiatomF2High        0.4963178      6   0.120 0.202513
ti(DateNum):DiatomF2Low          0.9783450      7   6.403 8.46e-12 ***
ti(DateNum):DiatomF2High         2.3307275      7  13.812  < 2e-16 ***
ti(LongNum):Icefar               1.1762236      7   1.098 0.003123 **
ti(LongNum):Icenear              0.9657436      7   1.248 0.001095 **
ti(YearNum2,DateNum):DiatomF2Low 0.5603925     42   0.015 0.314703
ti(YearNum2,DateNum):DiatomF2High 6.1099309    42   0.709 4.70e-06 ***
ti(YearNum2,Age):DiatomF2Low     0.4217443     41   0.015 0.209020
ti(YearNum2,Age):DiatomF2High    1.2371971     42   0.046 0.166386
ti(Age,BLm)                      0.4423555     47   0.013 0.189606
ti(YearNum2,BLm):DiatomF2Low     0.0003420     42   0.000 0.571188
ti(YearNum2,BLm):DiatomF2High    0.0002628     42   0.000 0.666280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =    0.7   Deviance explained = 70.6%
-REML = 1072.3  Scale est. = 0.32411   n = 1208
```

Figure 59: Summary of GAM model MW.BWt.M3, male animals captured in the West region.

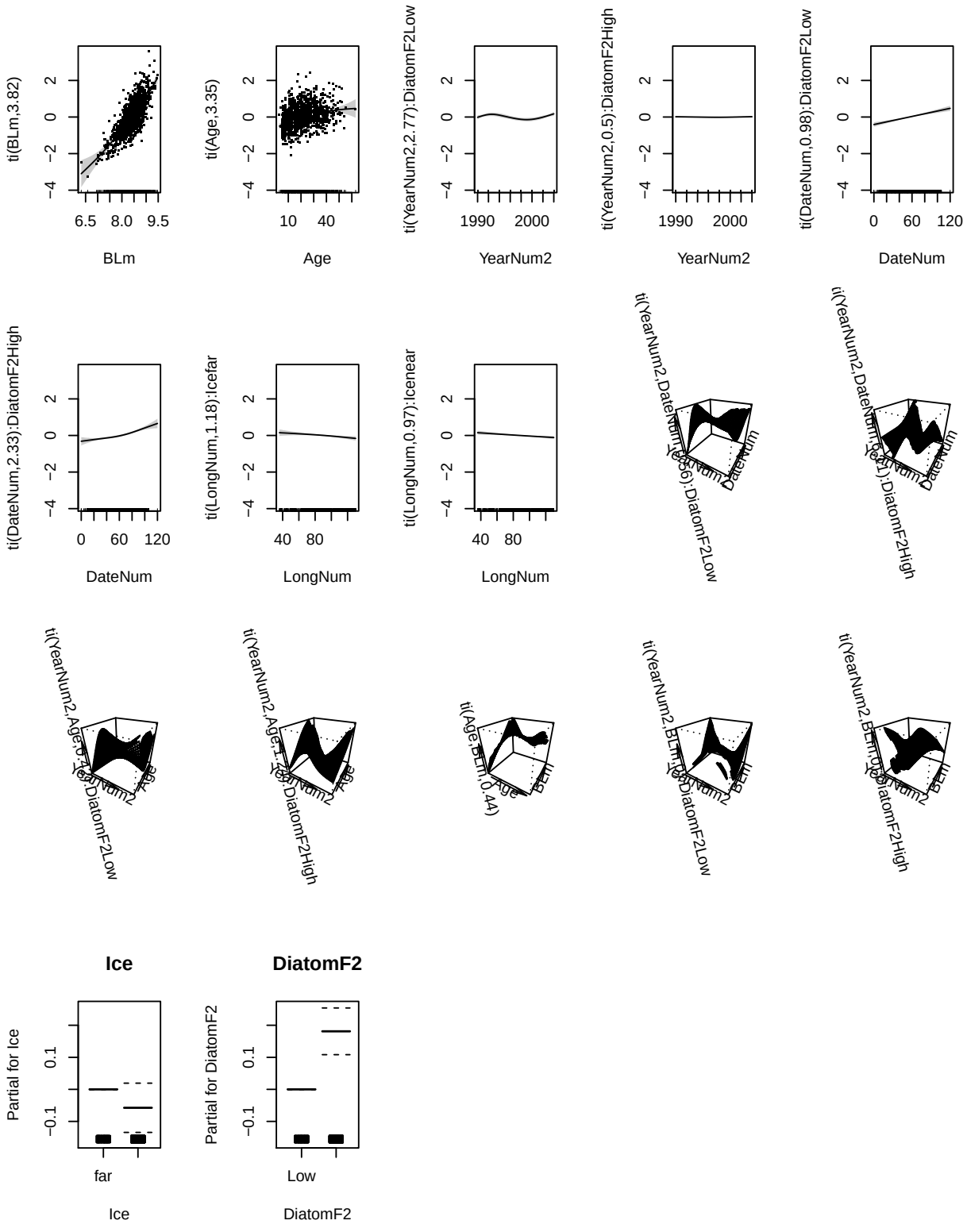

Figure 60: Partial effects plots for model MW.BWt.M3, gam fit to body weight for male animals captured in the West region.

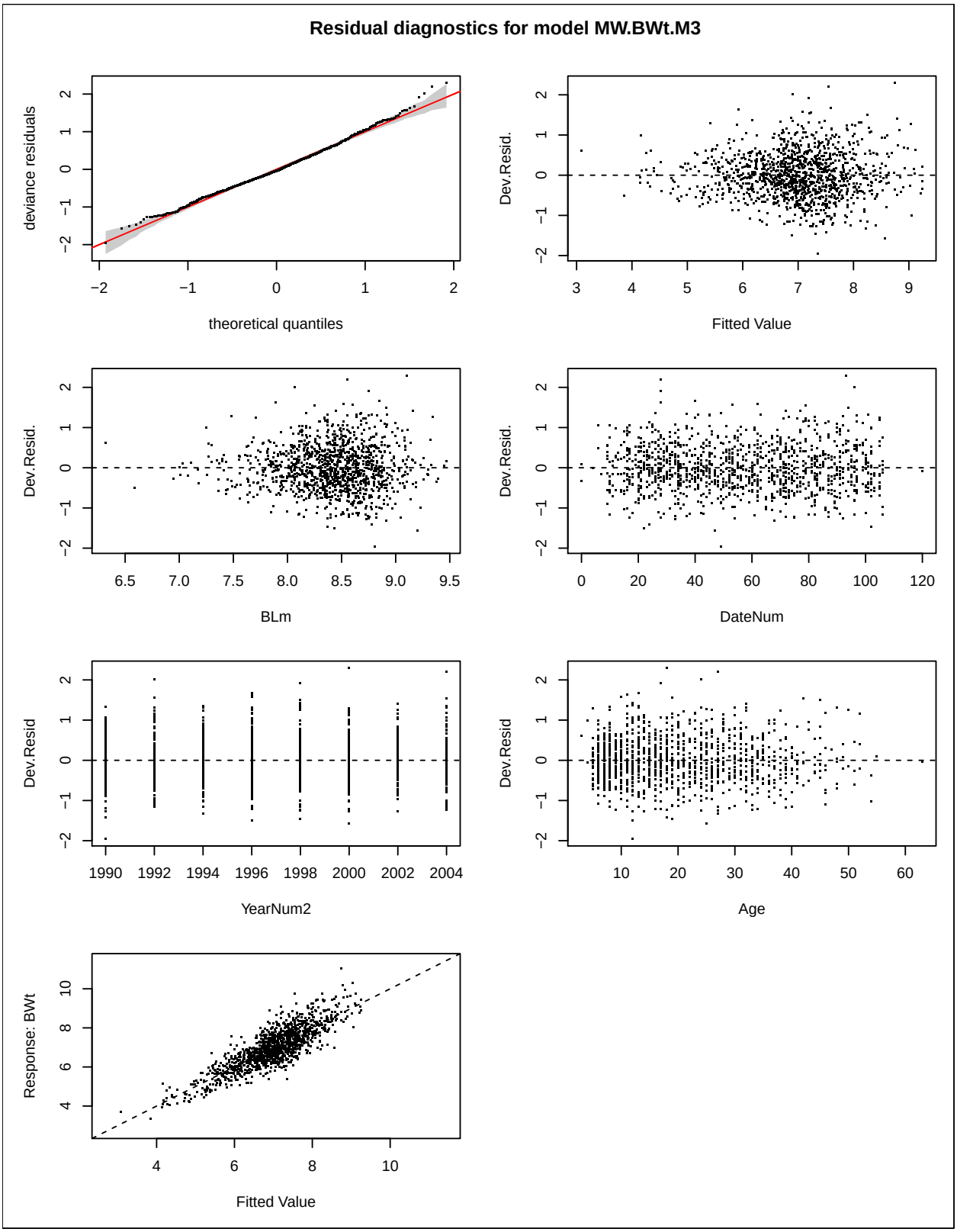

Figure 61: Residual diagnostics for model MW.BWt.M3 for male animals captured in the West region.

D.6 Males - West - Blubber thickness

Diagnostics are presented for gam models `MW.BT11.M3` and `MW.BT11.M4` for response `BT11`, first presented in Section C.2.

Model summaries for `MW.BT11.M3` and `MW.BT11.M4` are provided in Figures 62 and 63, respectively.

Partial effects plots are provided for models `MW.BT11.M3` and `MW.BT11.M4` in Figures 64 and 65.

A check of standard residual diagnostics for models `MW.BT11.M3` and `MW.BT11.M4` shows deviation from the assumed Gaussian error structure in the tails, as revealed by Q-Q plots, however this considered acceptable (if not perfect) given sample sizes and the degree of departure. No other lack-of-fit or departures from usual model assumptions seem apparent (Figures 66 and 67).

Summary of model MW.BT11.M3

```
Family: gaussian
Link function: identity

Formula:
BT11 ~ ti(BLm, k = MW.k1) + ti(Age, k = MW.k1) + ti(YearNum2,
    k = MW.k2, by = DiatomF2) + ti(DateNum, k = MW.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = MW.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(Age, BLm, k = MW.k1) +
    ti(YearNum2, BLm, k = c(MW.k2, MW.k1), by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.99951    0.04486  66.864   <2e-16 ***
Icenear       0.07600    0.04613   1.647   0.0997 .
DiatomF2High  0.37079    0.04304   8.615   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df      F  p-value
ti(BLm)                           2.031e+00      7  1.561 0.000817 ***
ti(Age)                           2.076e-04      7  0.000 0.654356
ti(YearNum2):DiatomF2Low          4.229e+00      6  6.449 1.12e-08 ***
ti(YearNum2):DiatomF2High         2.874e+00      6  5.304 9.41e-09 ***
ti(DateNum):DiatomF2Low           1.912e+00      7 13.009  < 2e-16 ***
ti(DateNum):DiatomF2High          2.008e+00      7 35.125  < 2e-16 ***
ti(LongNum):Icefar                1.864e-04      7  0.000 0.588235
ti(LongNum):Icenear               2.042e+00      7  1.120 0.006392 **
ti(YearNum2,DateNum):DiatomF2Low  1.894e-01     42  0.006 0.255789
ti(YearNum2,DateNum):DiatomF2High 1.191e+01     42  0.852 2.03e-05 ***
ti(YearNum2,Age):DiatomF2Low      3.292e-01     42  0.011 0.232200
ti(YearNum2,Age):DiatomF2High     9.893e-01     42  0.029 0.246681
ti(Age,BLm)                       1.669e+00     47  0.055 0.105965
ti(YearNum2,BLm):DiatomF2Low      4.251e-04     42  0.000 0.864111
ti(YearNum2,BLm):DiatomF2High     2.202e+00     39  0.120 0.080209 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.346   Deviance explained = 36.4%
-REML = 1264.9  Scale est. = 0.4426    n = 1208
```

Figure 62: Summary of GAM model MW.BT11.M3, male animals captured in the West region.

```
Family: gaussian
Link function: identity
```

Summary of model MW.BT11.M4

```
Formula:
BT11 ~ ti(BLm, k = MW.k1) + ti(Age, k = MW.k1) + ti(YearNum2,
    k = MW.k2, by = DiatomF2) + ti(DateNum, k = MW.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = MW.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(Age, BLm, k = MW.k1) +
    ti(YearNum2, BLm, k = c(MW.k2, MW.k1), by = DiatomF2) + ti(BWt,
    k = MW.k1, by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.03191    0.04168  72.740  < 2e-16 ***
Icenear       0.09783    0.04268   2.292   0.0221 *
DiatomF2High  0.29433    0.04050   7.267 6.69e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df      F  p-value
ti(BLm)                           2.073e+00      7  9.142  < 2e-16 ***
ti(Age)                           2.887e+00      7  3.557 1.08e-06 ***
ti(YearNum2):DiatomF2Low          3.081e+00      6  4.709 3.37e-07 ***
ti(YearNum2):DiatomF2High         2.809e+00      6  6.644 8.62e-11 ***
ti(DateNum):DiatomF2Low           1.647e+00      7  8.096 3.82e-15 ***
ti(DateNum):DiatomF2High          9.946e-01      7 25.025  < 2e-16 ***
ti(LongNum):Icefar                4.045e-04      7  0.000  1.00000
ti(LongNum):Icenear               2.207e+00      7  1.252  0.00478 **
ti(YearNum2,DateNum):DiatomF2Low  1.254e+00     42  0.042  0.19978
ti(YearNum2,DateNum):DiatomF2High 1.092e+01     42  0.883 4.50e-06 ***
ti(YearNum2,Age):DiatomF2Low      3.019e-01     42  0.009  0.23901
ti(YearNum2,Age):DiatomF2High     1.586e-03     42  0.000  0.40746
ti(Age,BLm)                       1.124e+00     47  0.031  0.17555
ti(YearNum2,BLm):DiatomF2Low      6.268e-04     42  0.000  0.87827
ti(YearNum2,BLm):DiatomF2High     2.278e+00     42  0.146  0.03997 *
ti(BWt):DiatomF2Low               9.897e-01      7 13.580  < 2e-16 ***
ti(BWt):DiatomF2High              9.947e-01      7 25.920  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.433   Deviance explained =   45%
```

Figure 63: Summary of GAM model MW.BT11.M4, male animals captured in the West region. This model is identical to model MW-BT11-M3 with the exception for the addition of BWt as an explanatory variable of blubber thickness.

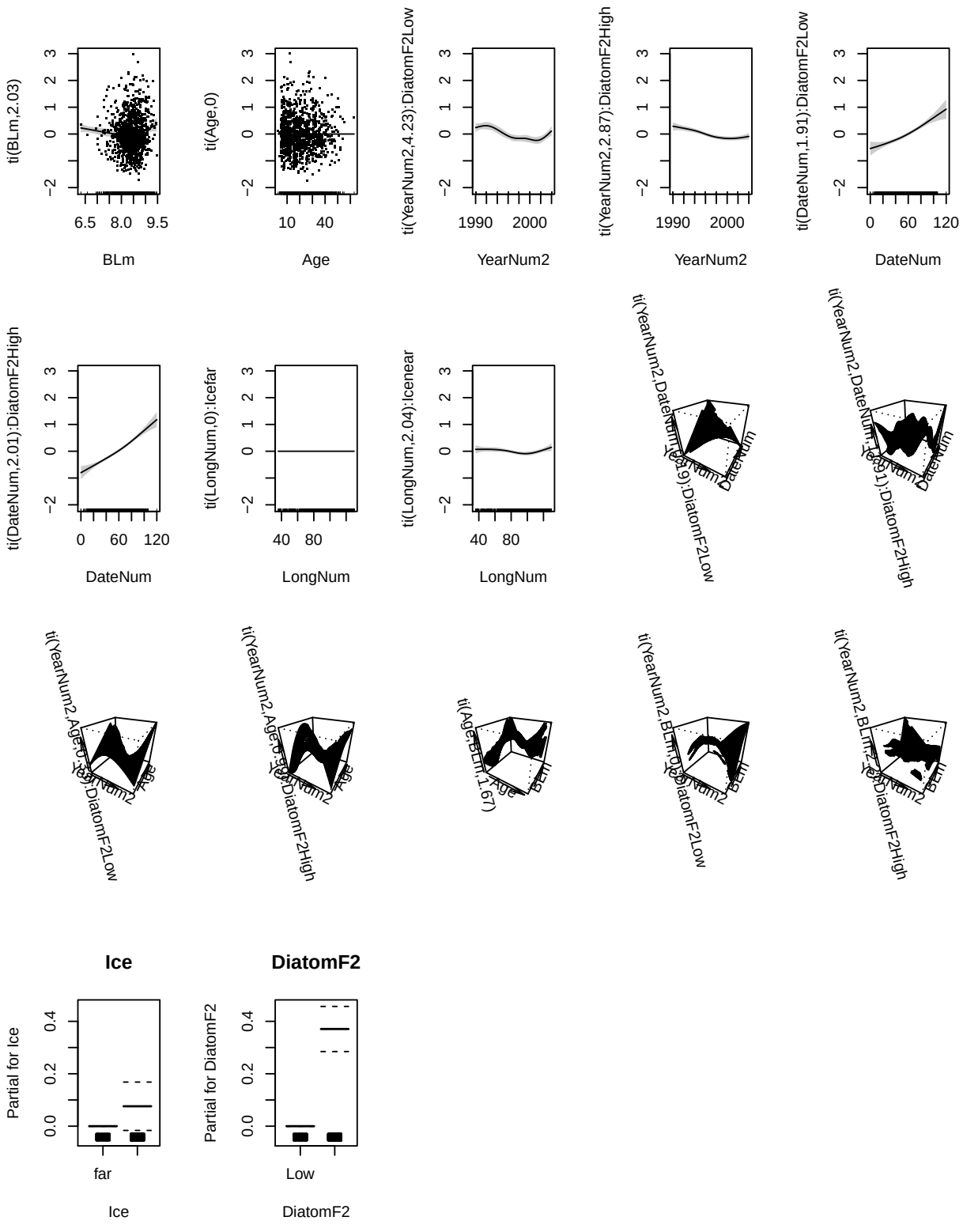

Figure 64: Partial effects plots for model MW.BT11.M3, gam fit to BT11 for male animals captured in the West region.

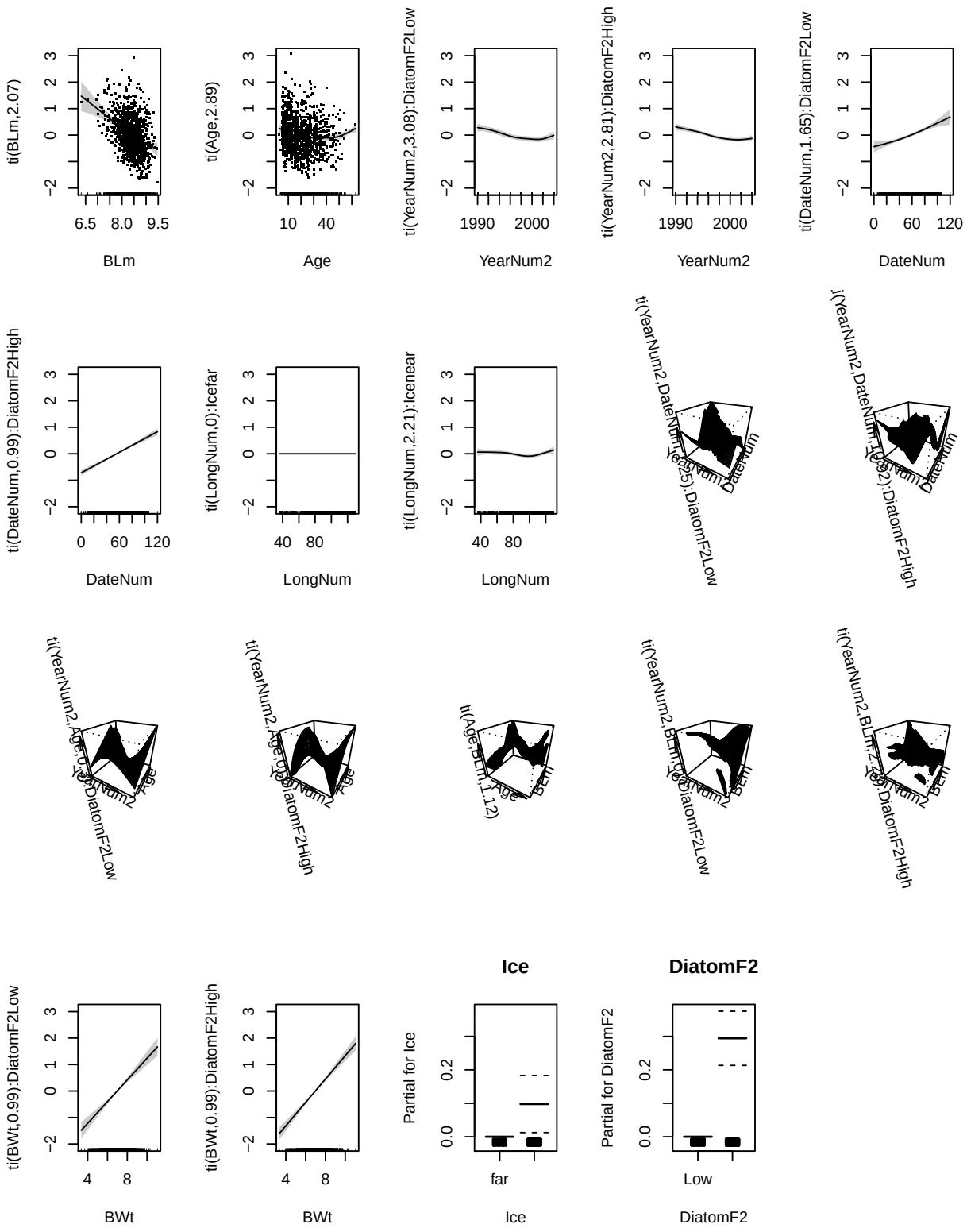

Figure 65: Partial effects plots for model MW.BT11.M4, gam fit to BT11 for male animals captured in the West region.

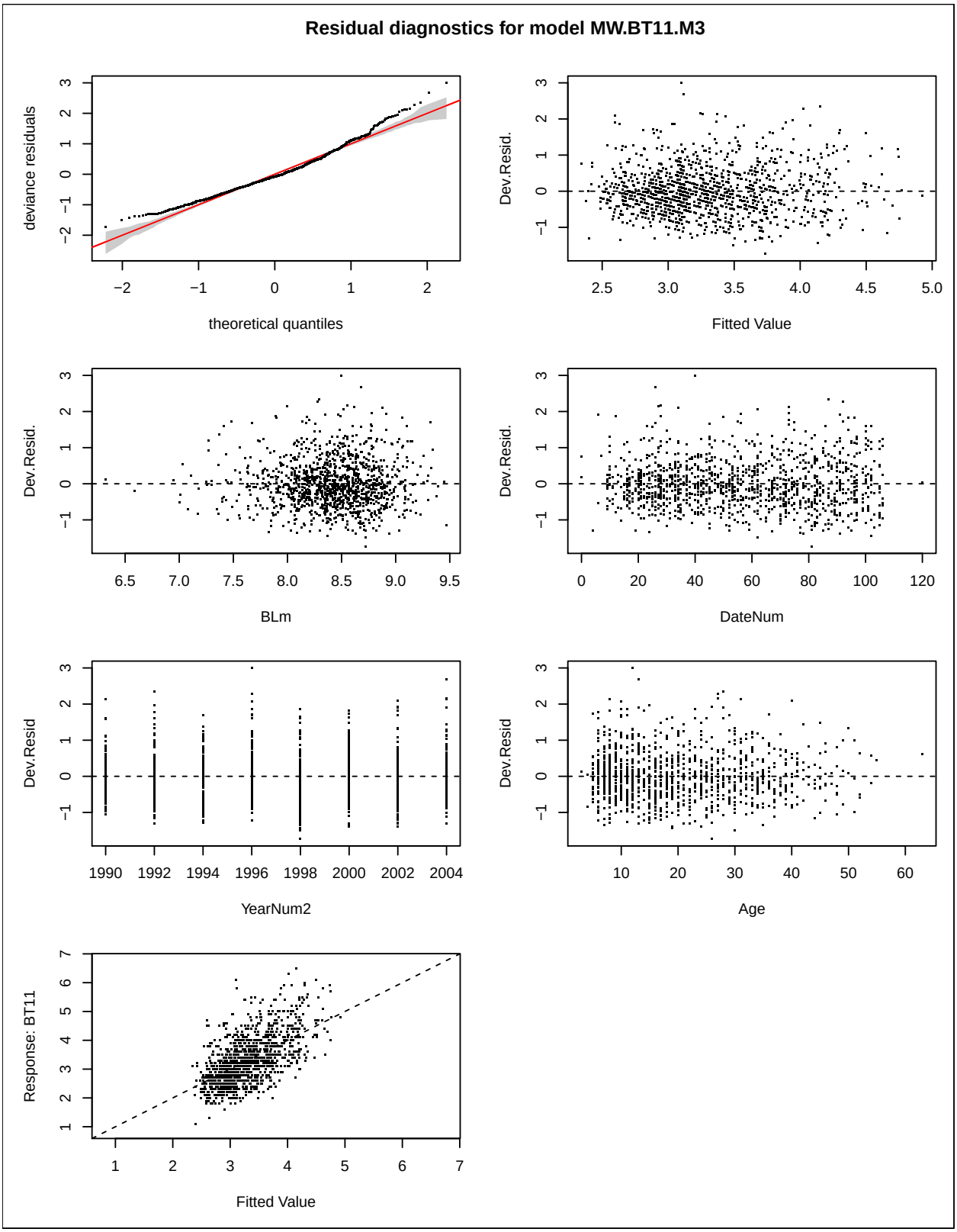

Figure 66: Residual diagnostics for model MW.BT11.M3 for male animals captured in the West region.

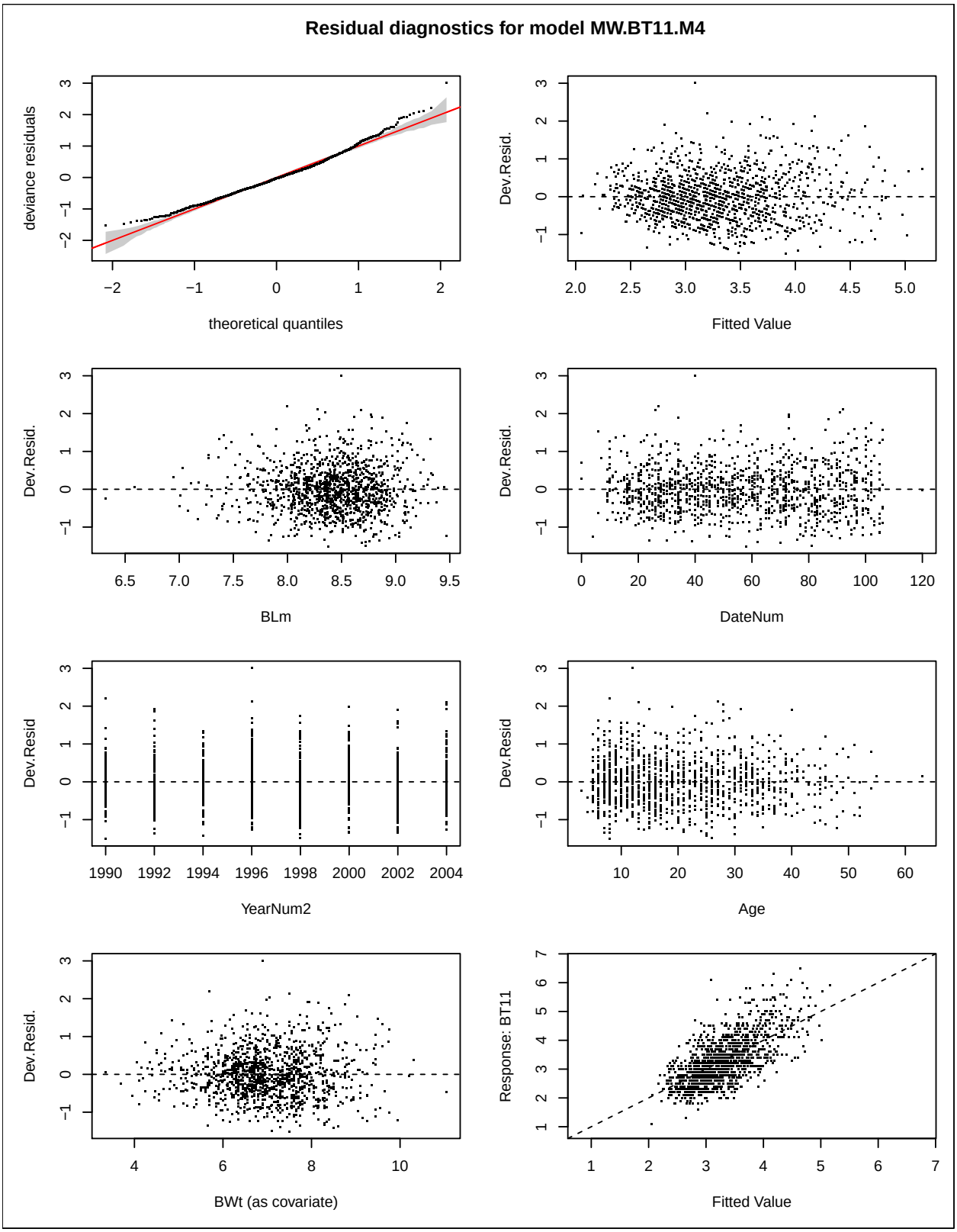

Figure 67: Residual diagnostics for model MW.BT11.M4 for male animals captured in the West region.

D.7 Males - East - Total weight

Diagnostics are presented for gam model `ME.BWt.M3` of body weight, first presented in Section C.3.

A model summary for `ME.BWt.M3` is provided in Figure 68.

Partial effects plots are provided for model `ME.BWt.M3` in Figure 69.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for brevity.

A check of standard residual diagnostics for model `ME.BWt.M3` revealed no appreciable lack-of-fit or departures from usual model assumptions (Figure 70).

Summary of model ME.BWt.M3

```
Family: gaussian
Link function: identity

Formula:
BWt ~ ti(BLm, k = ME.k1) + ti(Age, k = ME.k1) + ti(YearNum2,
    k = ME.k2, by = DiatomF2) + ti(DateNum, k = ME.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = ME.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(Age, BLm, k = ME.k1) +
    ti(YearNum2, BLm, k = c(ME.k2, ME.k1), by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.57494    0.03442 191.046  < 2e-16 ***
Icenear       0.04382    0.03631   1.207    0.228
DiatomF2High  0.23043    0.03483   6.617 5.36e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df       F  p-value
ti(BLm)                          9.989e-01      7 128.576  < 2e-16 ***
ti(Age)                          3.808e+00      7  17.334  < 2e-16 ***
ti(YearNum2):DiatomF2Low         4.815e-05      6   0.000  0.72447
ti(YearNum2):DiatomF2High        3.497e+00      6   4.779 7.14e-07 ***
ti(DateNum):DiatomF2Low          1.808e+00      7   8.791  < 2e-16 ***
ti(DateNum):DiatomF2High         1.054e+00      7  24.971  < 2e-16 ***
ti(LongNum):Icefar               7.413e-01      7   0.409  0.04479 *
ti(LongNum):Icenear              4.292e+00      7   5.368 1.49e-08 ***
ti(YearNum2,DateNum):DiatomF2Low 1.891e+00     42   0.227  0.00219 **
ti(YearNum2,DateNum):DiatomF2High 1.398e-01     42   0.004  0.28419
ti(YearNum2,Age):DiatomF2Low     6.910e-05     42   0.000  1.00000
ti(YearNum2,Age):DiatomF2High    8.409e-05     42   0.000  0.69530
ti(Age,BLm)                      2.491e+00     41   0.158  0.03588 *
ti(YearNum2,BLm):DiatomF2Low     7.332e-01     42   0.065  0.05178 .
ti(YearNum2,BLm):DiatomF2High    6.079e-01     40   0.039  0.11005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.696   Deviance explained = 70.1%
-REML = 1158.4  Scale est. = 0.32163   n = 1316
```

Figure 68: Summary of GAM model ME.BWt.M3, male animals captured in the East region.

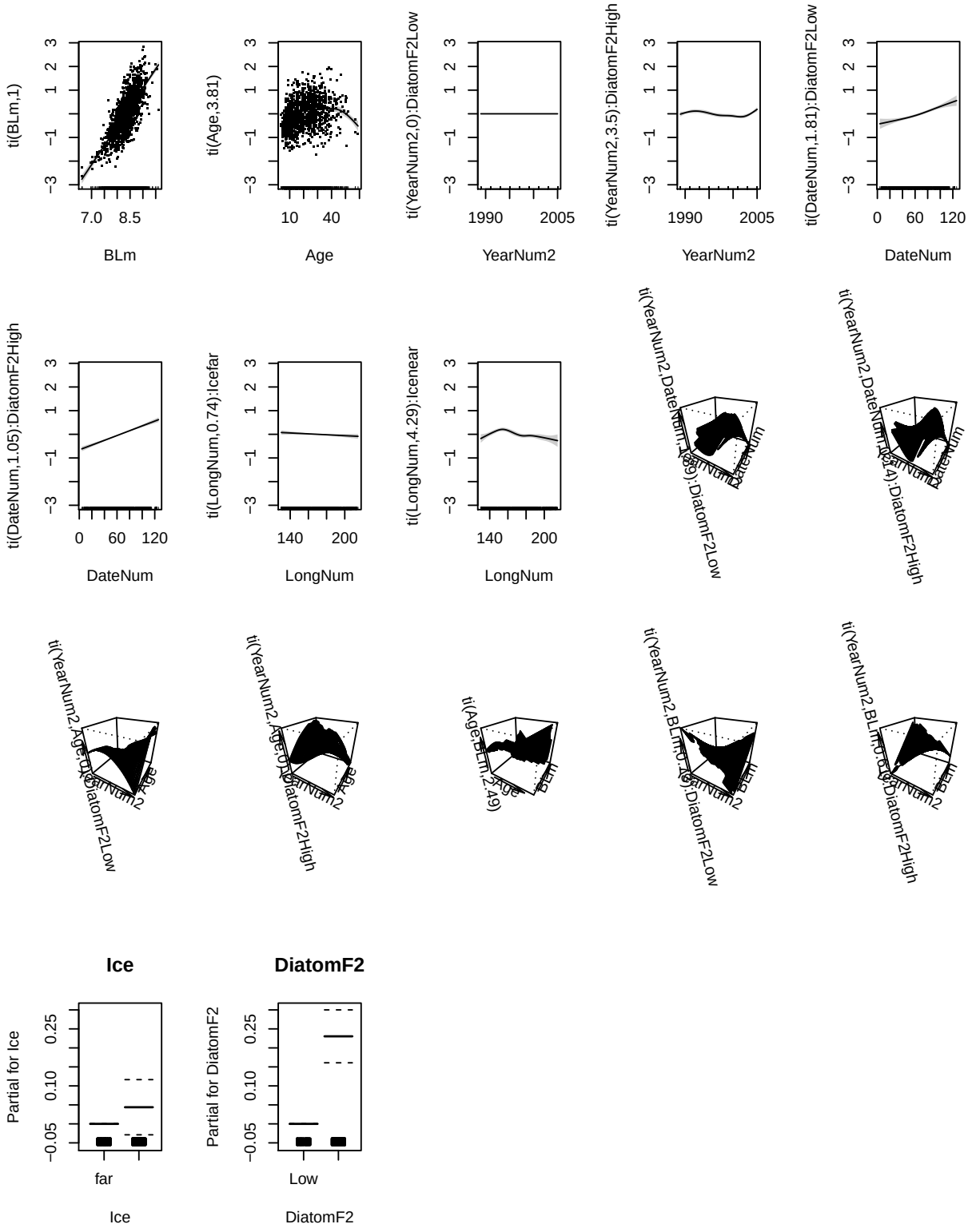

Figure 69: Partial effects plots for model ME.BWt.M3, gam fit to body weight for male animals captured in the East region.

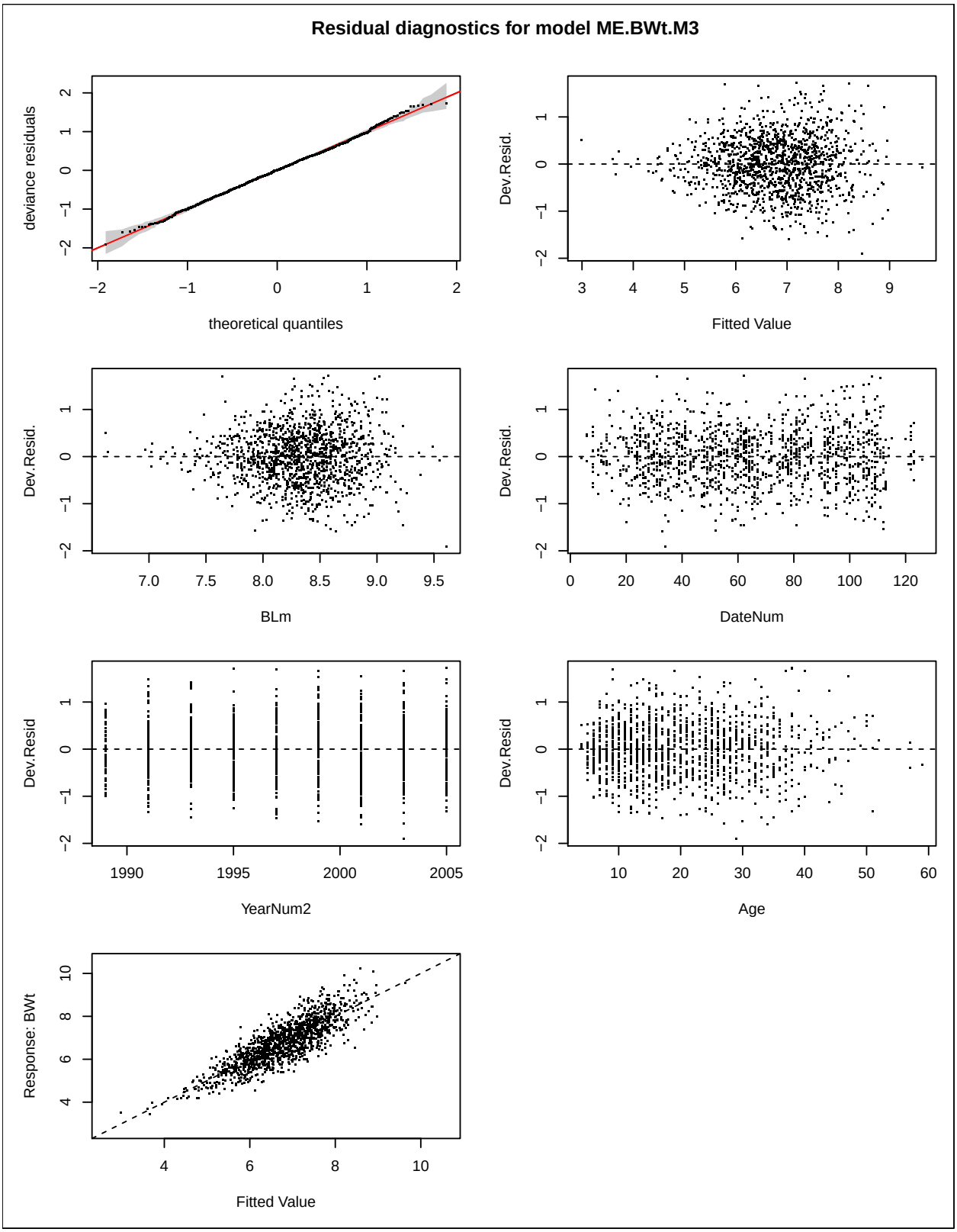

Figure 70: Residual diagnostics for model ME.BWt.M3 for male animals captured in the East region.

D.8 Males - East - Blubber thickness

Diagnostics are presented for gam models `ME.BT11.M3` and `ME.BT11.M4` of response `BT11`, first presented in Section C.4.

Summaries for models `ME.BT11.M3` and `ME.BT11.M4` are provided in Figures 71 and 72, respectively.

Partial effects plots are provided for models `ME.BT11.M3` and `ME.BT11.M4` in Figures 73 and 74.

Diagnostics assessing the sufficiency of the allowed basis dimension per smooth proved satisfactory, as did a check of the concurvity between model terms (the gam equivalent of colinearity in linear models). These results are not shown for brevity.

A check of standard residual diagnostics for models `ME.BT11.M3` and `ME.BT11.M4` shows, like almost all other models for `BT11`, some slight departures from distributional assumptions as evidenced in the tails of Q-Q plots. The size of the data set compared with the number of points exhibiting this behaviour leaves us confident this is not cause for concern. No other lack of fit was detected from those diagnostics examined.

Summary of model ME.BT11.M3

```
Family: gaussian
Link function: identity

Formula:
BT11 ~ ti(BLm, k = ME.k1) + ti(Age, k = ME.k1) + ti(YearNum2,
    k = ME.k2, by = DiatomF2) + ti(DateNum, k = ME.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = ME.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(Age, BLm, k = ME.k1) +
    ti(YearNum2, BLm, k = c(ME.k2, ME.k1), by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.35550    0.04236  79.211  < 2e-16 ***
Icenear       0.13223    0.04796   2.757  0.00591 **
DiatomF2High  0.48449    0.04367  11.095  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df      F  p-value
ti(BLm)                          0.3682393      7  0.064  0.26256
ti(Age)                          1.3356137      7  0.907  0.00740 **
ti(YearNum2):DiatomF2Low         0.0011358      6  0.000  0.56894
ti(YearNum2):DiatomF2High        2.4079392      6  0.936  0.05187 .
ti(DateNum):DiatomF2Low          2.5258041      7 21.513  < 2e-16 ***
ti(DateNum):DiatomF2High         4.0709952      7 48.404  < 2e-16 ***
ti(LongNum):Icefar               1.8731282      7  0.996  0.00605 **
ti(LongNum):Icenear              2.9377407      7  2.985 5.25e-06 ***
ti(YearNum2,DateNum):DiatomF2Low  2.4016145     42  0.179  0.01888 *
ti(YearNum2,DateNum):DiatomF2High 7.8672692     42  0.639 3.14e-05 ***
ti(YearNum2,Age):DiatomF2Low     0.0006954     42  0.000  0.84962
ti(YearNum2,Age):DiatomF2High    0.0017020     42  0.000  0.46426
ti(Age,BLm)                      0.0006066     48  0.000  0.93947
ti(YearNum2,BLm):DiatomF2Low     0.1544810     42  0.004  0.27512
ti(YearNum2,BLm):DiatomF2High    2.4325705     42  0.134  0.03322 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.468   Deviance explained =   48%
-REML = 1427.3  Scale est. = 0.48321   n = 1316
```

Figure 71: Summary of GAM model ME.BT11.M3, male animals captured in the East region.

```
Family: gaussian
Link function: identity
```

Summary of model ME.BT11.M4

```
Formula:
BT11 ~ ti(BLm, k = ME.k1) + ti(Age, k = ME.k1) + ti(YearNum2,
    k = ME.k2, by = DiatomF2) + ti(DateNum, k = ME.k1, by = DiatomF2) +
    Ice + ti(LongNum, k = ME.k1, by = Ice) + DiatomF2 + ti(YearNum2,
    DateNum, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(YearNum2,
    Age, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(Age, BLm, k = ME.k1) +
    ti(YearNum2, BLm, k = c(ME.k2, ME.k1), by = DiatomF2) + ti(BWt,
    k = ME.k1, by = DiatomF2)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.42138    0.03876  88.263  < 2e-16 ***
Icenear       0.14012    0.04374   3.203  0.00139 **
DiatomF2High  0.38608    0.04055   9.521  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                       edf Ref.df      F  p-value
ti(BLm)                          0.9873308      7 10.865  < 2e-16 ***
ti(Age)                          0.8334599      7  0.714  0.01387 *
ti(YearNum2):DiatomF2Low         0.3802272      6  0.099  0.20553
ti(YearNum2):DiatomF2High        3.6849158      6  3.159  0.00015 ***
ti(DateNum):DiatomF2Low          1.1302424      7 14.647  < 2e-16 ***
ti(DateNum):DiatomF2High         3.7245381      7 30.680  < 2e-16 ***
ti(LongNum):Icefar               1.7582513      7  2.471 9.72e-06 ***
ti(LongNum):Icenear              2.1230788      7  4.236 6.33e-09 ***
ti(YearNum2,DateNum):DiatomF2Low  0.0012789     42  0.000  0.69356
ti(YearNum2,DateNum):DiatomF2High 6.1079321     42  0.570 3.18e-05 ***
ti(YearNum2,Age):DiatomF2Low     0.0009582     42  0.000  0.66975
ti(YearNum2,Age):DiatomF2High    0.5611445     42  0.023  0.15402
ti(Age,BLm)                      0.0006505     48  0.000  0.83066
ti(YearNum2,BLm):DiatomF2Low     0.7962154     42  0.063  0.05918 .
ti(YearNum2,BLm):DiatomF2High    4.6318492     42  0.252  0.01140 *
ti(BWt):DiatomF2Low              2.4702257      7 16.265  < 2e-16 ***
ti(BWt):DiatomF2High             2.2049634      7 24.679  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.539   Deviance explained = 55.1%
```

Figure 72: Summary of GAM model ME.BT11.M4, male animals captured in the East region. This model is identical to model ME-BT11-M3 with the exception for the addition of BWt as an explanatory variable of blubber thickness.

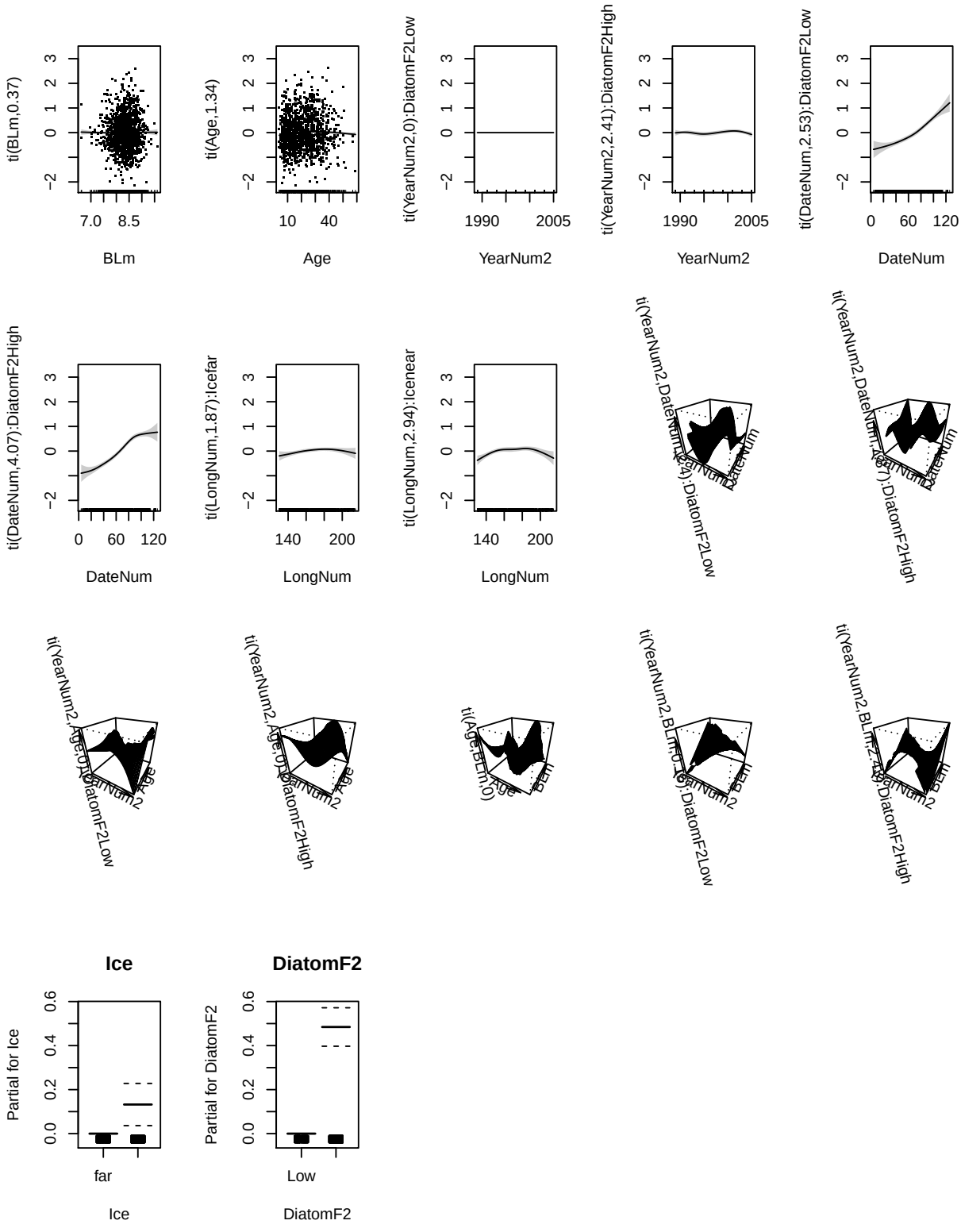

Figure 73: Partial effects plots for model ME.BT11.M3, gam fit to BT11 for male animals captured in the East region.

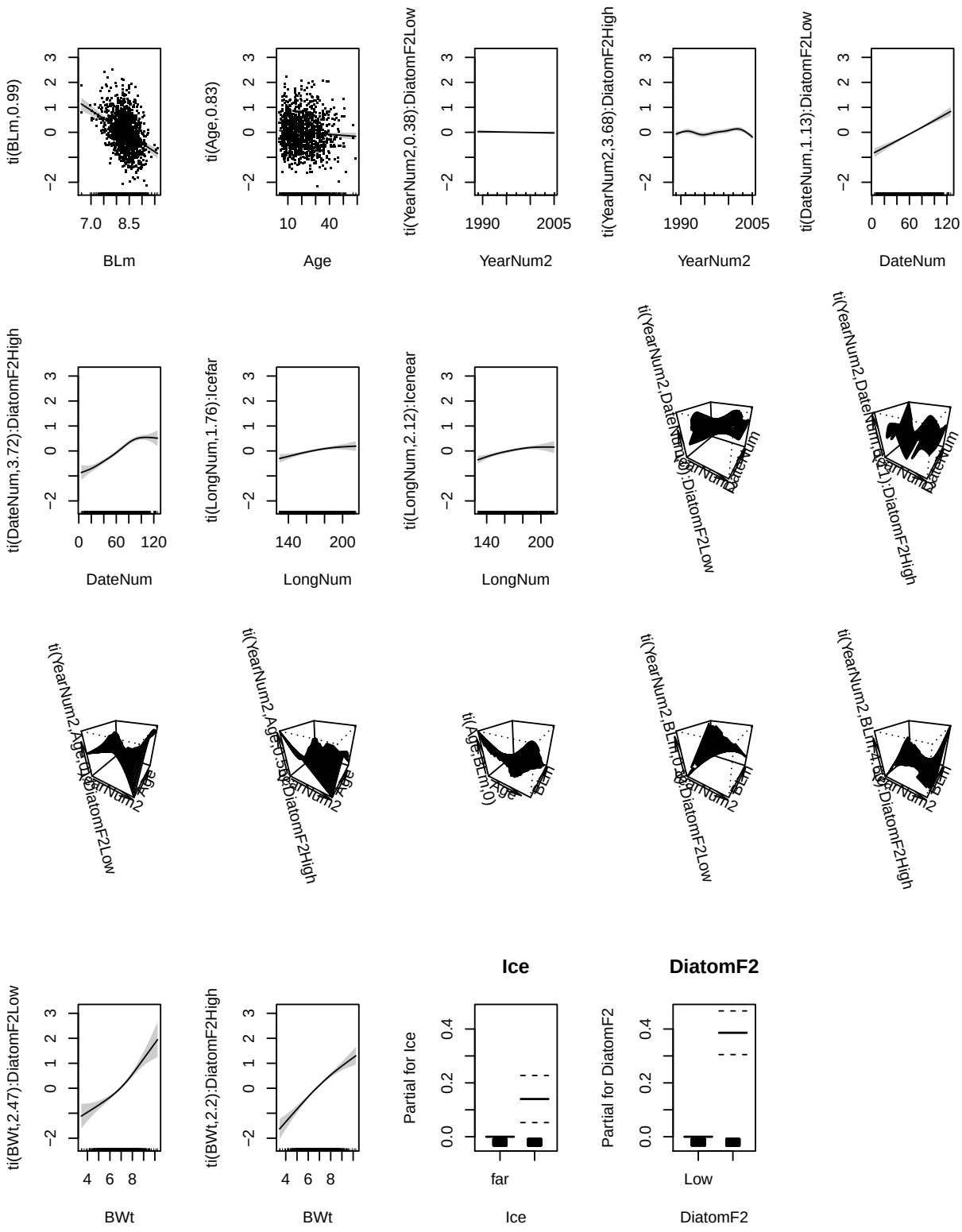

Figure 74: Partial effects plots for model ME.BT11.M4, gam fit to BT11 for male animals captured in the East region.

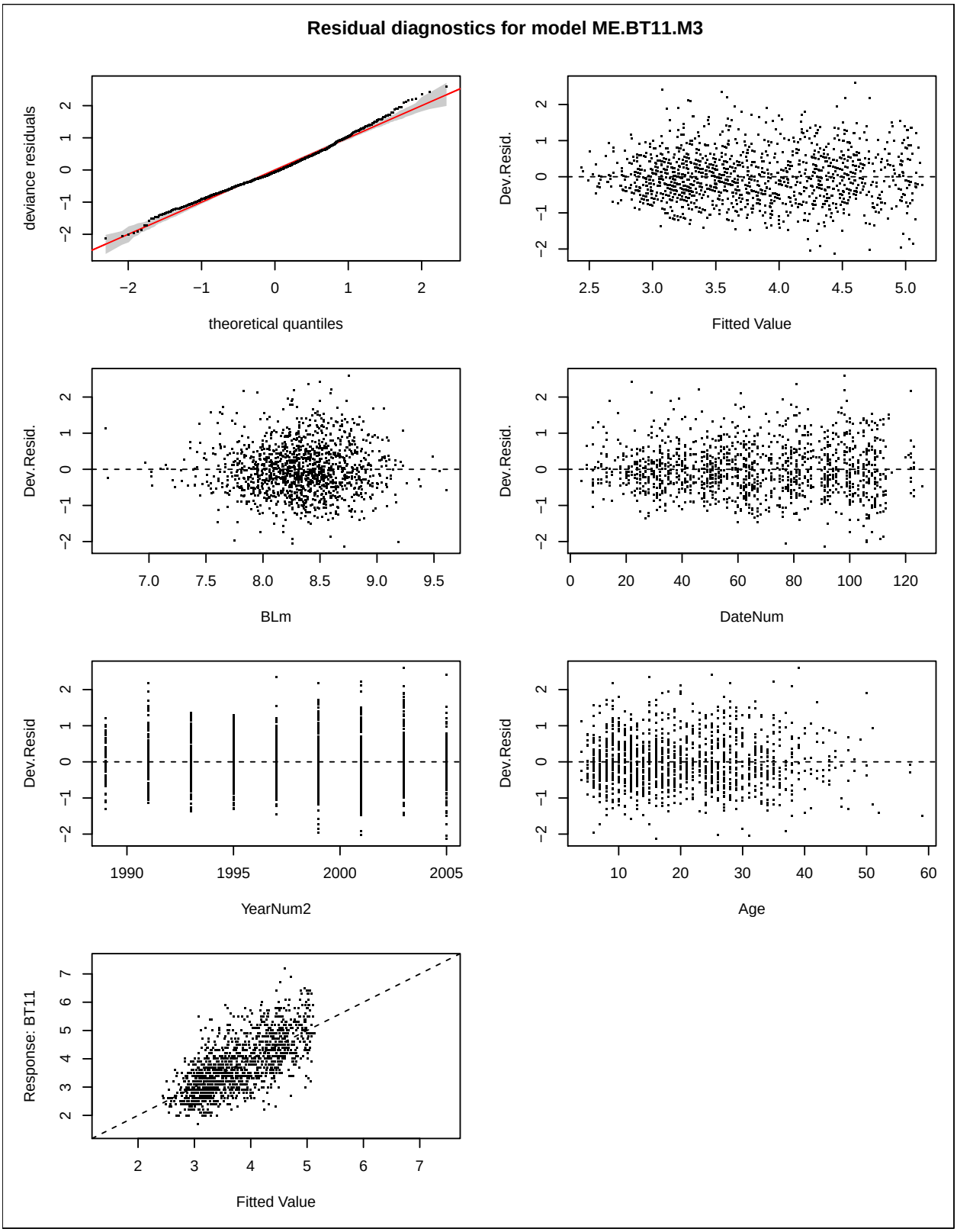

Figure 75: Residual diagnostics for model ME.BT11.M3 for male animals captured in the East region.

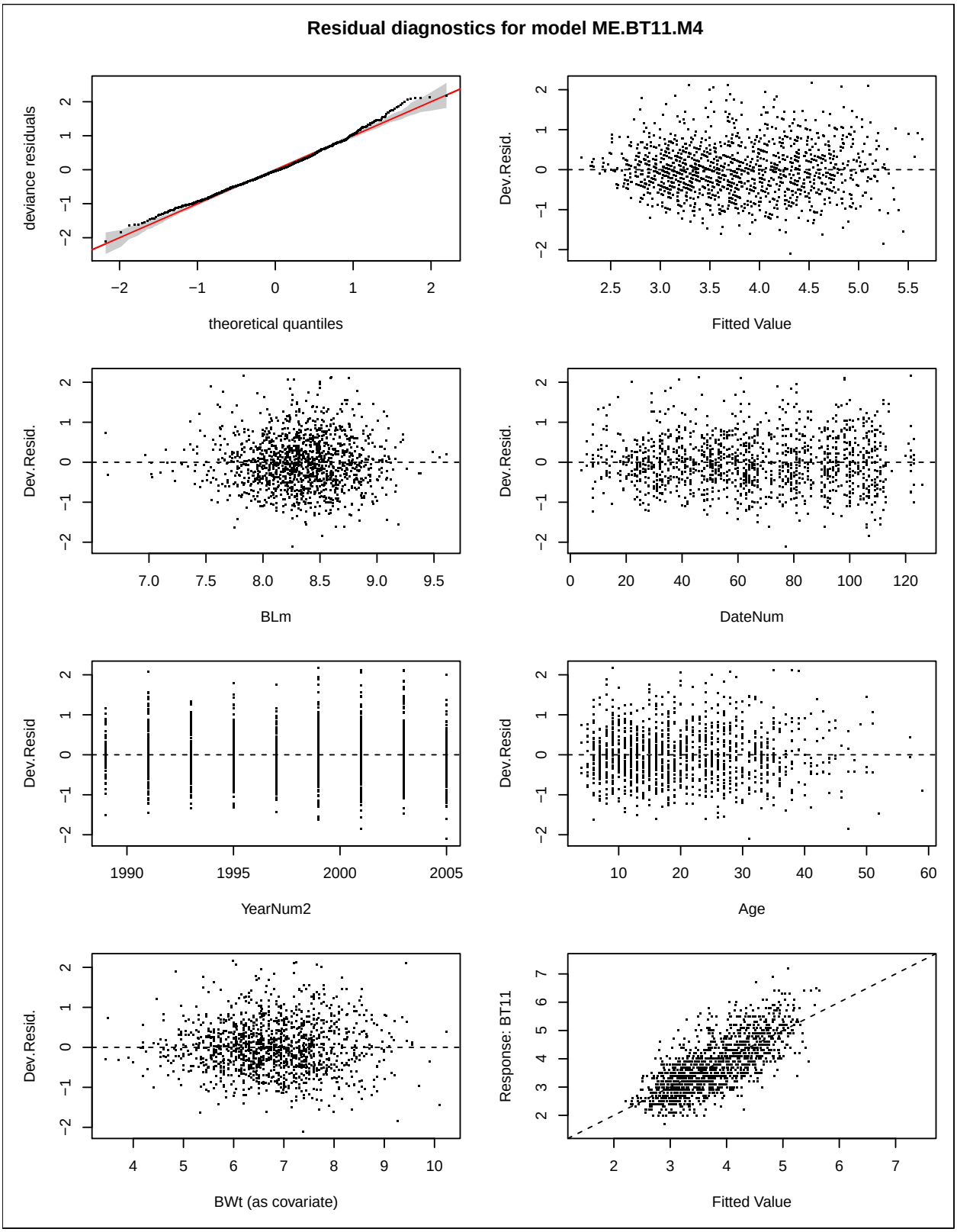

Figure 76: Residual diagnostics for model ME.BT11.M4 for male animals captured in the East region.

E R Session Information

```
## Session info -------------------------------------------------------------------------------------------------------

##  setting  value
##  version  R version 3.4.1 (2017-06-30)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  English_Australia.1252
##  tz       Australia/Hobart
##  date     2018-02-16

## Packages -----------------------------------------------------------------------------------------------------------

##  package      * version date       source
##  backports      1.1.0   2017-05-22 CRAN (R 3.4.1)
##  base         * 3.4.1   2017-06-30 local
##  bitops         1.0-6   2013-08-17 CRAN (R 3.1.0)
##  bookdown       0.5     2017-08-20 CRAN (R 3.4.1)
##  caTools        1.17.1  2014-09-10 CRAN (R 3.1.1)
##  codetools      0.2-15  2016-10-05 CRAN (R 3.2.5)
##  compiler       3.4.1   2017-06-30 local
##  datasets     * 3.4.1   2017-06-30 local
##  devtools       1.13.4  2017-11-09 CRAN (R 3.4.2)
##  digest         0.6.12  2017-01-27 CRAN (R 3.4.1)
##  evaluate       0.10.1  2017-06-24 CRAN (R 3.4.1)
##  gdata          2.18.0  2017-06-06 CRAN (R 3.4.1)
##  gplots       * 3.0.1   2016-03-30 CRAN (R 3.2.5)
##  graphics     * 3.4.1   2017-06-30 local
##  grDevices    * 3.4.1   2017-06-30 local
##  grid         * 3.4.1   2017-06-30 local
##  gtools         3.5.0   2015-05-29 CRAN (R 3.2.1)
##  highr          0.6     2016-05-09 CRAN (R 3.2.5)
##  htmltools      0.3.6   2017-04-28 CRAN (R 3.4.1)
##  KernSmooth     2.23-15 2015-06-29 CRAN (R 3.2.1)
##  knitr        * 1.17    2017-08-10 CRAN (R 3.4.1)
##  lattice      * 0.20-35 2017-03-25 CRAN (R 3.4.1)
##  latticeExtra * 0.6-28  2016-02-09 CRAN (R 3.2.5)
##  magrittr       1.5     2014-11-22 CRAN (R 3.2.1)
##  MASS         * 7.3-47  2017-04-21 CRAN (R 3.4.1)
##  Matrix         1.2-11  2017-08-16 CRAN (R 3.4.1)
##  memoise        1.1.0   2017-04-21 CRAN (R 3.4.1)
##  methods      * 3.4.1   2017-06-30 local
##  mgcv         * 1.8-22  2017-09-19 CRAN (R 3.4.2)
##  nlme         * 3.1-131 2017-02-06 CRAN (R 3.4.1)
##  RColorBrewer * 1.1-2   2014-12-07 CRAN (R 3.4.1)
##  Rcpp           0.12.12 2017-07-15 CRAN (R 3.4.1)
##  rmarkdown      1.8     2017-11-17 CRAN (R 3.4.2)
##  rprojroot      1.2     2017-01-16 CRAN (R 3.4.1)
##  stats        * 3.4.1   2017-06-30 local
##  stringi        1.1.5   2017-04-07 CRAN (R 3.4.1)
##  stringr        1.2.0   2017-02-18 CRAN (R 3.4.1)
##  texreg       * 1.36.23 2017-03-03 CRAN (R 3.4.1)
##  tools          3.4.1   2017-06-30 local
##  utils        * 3.4.1   2017-06-30 local
##  withr          2.0.0   2017-07-28 CRAN (R 3.4.1)
##  xtable       * 1.8-2   2016-02-05 CRAN (R 3.2.5)
##  yaml           2.1.14  2016-11-12 CRAN (R 3.2.5)
```

References

Augustin, N.H., Trenkel, V.M., Wood, S.N., Lorance, P., 2013. Space-time modelling of blue ling for fisheries stock management. Environmetrics 24, 109–119. https://doi.org/10.1002/env.2196

Cunen, C., 2017. No Paradox: Explaining how we can observe a decrease in fat weight, while the total weight appears to remain unchanged. *IWC Document SC/67A/EM/07*: 8pp.

Cunen, C., Walloe, L., Hjort, N.L., 2018. Focused model selection for linear mixed models, with an application to whale ecology. Annals of Applied Statistics *Submitted*, 38pp.

Cunen, C., Walloe, L., Hjort, N.L., 2017. Decline in energy storage in Antarctic minke whales during the JARPA period: Assessment via the Focused Information Criterion (FIC). *IWC Document SC/67A/EM/04*: 56pp.

de la Mare, W.K., 2012. Lurking variables and the interpretation of statistical analyses of data collected under JARPA. *IWC Document SC/64/EM/03*: 65pp.

de la Mare, W.K., 2011. Are reported trends in Antarctic minke whale body condition reliable? *IWC Document SC/63/O16* : 25pp.

de la Mare, W.K., Candy, S.G., McKinlay, J.P., Wotherspoon, S.J., Double, M.C., 2014. What can be concluded from the statistical analyses of JARPA/JARPA II body condition data? *Paper Presented to the IWC Expert Panel Review of JARPA II, SC/F14/O06* : 58pp.

de la Mare, W.K., McKinlay, J.P., Welsh, A.H., 2017. Analyses of the JARPA Antarctic minke whale fat weight data set. *IWC Document SC/67A/EM/01*: 57pp.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. Chapman & Hall.

Hedley, S.L., Bravington, M.V., 2014. Comments on design-based and model-based abundance estimates for the RMP and other contexts *IWC Document SC/65B/RMP/11*: 33pp.

IWC, 2017. Report of the Scientific Committee. Annex L: Report of the Working Group on Ecosystem Modelling *IWC/67/Rep01(2017) Annex L., Bled, Slovenia, 9-21 May 2017.*

Kahneman, D., Frederick, S., 2005. A model of heuristic judgment, in: The Cambridge Handbook of Thinking and Reasoning. Cambridge University Press, pp. 267–293.

Kleiber, M., 1961. The fire of life: An introduction to animal energetics. Wiley, New York.

Konishi, K., Hakamada, T., Kiwada, H., Kitakado, T., Walløe, L., 2014. Decrease in stomach contents in the Antarctic minke whale (Balaenoptera bonaerensis) in the Southern Ocean. Polar Biology 37, 205–215. https://doi.org/10.1007/s00300-013-1424-3

Konishi, K., Tamura, T., Zenitani, R., Bando, T., Kato, H., Walløe, L., 2008. Decline in energy storage in the Antarctic minke whale (Balaenoptera bonaerensis) in the Southern Ocean. Polar Biology 31, 1509–1520. https://doi.org/10.1007/s00300-008-0491-3

Konishi, K., Walløe, L., 2015. Substantial decline in energy storage and stomach fullness in Antarctic minke whales (Balaenoptera bonaerensis) during the 1990s. J. Cetacean Res. Manage 15, 77–92.

Leaper, R., Lavigne, D., 2007. How much do large whales eat? J. Cetacean Res. Manage 9, 179–188.

Lockyer, C., 1981. Estimation of the energy costs of growth, maintenance and reproduction in the female minke whale (Balaenoptera acutorostrata), from the Southern Hemisphere. Rep. Int. Whal. Commn. 31, 337–343.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis 55, 2372–2387. https://doi.org/10.1016/j.csda.2011.02.004

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall/CRC.

McKinlay, J.P., de la Mare, W.K., Welsh, A.H., 2017. A re-examination of minke whale body condition as

reflected in the data collected under JARPA. *IWC Document SC/67A/EM/02*: 108pp.

Miller, D.L., Burt, M.L., Rexstad, E.A., Thomas, L., 2013. Spatial models for distance sampling data: Recent developments and future directions. Methods in Ecology and Evolution 4, 1001–1010. https://doi.org/10.1111/2041-210X.12105

Pinkerton, M.H., Bradford-Grieve, J.M., 2014. Characterizing foodweb structure to identify potential ecosystem effects of fishing in the Ross Sea, Antarctica. ICES Journal of Marine Science 71, 1542–1553. https://doi.org/10.1093/icesjms/fst230

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sarkar, D., 2008. Lattice. Springer New York, New York, NY.

Shah, A.K., Oppenheimer, D.M., 2008. Heuristics made easy: An effort-reduction framework. Psychological Bulletin 134, 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Tamura, T., Konishi, K., 2014. Prey composition and consumption rate by Antarctic minke whales based on JARPA and JARPA II data (No. SC-F14-J15). Paper SC.

Williams, R., Vikingsson, G.A., Gislason, A., Lockyer, C., New, L., Thomas, L., Hammond, P.S., 2013. Evidence for density-dependent changes in body condition and pregnancy rate of North Atlantic fin whales over four decades of varying environmental conditions. ICES Journal of Marine Science 70, 1273–1280. https://doi.org/10.1093/icesjms/fst059

Wood, S.N., 2017. Generalized Additive Models: An introduction with R, 2nd ed. Chapman and Hall.

Wood, S.N., 2011. Fast Stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological) 73, 3–36.

Wotherspoon, S.J., Double, M.C., McKinlay, J.P., Candy, S.G., Andrews-Goff, V., de la Mare, W.K., 2014. JARPA and JARPA II cannot monitor trends in the Antarctic ecosystem due to flawed sampling strategies *Paper Presented to the IWC Expert Panel Review of JARPA II, SC/F14/O05* : 12pp.

Xie, Y., 2016. Bookdown: Authoring Books and Technical Documents with R Markdown. CRC Press.