# Supporting information

for:

Ute Bradter, John D. Altringham, William E. Kunin, Tim J. Thom,Jerome O'Connell & Tim G. Benton: Variable ranking and selection with random forest for unbalanced data. Environmental Data Science.

## Fig S1: The case study area and the location of survey transects

a) b)



a) The study area (black line) within Great Britain (grey line).

b) Wader surveys were carried out from 61 transects (black lines) focused on areas below 500 m in elevation (shaded in grey) within the study area, which was located within the Yorkshire Dales, an upland area of the UK.

Outline of Great Britain from Strategi data downloaded from the EDINA Digimap OS service. © Crown Copyright/database right 2009. An Ordnance Survey/EDINA supplied service.

## S1: Additional information to the case studies

### Bird surveys

The transect locations were selected by superimposing a 4 x 4 km grid over the study area. Grid intersections within 500 m of areas above 500 m in elevation were omitted to focus available survey time on areas of interest for one of the funding organisations. For the remaining intersections, we selected sections of public rights of way (typically footpaths; sometimes shared with other users such as horse riders) which were 2 km in length, at elevations below 500 m, as close as possible to the grid intersections while being as straight as possible, within 1 km of car parking, no closer than 700 m to another transect and not within 1.5 km to larger woods or reservoirs. When several neighbouring transects had the same orientation, we changed the orientation of every second transect if possible (for example, in a river valley several transects were oriented west-east following the course of the river and for every second a more north-south orientation (thus oriented uphill) was chosen).

Each transect was surveyed three times, once each between 2nd - 30th April, 3rd - 22nd May and 2nd June – 1st July. This covered much of the incubation and chick rearing period of the species. First clutches in the UK are laid on average around 12 April for Northern lapwing (25 March - 25 May), 30 April for common snipe (31 March - 24 June) and 1 May for common redshank (14 April - 7 June) (Robinson 2017). Incubation lasts about 25 - 34, 18 - 20 and 24 days, respectively and fledging occurs after about 35 - 40, 19 - 20 and 25 - 35 days, respectively (Robinson 2017).

The sequence in which transects were surveyed was stratified by location so that each transect cluster contained transects surveyed early, late and in the middle part of each month. The time of day in which each transect was surveyed was varied between repeat surveys. Bird surveys are often carried out exclusively in the morning or evening hours due to generally greater activity levels of birds. This would considerably have restricted the number of transects we would have been able to cover due to the travelling times between transects. As the ability to detect Northern lapwing and common redshank was judged reasonable even if individuals were less active and as we carried out repeat surveys to reduce the risk of recording false absences, we carried out surveys any time of day though with an emphasis on mornings. Common snipe will be the least easy to detect when more inactive as they are then often silent and hidden in vegetation. Therefore, if the time of day caused any bias in our surveys, this would likely be strongest for common snipe. Surveys were not carried out in continuous rain, poor visibility or wind of Beaufort scale 5 or more. Surveys were carried out by a single observer, so there is no bias introduced from using multiple observers.

Each transect was walked slowly and all areas visible from the transect were scanned with binoculars (Zeiss 10 x 40 B). The visible area was approximately marked on a field map and later digitized. All individuals of the species were recorded per observation unit.

Records of flocks for all species and records of Northern lapwing from the third repeat survey were omitted as they are likely to include a high proportion of dispersed breeders and fledged birds. Early surveys of common snipe risk recording common snipe displaying on spring migration and surveys targeting common snipe are often not conducted before mid-April (Hoodless, Ewald et al. 2007). Two out of 44 observation units which contained common snipe presences were due to common snipe records from the first half of April only. However, as common snipe was recorded in the adjacent observation unit in another month, we did not exclude these records.

### References

Hoodless, A. N., J. A. Ewald and D. Baines (2007). "Habitat use and diet of Common Snipe *Gallinago gallinago* breeding on moorland in northern England." Bird Study **54**: 182-191.

Robinson, R. A. (2017). BirdFacts: profiles of birds occurring in Britain & Ireland., BTO Research Report 407. BTO, Thetford, <http://bto.org/birdfacts>. Accessed 23 January 2017.

### Covariates and data sources

1) Area of land of six elevation categories at the scales: 0.25, 1, 2.5, 5, 7, and 10 km. Category divisions: 0-200, 200-300, 300-400, 400-500, 500-600, 600-850m.

Data source: 50 m resolution Digital Terrain Model (DTM) (Land-Form PANORAMA downloaded from the EDINA Digimap OS service; http://edina.ac.uk/digimap.© Crown Copyright/database right 1993).

2) Area of land of nine aspect categories at the scales: 0.25, 1, 2.5, 5, 7, and 10 km. Category divisions: flat, north, north-east, east, south-east, south, south-west, west, north-west.

Data source: calculated from the 50 m resolution DTM.

3) Area of land of six slope categories at the scales: 0.25, 1, 2.5, 5, 7, and 10 km. Category divisions: 0-2, 2-5, 5-10, 10-15, 15-25, 25-60°.

Data source: calculated from the 50 m resolution DTM.

4) Area of land with specific soil characteristics (13 categories) at the scales: 0.25, 1, 2.5, 5, 7, and 10 km. Category divisions: these were formed using the attributes of the data (NSRI 2009):

Texture: peat, loam

Fertility and lime status: lime-rich, very low, low, moderate, high fertility. Very low and low were further grouped as unfertile soils; moderate and high as fertile soils.

Drainage: well drained, impeded drainage, wet soils. Soils with impeded drainage and wet soils were further grouped as moist soils.

Data source: NATMAP soilscapes, the simplified version of the National Soil Map data set, 1:125000 vector map (NATMAP soilscapes © Cranfield University (NSRI) and for the Controller of HMSO 2009).

5) Area of settlements (houses, gardens, etc.) at the scales: 0.25, 1, 2.5, 5, 7, and 10 km.

Data source: OS MasterMap, downloaded from the EDINA Digimap OS service (http://edina.ac.uk/digimap. © Crown Copyright/database right 2007. An Ordnance Survey/EDINA supplied service).

6) Length of paths/roads at the scales 0.25 and 1 km.

Data source: OS MasterMap and a database on tracks and paths held by the Yorkshire Dales National Park Authority.

7) Length of field walls (Obstructing features in the OS MasterMap) at the scales 0.25 and 1 km.

Data source: OS MasterMap

8) Viewshed (the area visible from observation points) at the scales 0.25 and 1 km. It was calculated for four random points per observation unit and averaged per observation unit and spatial scale. Random points had to be at least 120 m apart.

Data source: DTM at 10 m resolution (Land-Form Profile downloaded from the EDINA Digimap OS service; http://edina.ac.uk/digimap.© Crown Copyright/database right 2003) onto which 2 m for all field walls and 7 m for all buildings were added.

9) Average annual rainfall for the period 1961-90 at a single scale, the resolution of the data (5 km).

Data source: www.metoffice.gov.uk.

10) Numbers of sheep and cattle in 2004 at a single scale, the resolution of the data (2 km).

Data source: AG Census downloaded from the EDINA Digimap OS service (http://edina.ac.uk/digimap.© Crown Copyright/database right 2009).

11) Satellite imagery: within each observation unit: the mean, min, max and standard deviation of each band and the following vegetation indices used in some studies on wetlands (e.g. Ozesmi and Bauer 2002, Huang, Peng et al. 2014): normalized difference vegetation index (NDVI: (Band 4 - Band 3) / (Band 4 + Band 3)),

normalized difference wetness index (NDWI in two forms: (Band 4 - Band 5) / (Band 4 + Band 5) and (Band 3 - Band 5) / (Band 3 + Band 5)),

infrared-visible ratio (IVR: Band 5 / Band 2),

infrared ratio (IR: (Band 5 - Band 7) / (Band 5 + Band 7)),

ratio between bands 4 and 2.

Data source: 30 m resolution (60 m for thermal band) imagery from Landsat 7 (22 March 2003) and Landsat 5 (10 June 2006). Landsat Surface Reflectance products courtesy of the U.S. Geological Survey Earth Resources Observation and Science Center. Downloaded from http://landsat.usgs.gov on 20 May 2014. Acquisition dates of satellite imagery were chosen from days with minimal cloud cover and with no or very little rain on the day of acquisition and the five days prior. Rainfall data extracted from the MetOffice MIDAS Land and Marine Surface Station Data.

12) Number of walking groups per transect averaged over the three repeat surveys.

Data source: Wader survey.

13) Area of observation unit visible from transect route.

Data source: Digitized from field maps.

### References

Huang, C., Y. Peng, M. Lang, I.-Y. Yeo and G. McCarty (2014). "Wetland inundation mapping and change monitoring using Landsat and airborne LiDAR data." Remote Sensing of Environment **141**: 231-242.

NSRI. (2009). "NATMAPsoilscapes - The simplified digital national soil map. 1:250,000 scale." Cranfield University Retrieved 2 June 2009, from www.landis.org.uk.

Ozesmi, S. L. and M. E. Bauer (2002). "Satellite remote sensing of wetlands." Wetlands Ecology and Management **10**: 381-402.

### Proxy covariates used and their potential influences on the focal species

Food availability

Aspect, elevation, slope and soil type might influence soil invertebrate abundance through soil fertility, soil moisture, soil pH and temperature (Curry 1994, Curry 2004). Wetter soils, in areas with higher rainfall or more moisture retaining soils, may increase food availability for wader species which probe the soil (Green, Hirons et al. 1990).

Microclimatic conditions

Aspect and elevation might influence microclimate through temperature, wind exposure and solar irradiation and therefore thermoregulation and water loss of individuals (Wolf and Walsberg 1996). Long periods of rain can be detrimental to chick body condition (Eglington, Bolton et al. 2010).

Habitat structure

The structure or type of vegetation can affect habitat selection patterns or nest success (Martin 1993, van der Wal and Palmer 2008, Chalfoun and Martin 2009). We used soil characteristics and remotely sensed imagery as proxies as they can predict vegetation type (Bradter, Thom et al. 2011). We also included livestock numbers (sheep and cattle) as livestock grazing can influence vegetation structure and type (McCracken and Tallowin 2004) and occupancy patterns of birds. Northern lapwing for example prefer areas with short sward (O'Brien 2002). Furthermore, livestock grazing can affect reproductive success of ground nesting waders (Bairlein and Bergner 1995).

Wet areas

For the bands and vegetation indices useful for the detection of wet areas, see the previous section entitled ‘covariates and data sources’.

Disturbance and perceived predation risk

Paths, roads and railways and human settlements may be potential sources of disturbance. Human settlements may also be potential sources of predators (Marzluff and Neatherlin 2006). We included the length of walls as a correlate for field size. Field size has been found to have a positive effect on the presence of Northern lapwing possibly because larger fields provide larger areas with unrestricted views which may aid predator detection (Small 2002). Viewshed, the area visible from a fixed point, might influence predator detection or perceived predation risk (Devereux, Whittingham et al. 2006).

### References

Bairlein, F. and G. Bergner (1995). "Vorkommen und Bruterfolg von Wiesenvögeln in der nördlichen Wesermarsch, Niedersachsen." Vogelwelt **116**: 53-59. In German with English summary.

Bradter, U., T. J. Thom, J. D. Altringham, W. E. Kunin and T. G. Benton (2011). "Prediction of National Vegetation Classification communities in the British uplands using environmental data at multiple spatial scales, aerial images and the classifier random forest." Journal of Applied Ecology **48**: 1057-1065.

Chalfoun, A. D. and T. E. Martin (2009). "Habitat structure mediates predation risk for sedentary prey: experimental tests of alternative hypotheses." Journal of Animal Ecology **78**(3): 497-503.

Curry, J. P. (1994). Grassland Invertebrates - Ecology, influence on soil fertility and effects on plant growth. London, Chapman & Hall.

Curry, J. P. (2004). Factors affecting the abundance of earthworms in soils. Earthworm ecology. C. A. Edwards. London, CRC Press**:** 91-113.

Devereux, C. L., M. J. Whittingham, E. Fernández-Juricic, J. A. Vickery and J. R. Krebs (2006). "Predator detection and avoidance by starlings under differing scenarios of predation risk." Behavioral Ecology **17**: 303-309.

Eglington, S. M., M. Bolton, M. A. Smart, W. J. Sutherland, A. R. Watkinson and J. A. Gill (2010). "Managing water levels on wet grasslands to improve foraging conditions for breeding northern lapwing *Vanellus vanellus*." Journal of Applied Ecology **47**(2): 451-458.

Green, R. E., G. J. M. Hirons and B. H. Cresswell (1990). "Foraging habitats of female common snipe *Gallinago gallinago* during the incubation period." Journal of Applied Ecology **27**(1): 325-335.

Martin, T. E. (1993). "Nest predation and nest sites - New perspectives on old patterns." BioScience **43**(8): 523-532.

Marzluff, J. M. and E. Neatherlin (2006). "Corvid response to human settlements and campgrounds: Causes, consequences, and challenges for conservation." Biological Conservation **130**: 301-314.

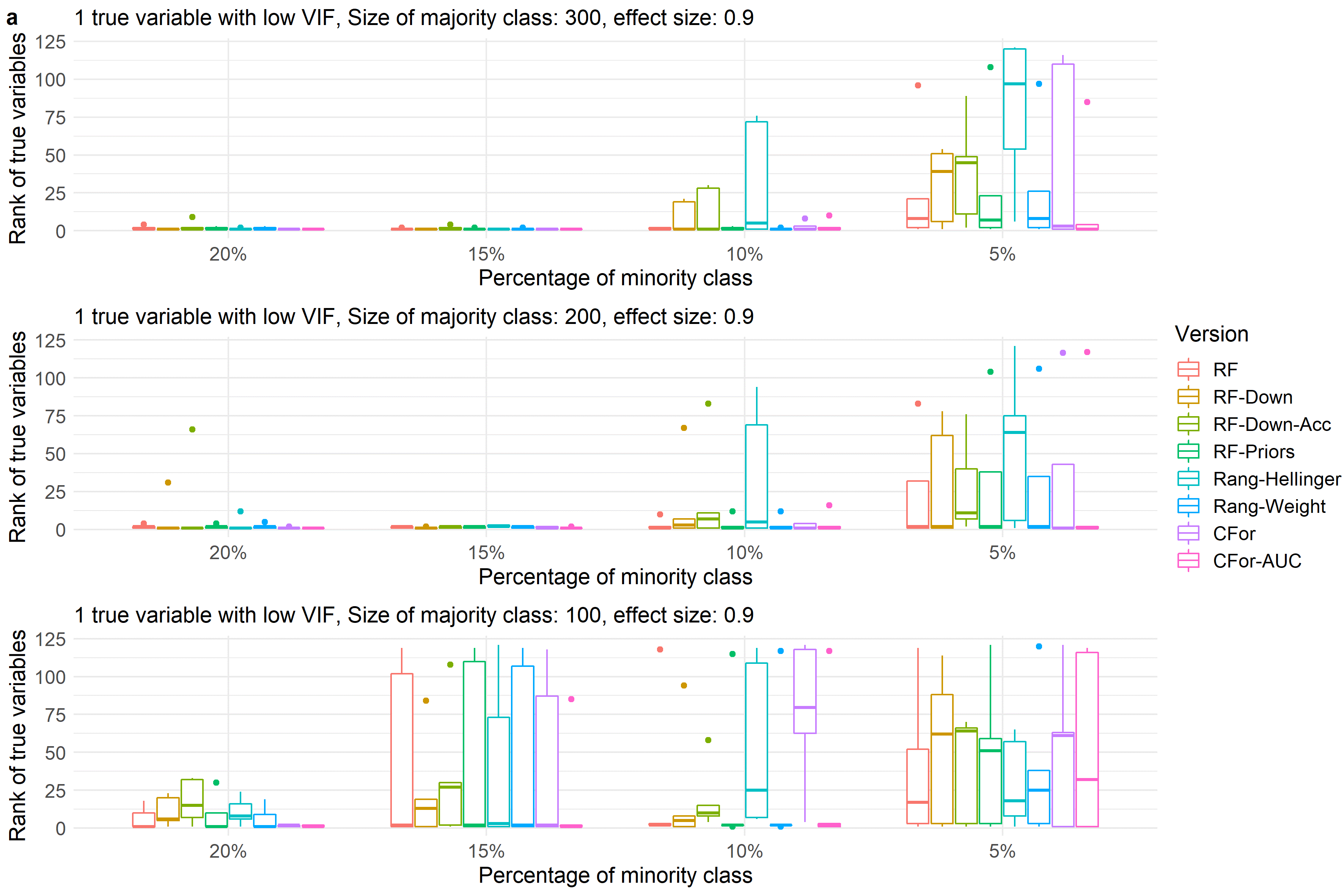
McCracken, D. I. and J. R. Tallowin (2004). "Swards and structure: the interactions between farming practices and bird food resources in lowland grasslands." Ibis **146**: 108-114.

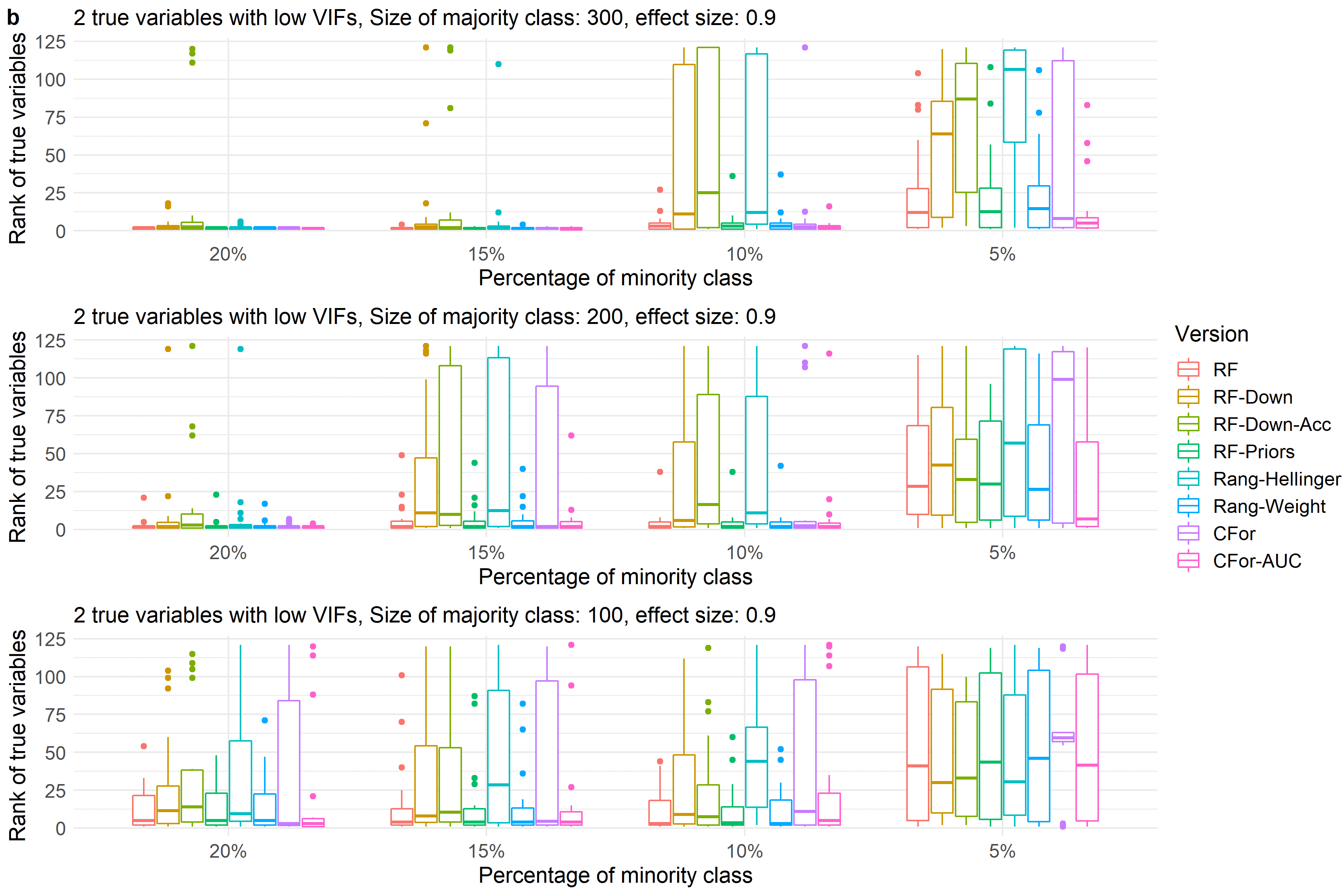
O'Brien, M. (2002). "The relationship between field occupancy rates by breeding lapwing and habitat management on upland farmland in Northern Britain." Aspects of Applied Biology **67**: 85-92.

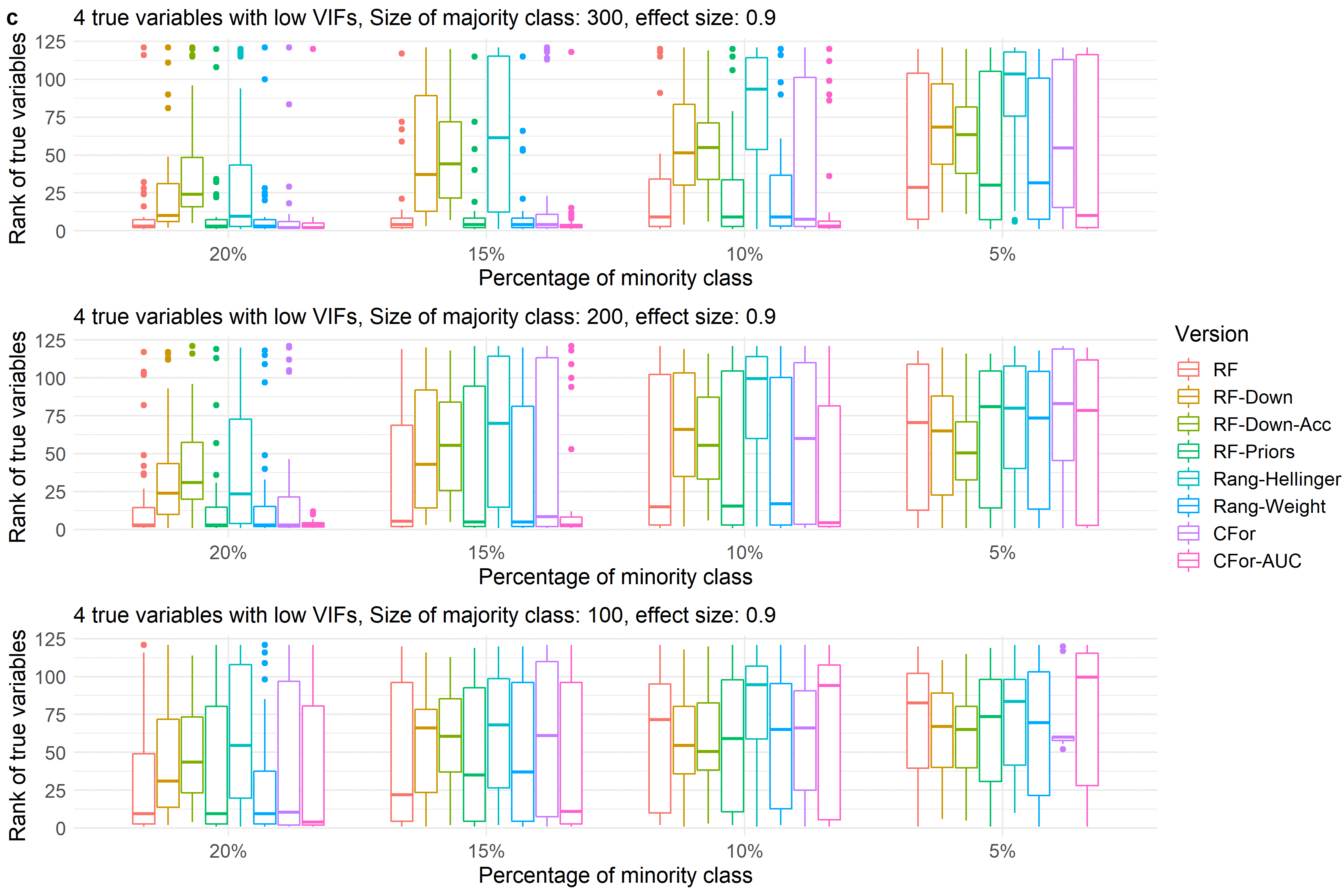
Small, C. J. (2002). Waders, habitats and landscape in the Pennine Dales. PhD, University of Lancaster.

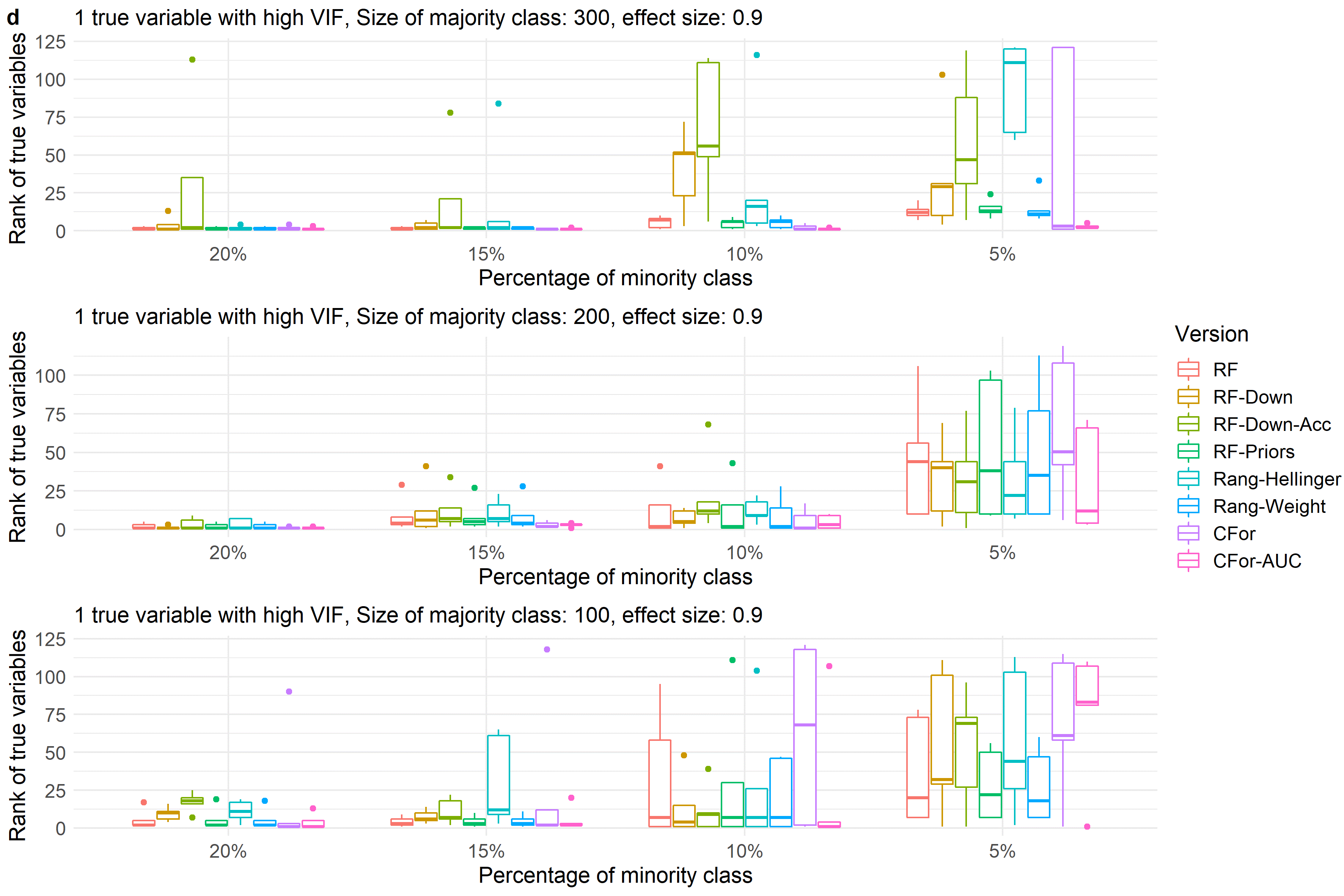
van der Wal, R. and S. C. F. Palmer (2008). "Is breeding of farmland wading birds depressed by a combination of predator abundance and grazing?" Biology Letters **4**: 256-258.

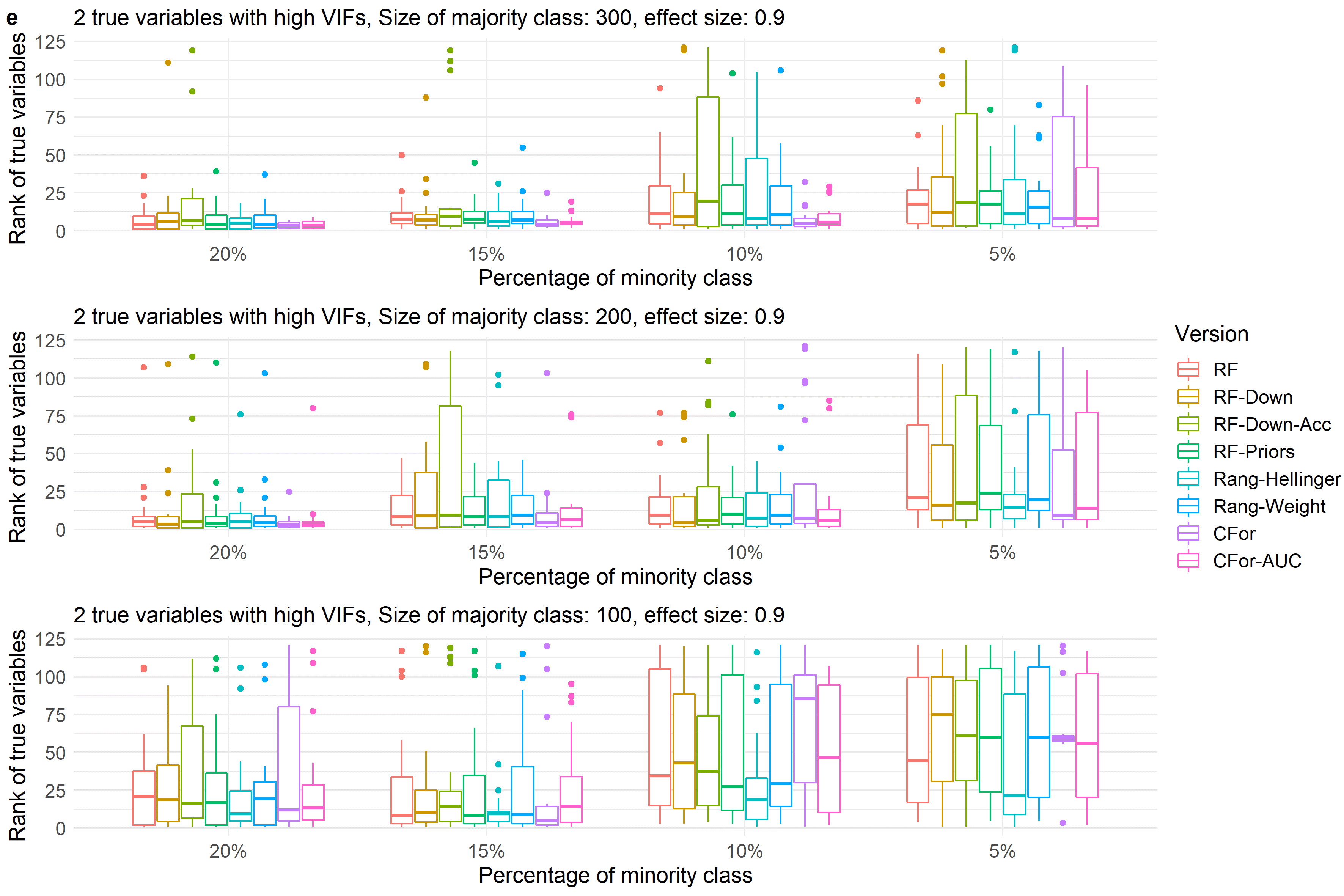
Wolf, B. O. and G. E. Walsberg (1996). "Thermal effects of radiation and wind on a small bird and implications for microsite selection." Ecology **77**(7): 2228-2236.

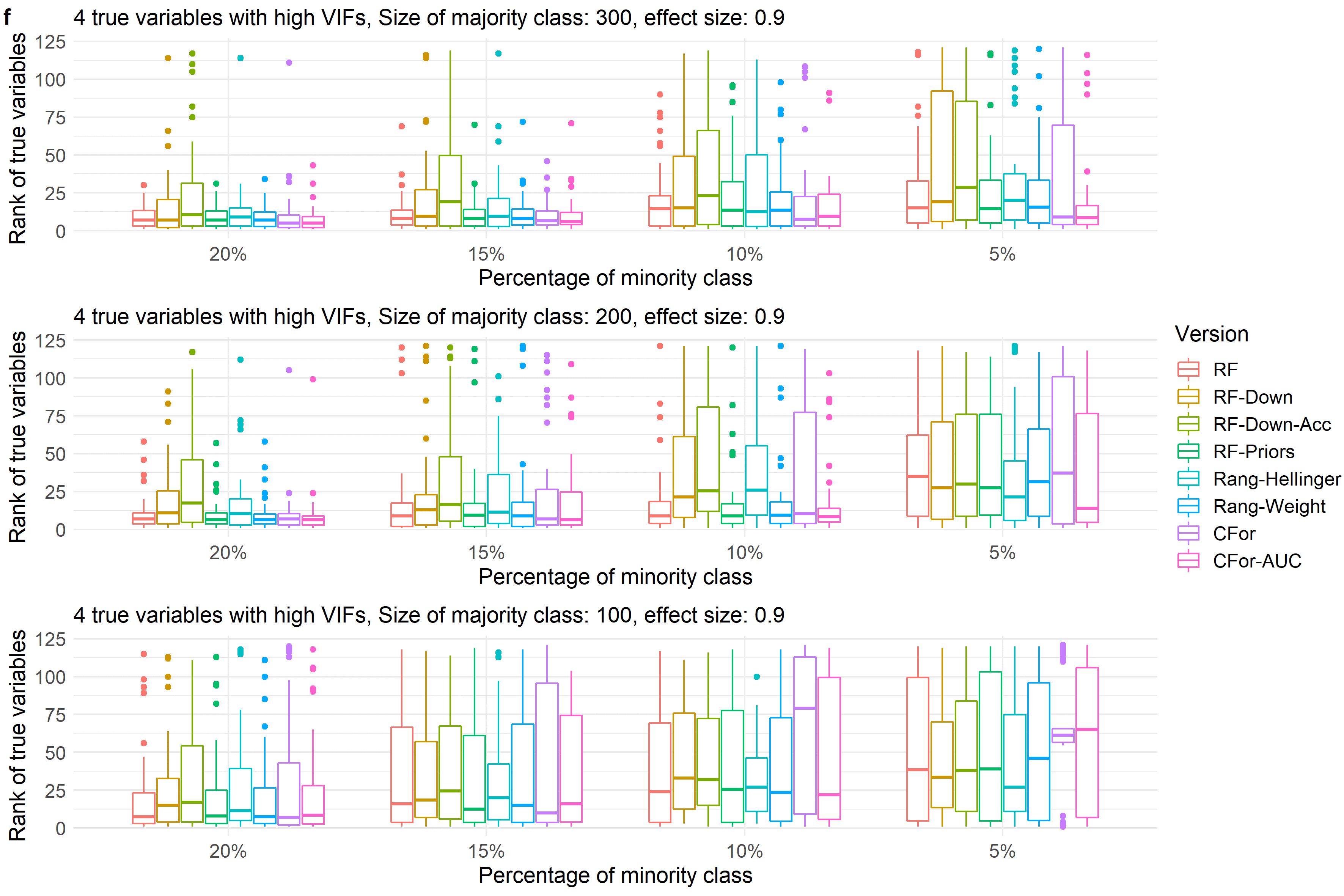








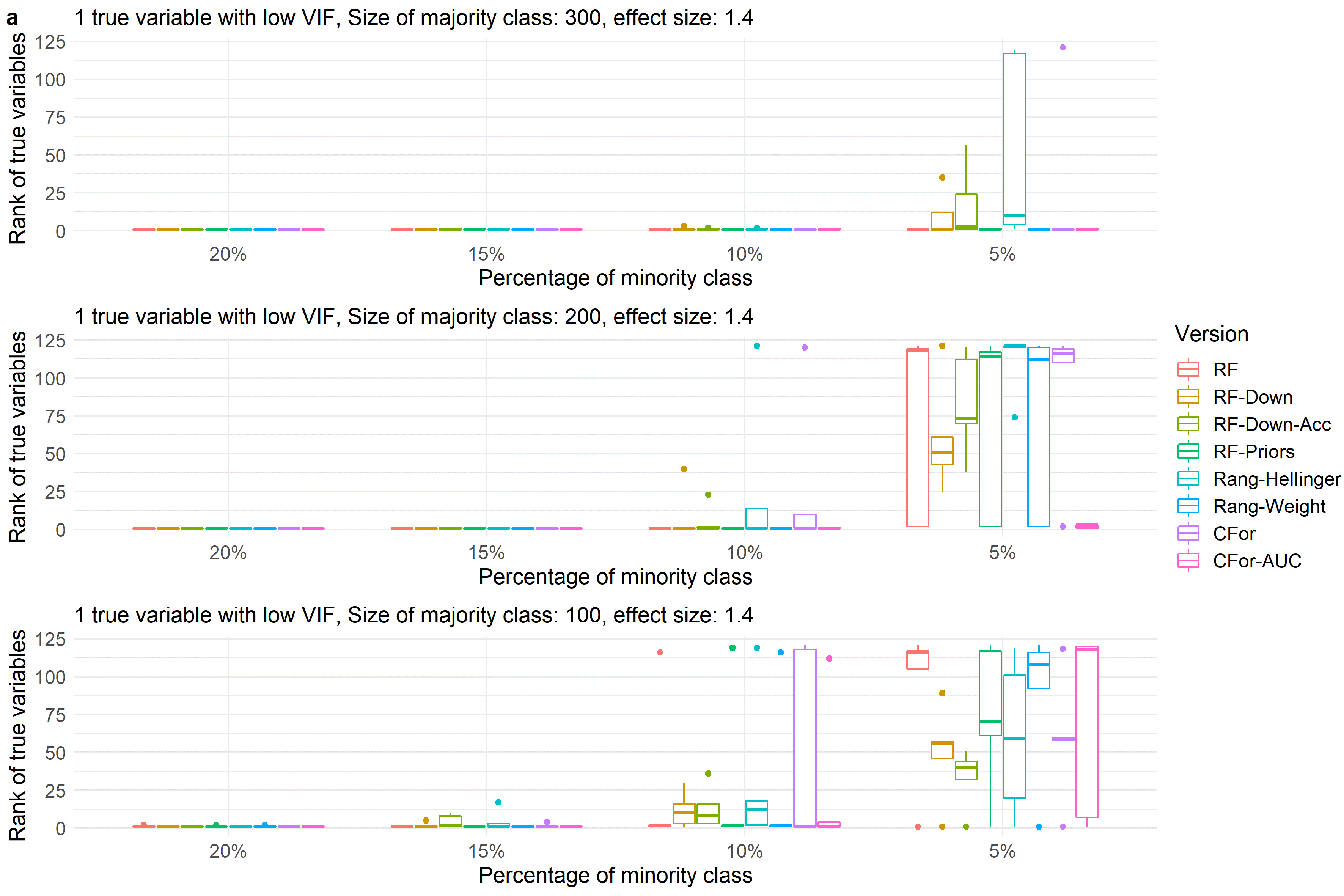


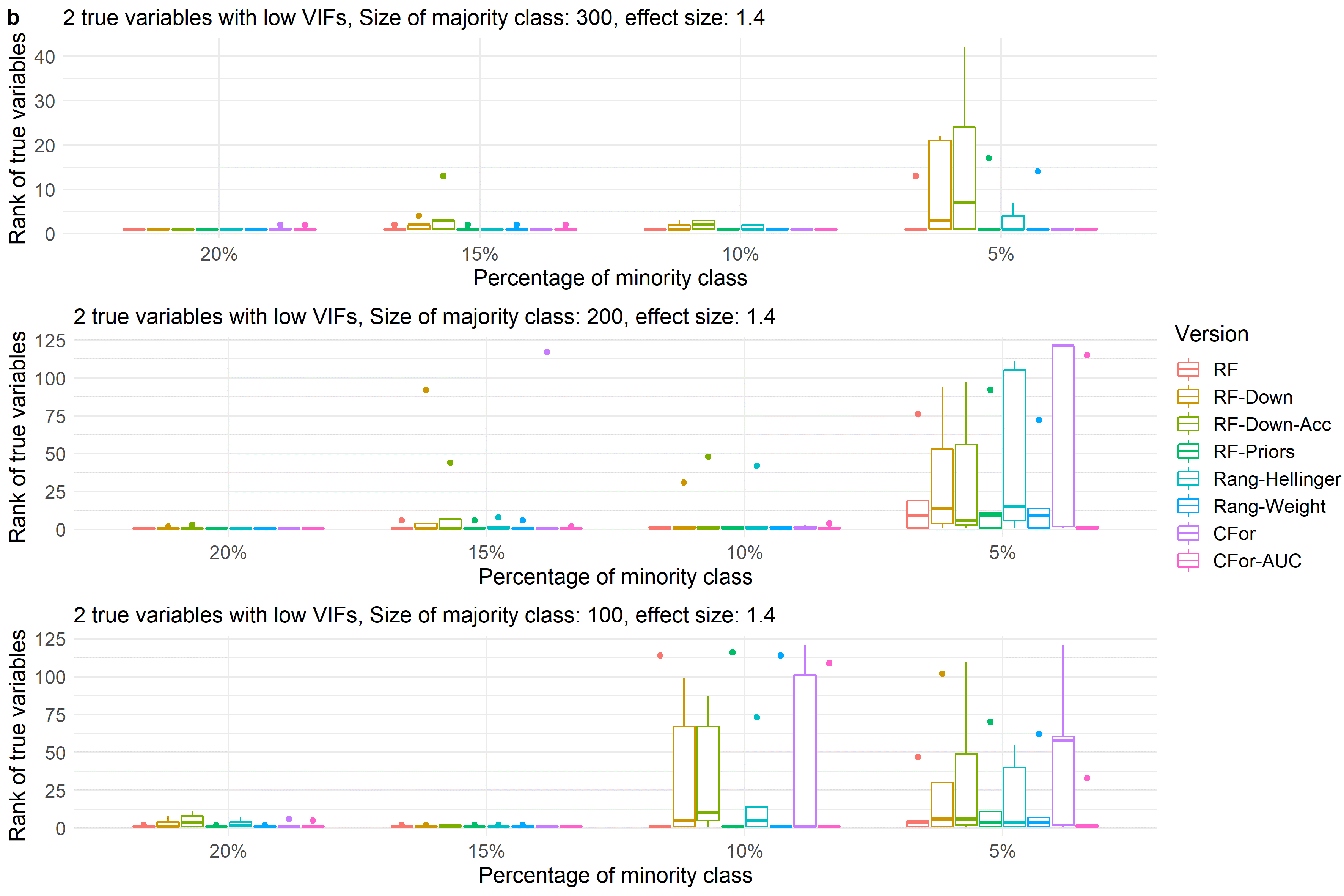


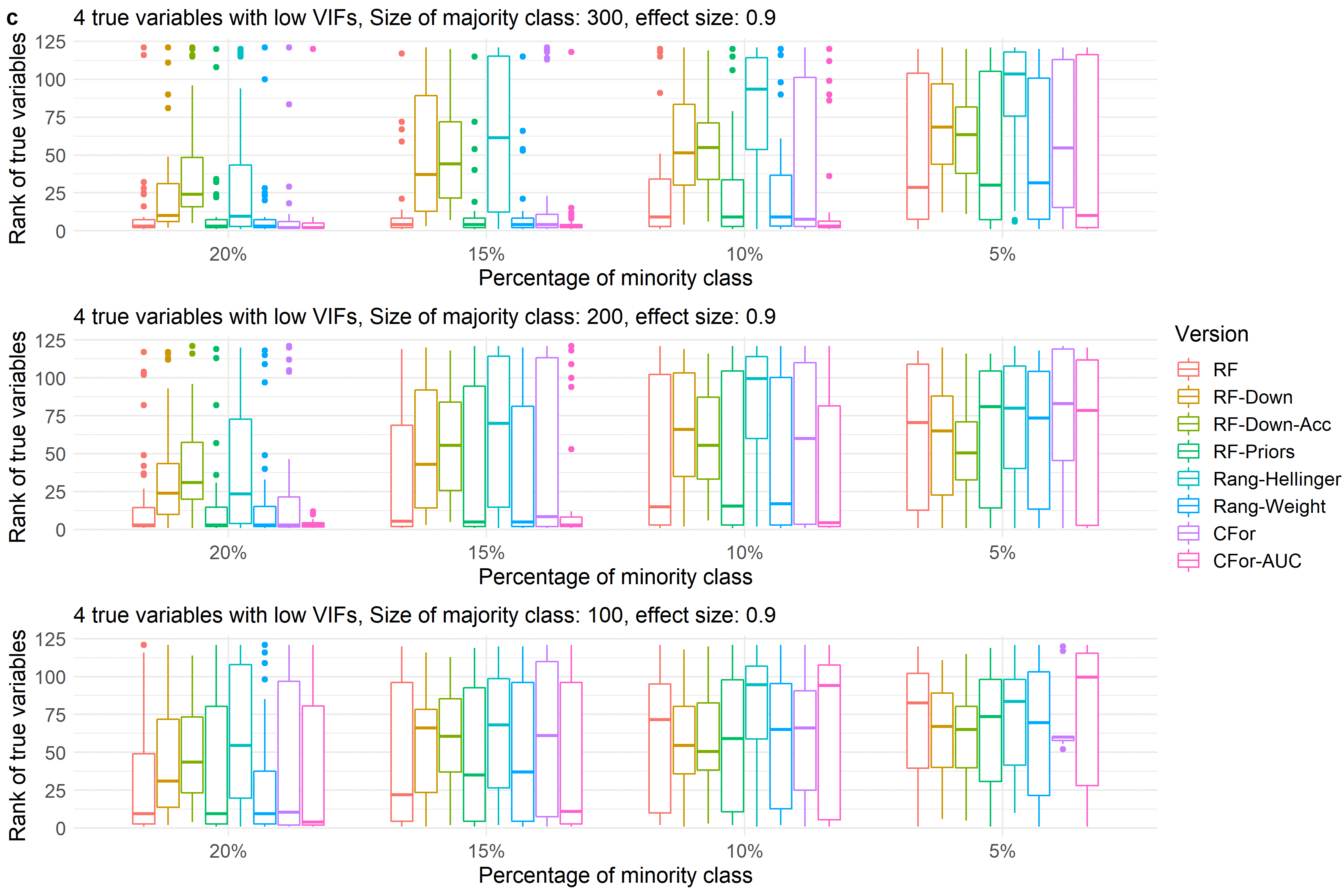
## Fig. S2

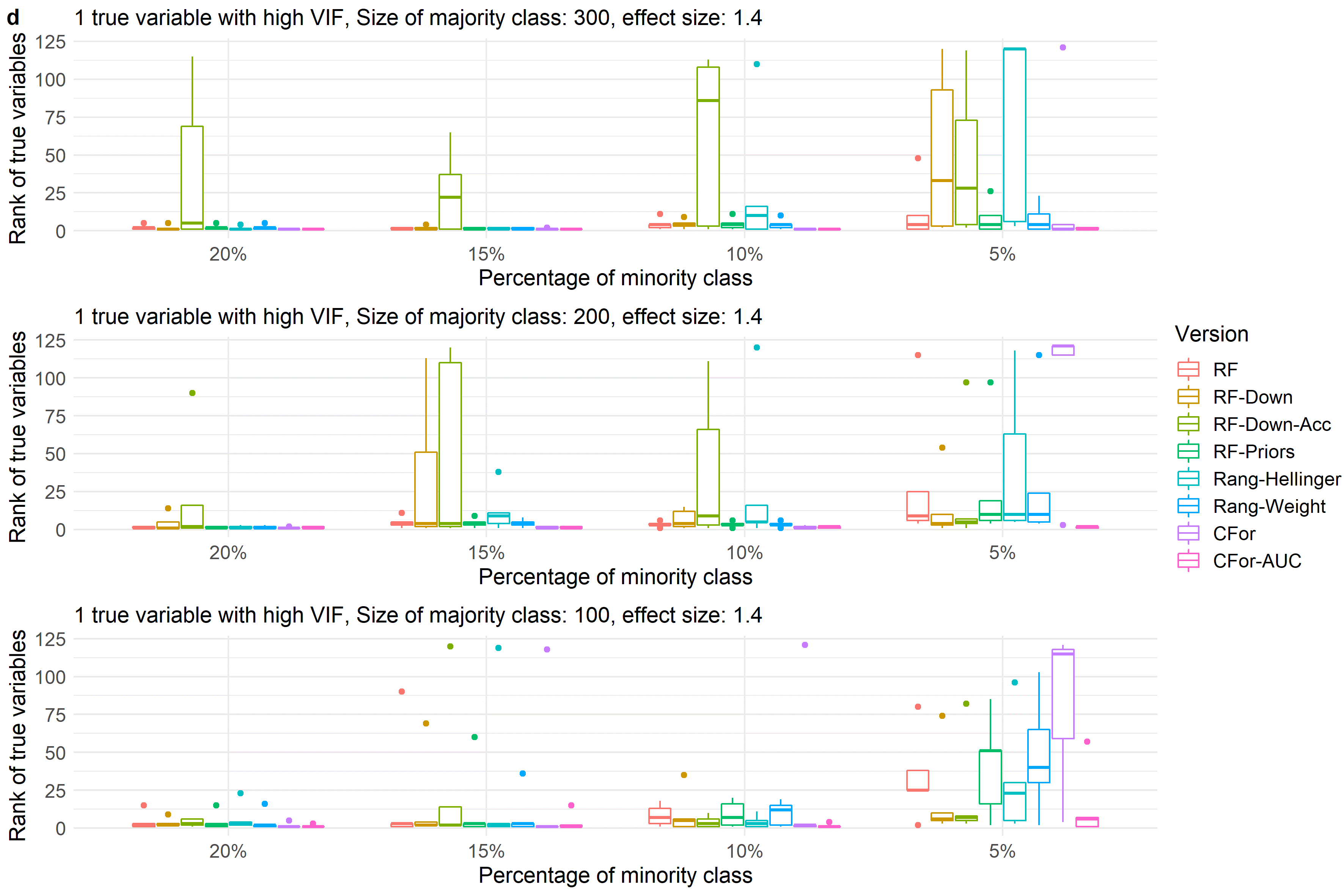
Boxplots showing ranks of permutation importance for true variables with a medium effect size of 0.9 for simulated data with a) one true variable with low VIF, b) two true variables with low VIFs, c) four true variables with low VIFs, d) one true variable with high VIF, e) two true variables with high VIFs, f) four true variables with high VIFs.

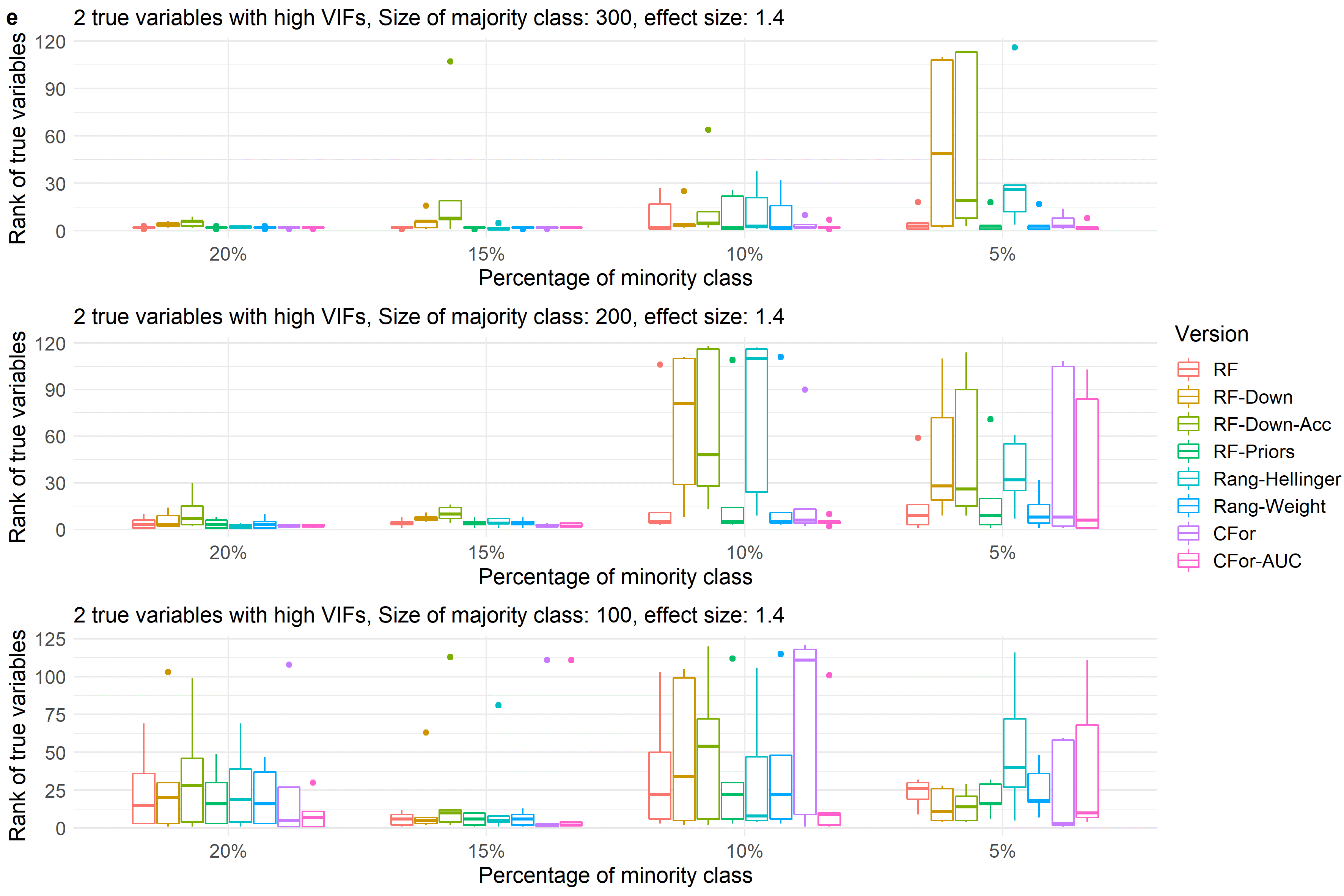
For each of the 60 simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors was ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 \* interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right).

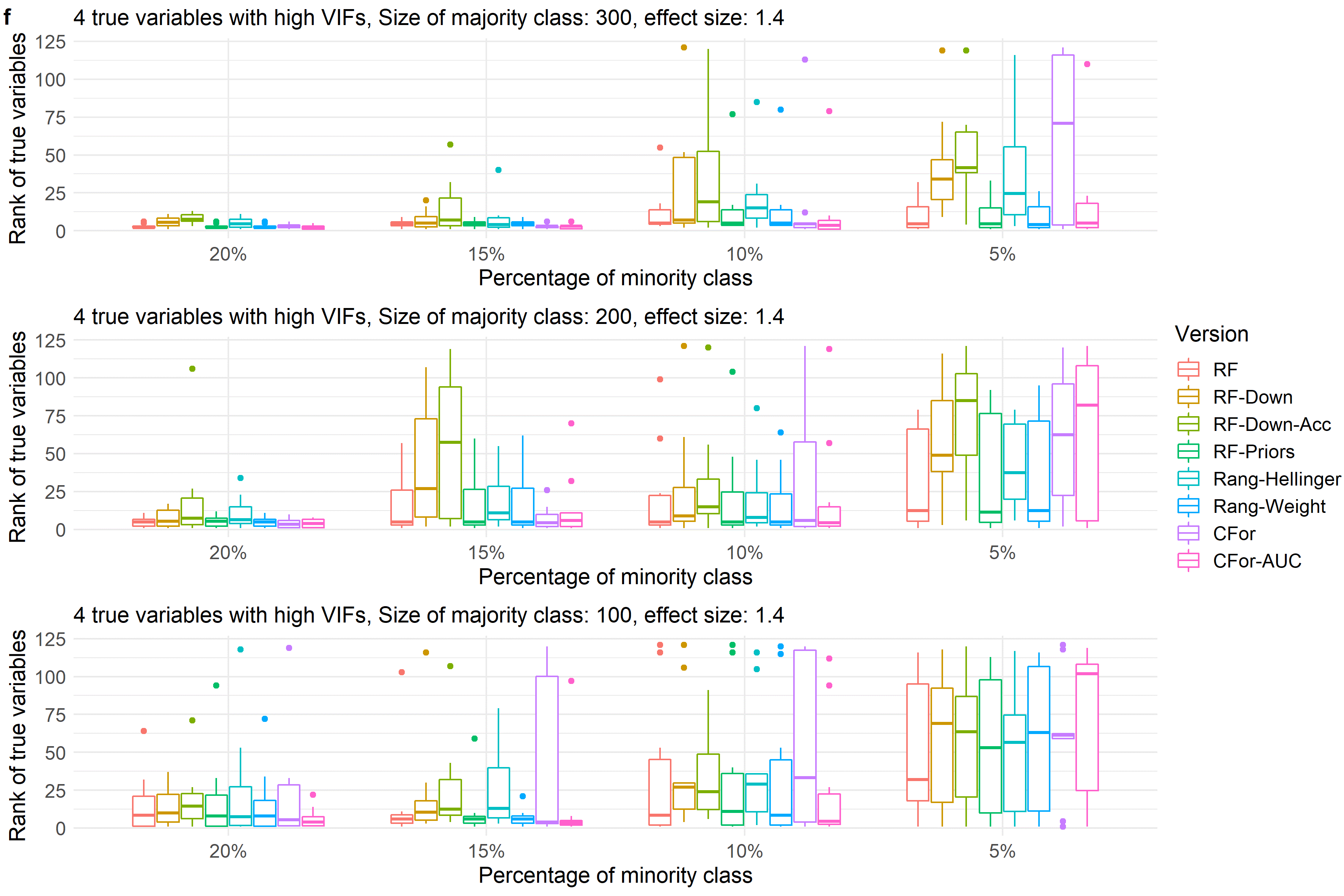








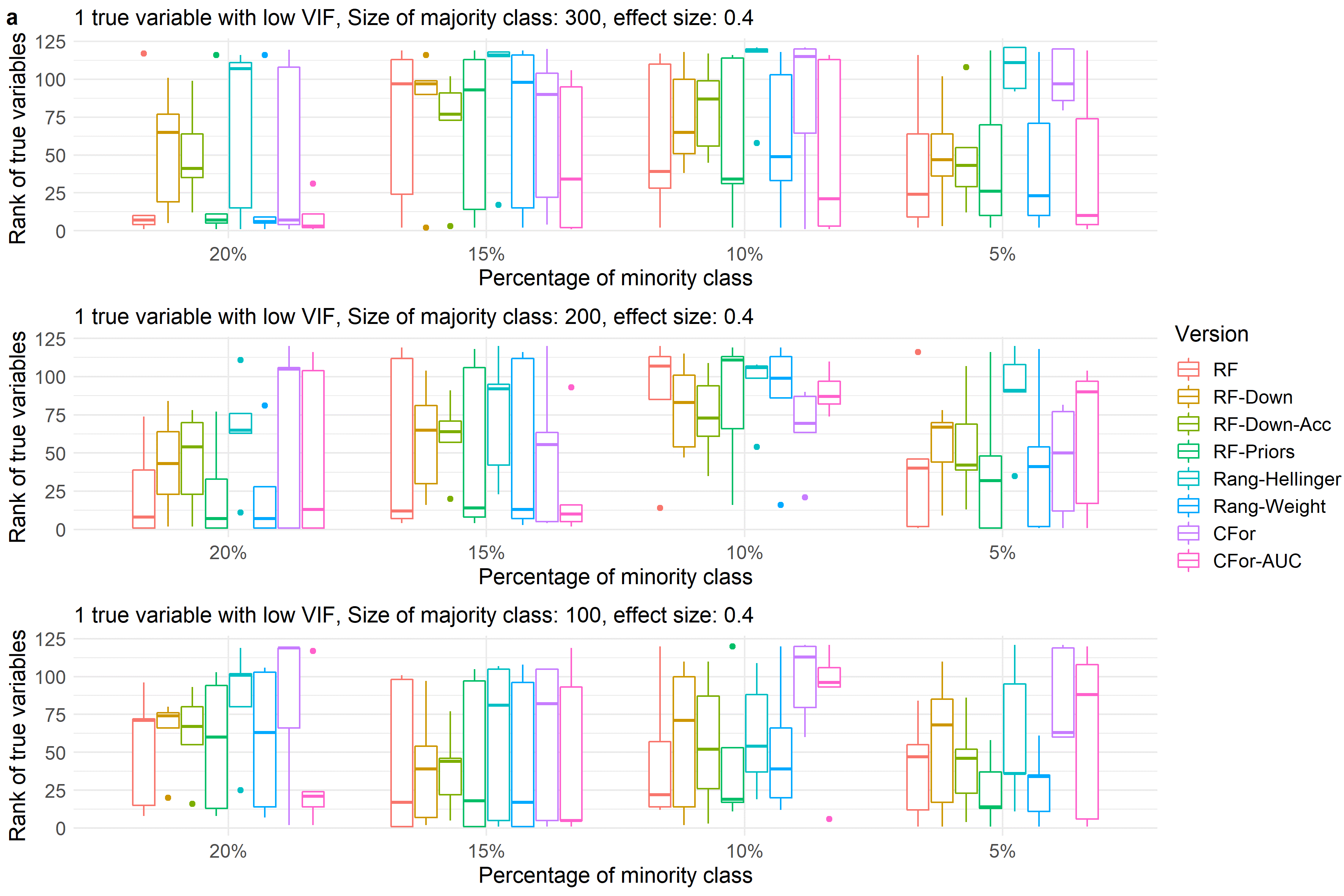


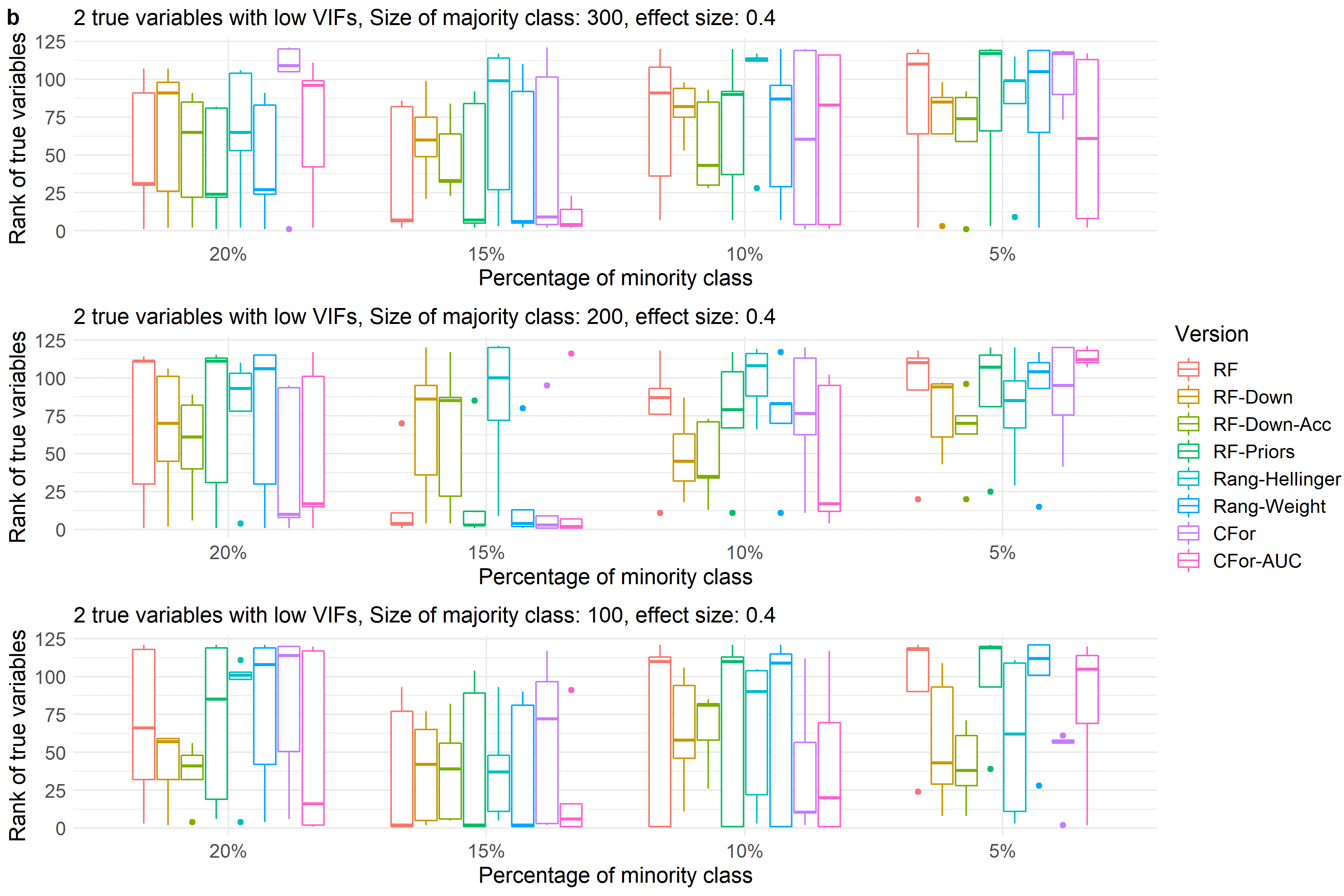


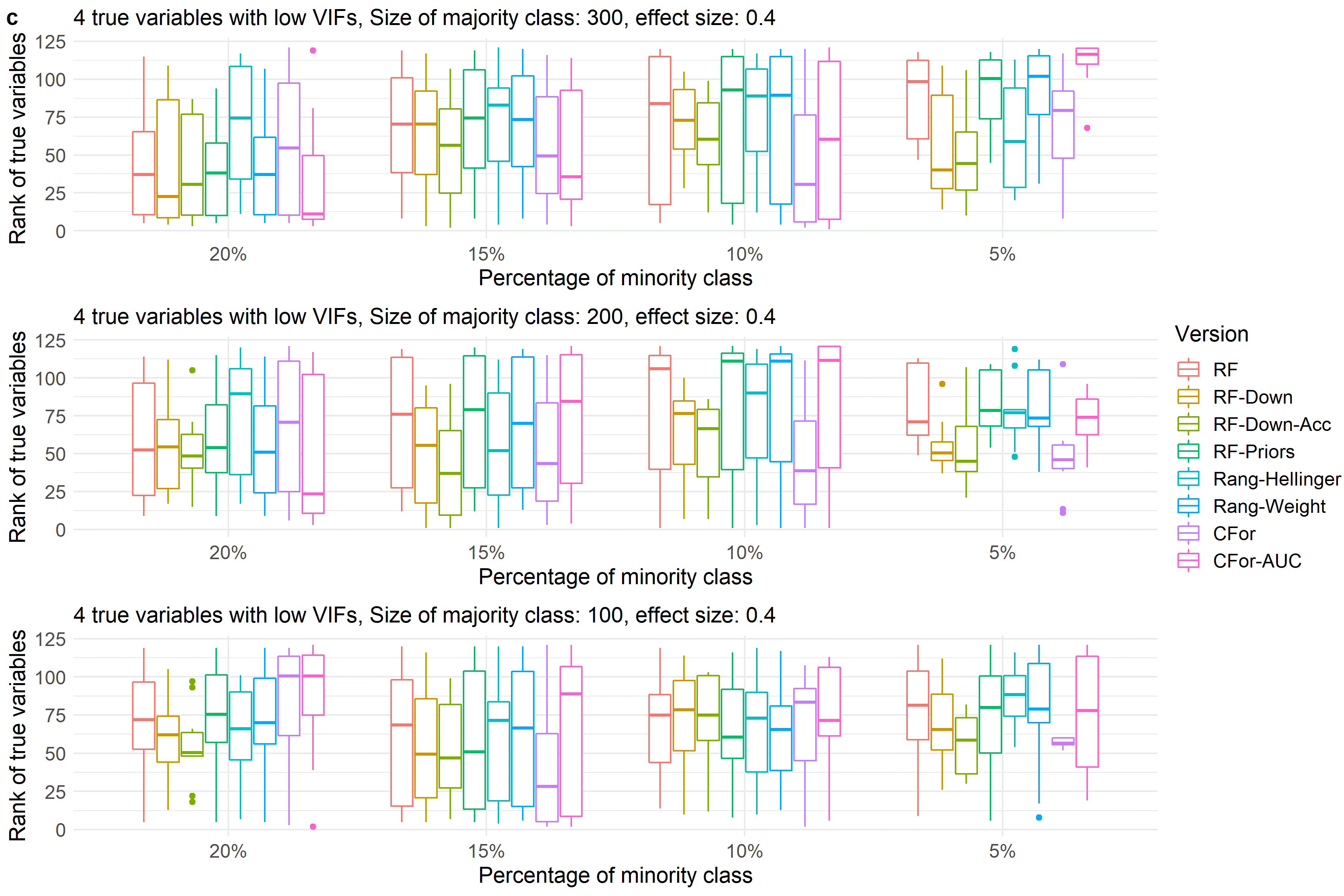
## Fig. S3

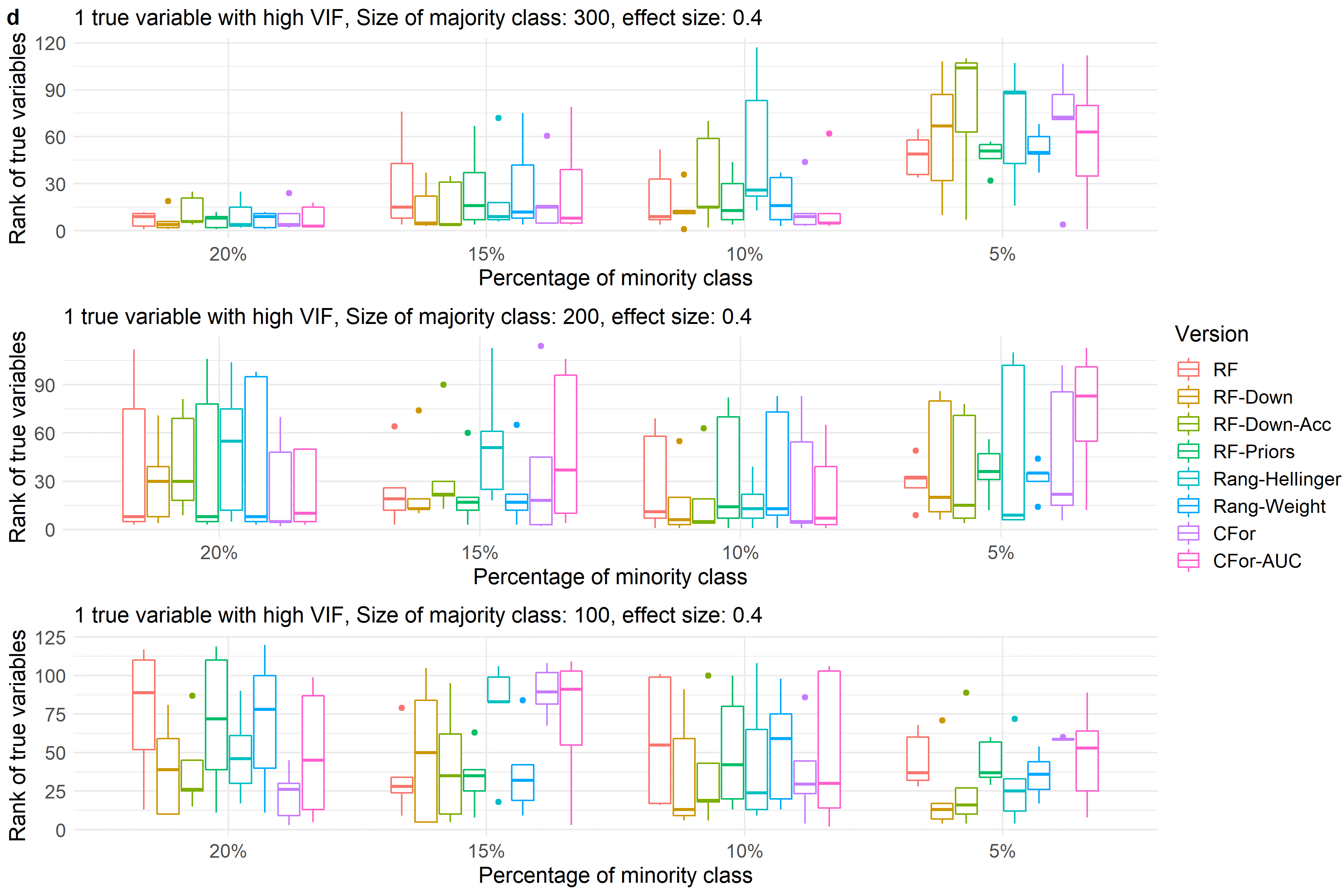
Boxplots showing ranks of permutation importance for true variables with a high effect size of 1.4 for simulated data with a) one true variable with low VIF, b) two true variables with low VIFs, c) four true variables with low VIFs, d) one true variable with high VIF, e) two true variables with high VIFs, f) four true variables with high VIFs.

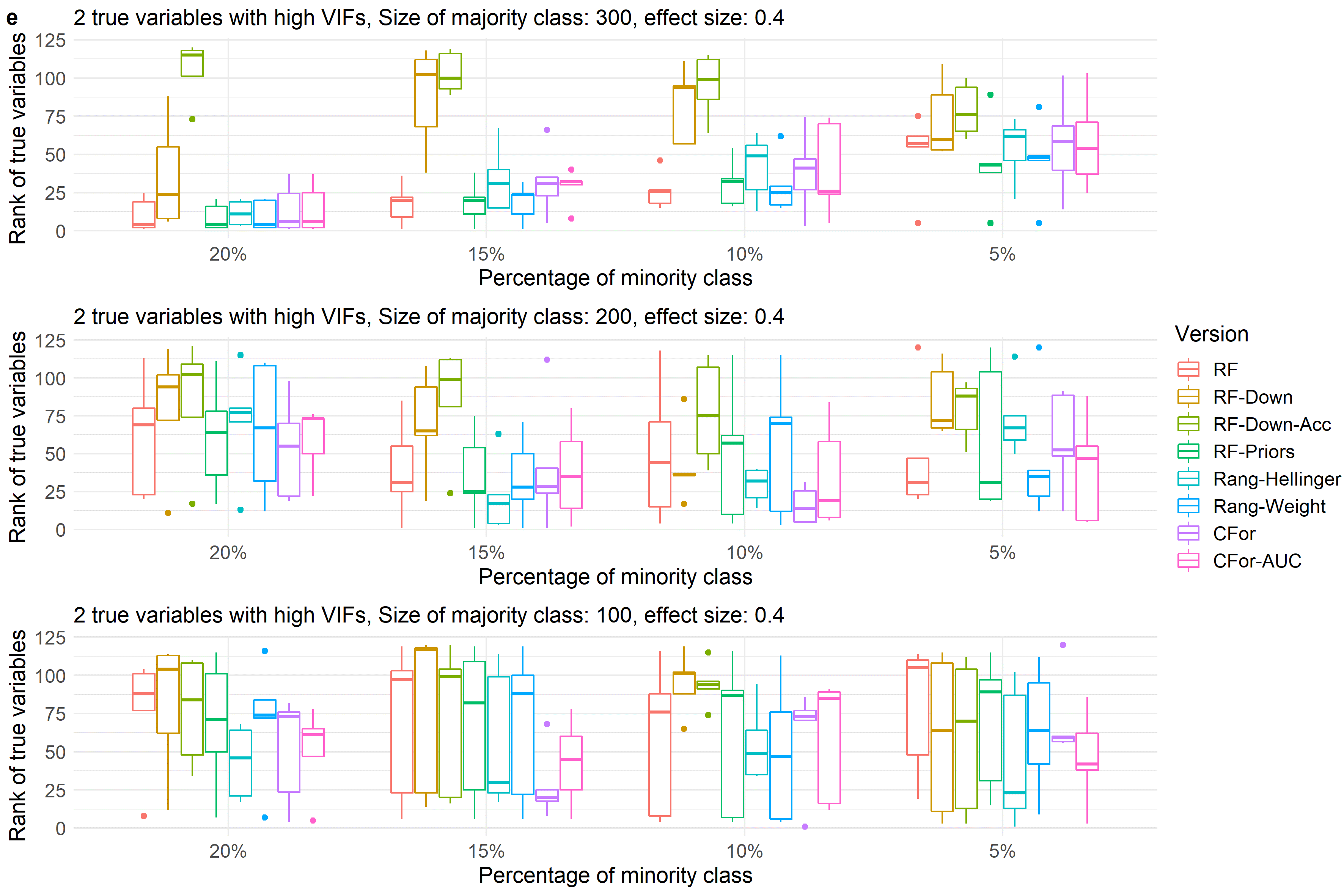
For each of the 60 simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors was ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 \* interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right).

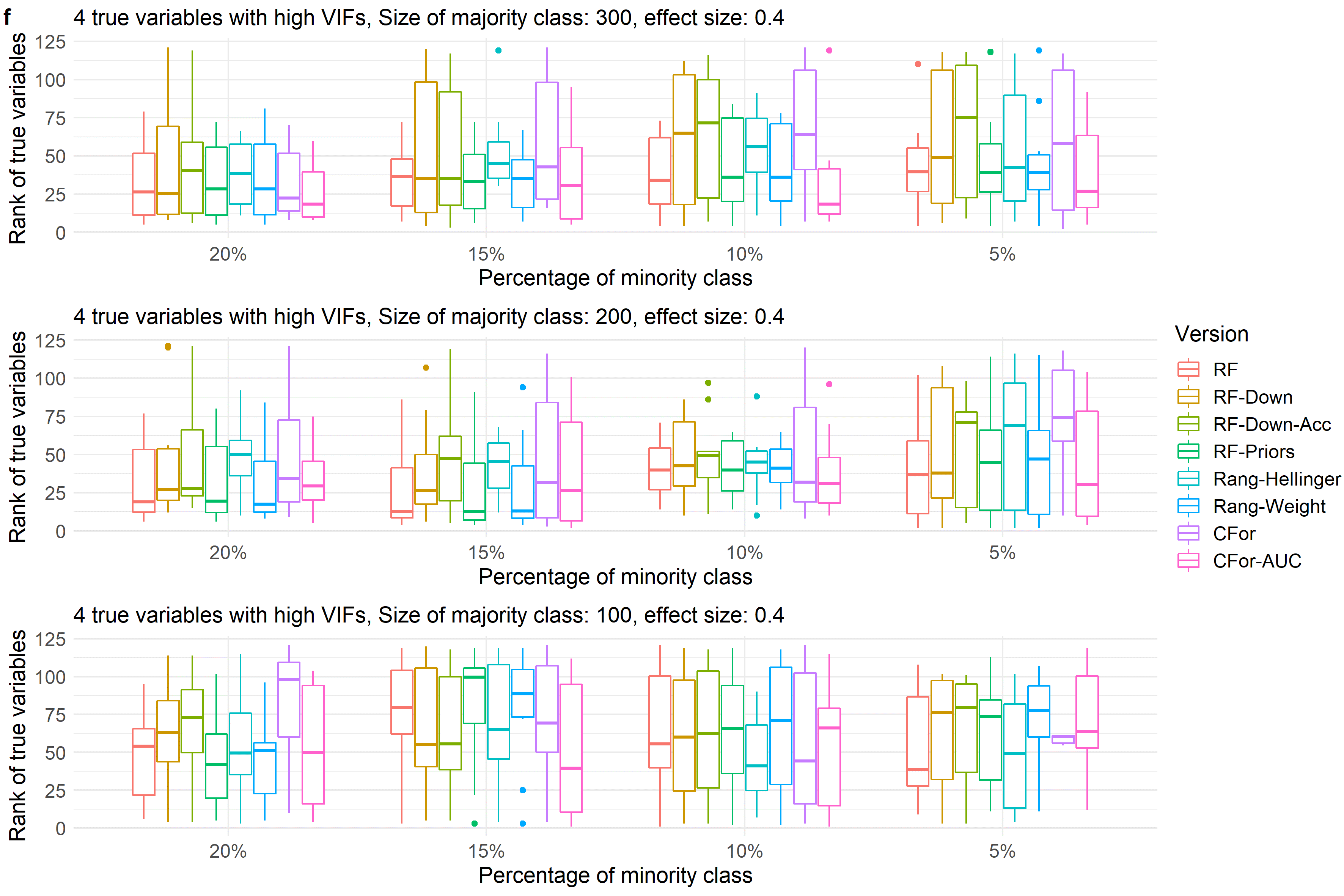








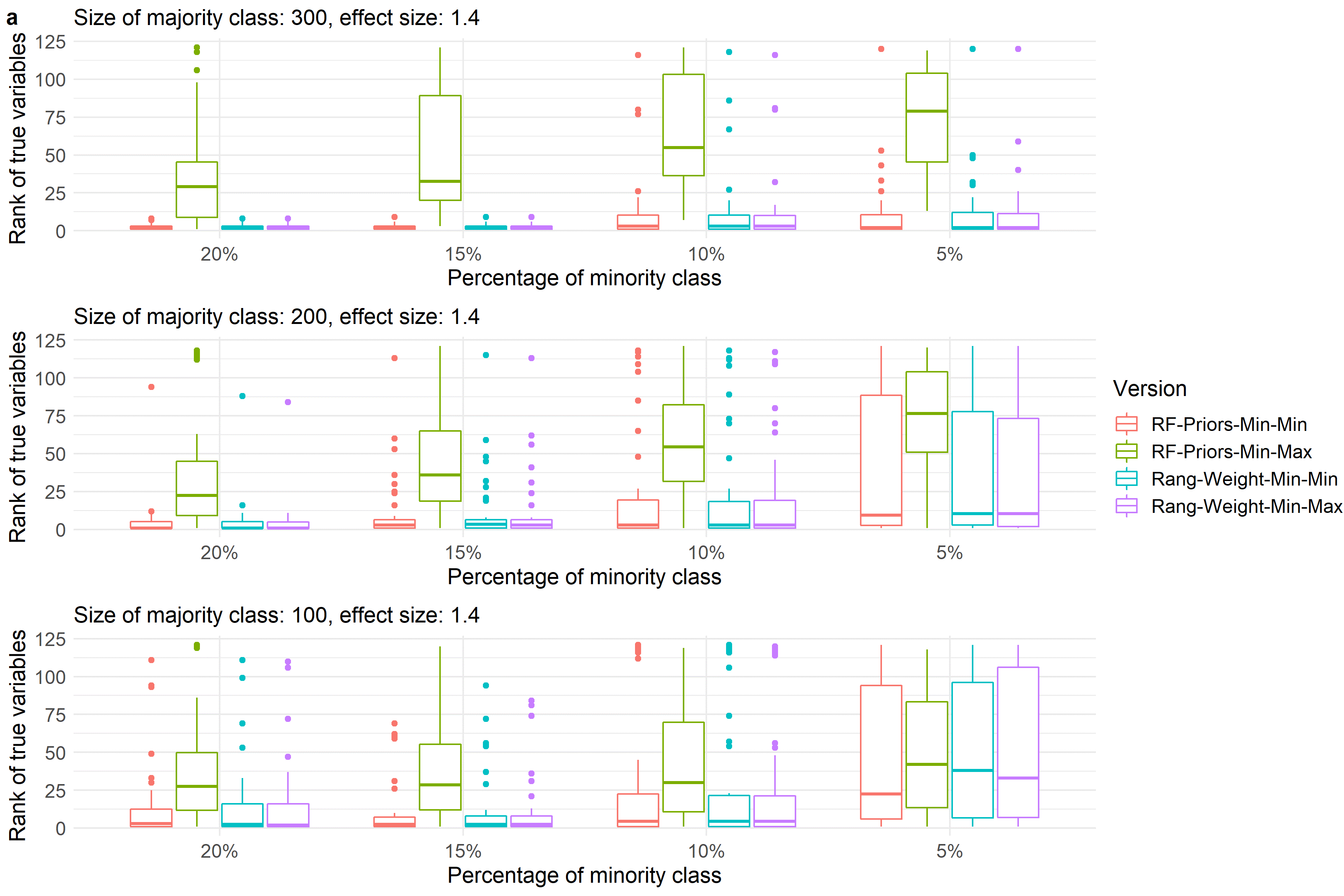


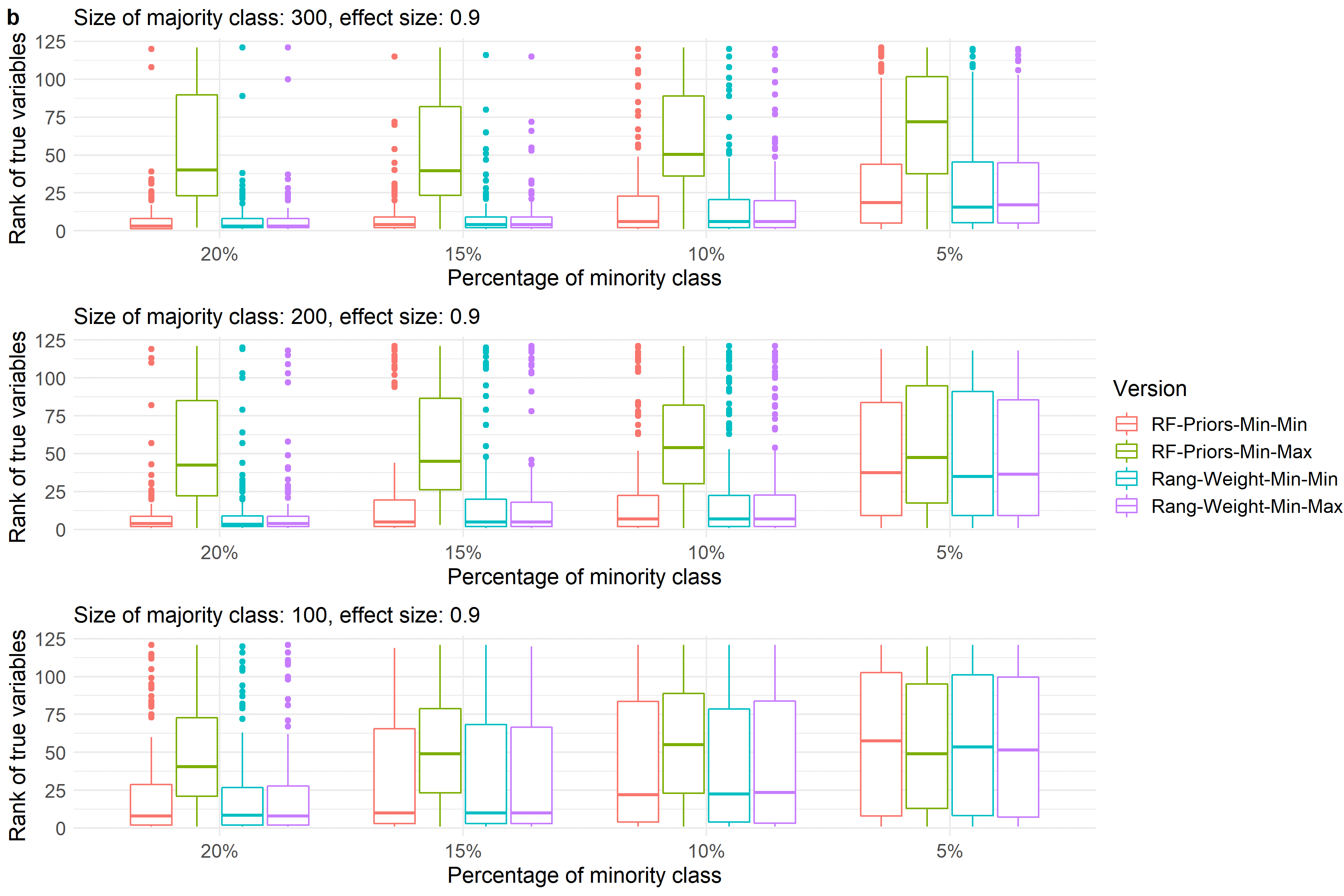


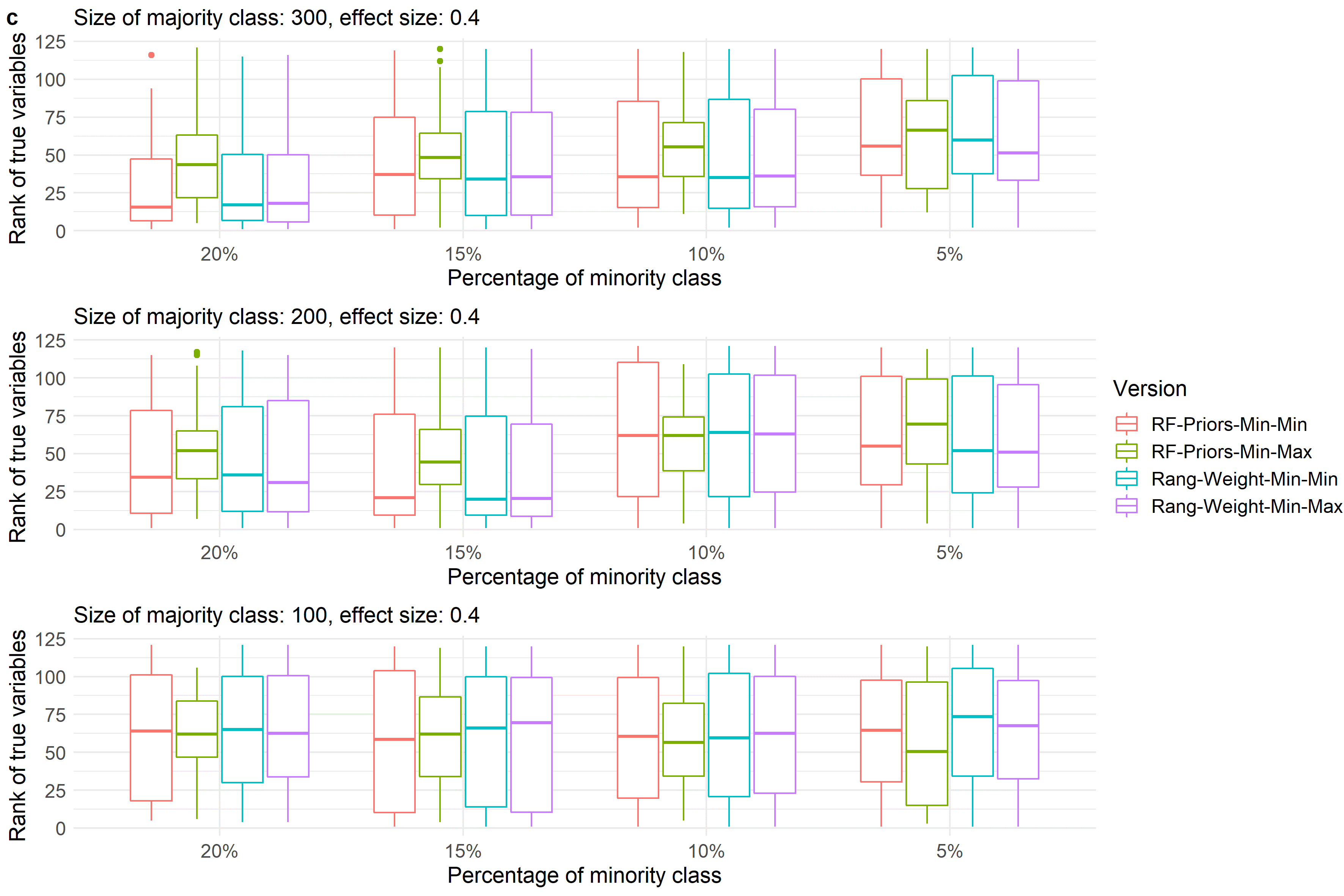
## Fig. S4

Boxplots showing ranks of permutation importance for true variables with a low effect size of 0.4 for simulated data with a) one true variable with low VIF, b) two true variables with low VIFs, c) four true variables with low VIFs, d) one true variable with high VIF, e) two true variables with high VIFs, f) four true variables with high VIFs.

For each of the 60 simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the eight algorithm versions. The true predictors was ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 \* interquartile range. Boxes represent the eight algorithm versions: default RF (RF), two downsampled versions (RF-Down, RF-Down-Acc), priors representing the proportion of class sizes (RF-Priors), Hellinger distance as splitting criterion (Rang-Hellinger), weighted random forest (Rang-Weight), random forest based on conditional inference trees (CFor), AUC permutation importance (CFor-AUC; from left to right).







## Fig. S5

Boxplots showing ranks of permutation importance for true variables with a) a high effect size of 1.4, b) a medium effect size of 0.9 and c) a low effect size of 0.4. Boxplots summarize ranks for true variables across simulated data with one true variable with low VIF, two true variables with low VIFs, four true variables with low VIFs, one true variable with high VIF, two true variables with high VIFs and four true variables with high VIFs.

For each of the simulated test sets with a total of 121 covariates, permutation importance was calculated 10 times for each of the four algorithm versions. The true predictors were ranked by the mean permutation importance averaged over the 10 repetitions. Boxes showing interquartile range and median; whiskers show the maximum of 1.5 \* interquartile range. Boxes represent the two algorithm versions with priors (RF-Priors) or weights (RF-Weight). For Min-Min the proportion of the minority class (Number of presences / Number of presences and absences) was assigned as the prior or weight for the minority class (presences) while the proportion of the majority class (Number of absences / Number of presences and absences) was assigned as the prior or weight for the majority class (absences). For Min-Max the proportion of the minority class was assigned as the prior or weight for the majority class (absences) while the proportion of the majority class was assigned as the prior or weight for the minority class (presences). From left to right, the boxes refer to RF-Priors-Min-Min, RF-Priors-Min-Max, Rang-Weight-Min-Min, Rang-Weight-Min-Max.