Calibration ensemble training on random subsets of different sizes

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

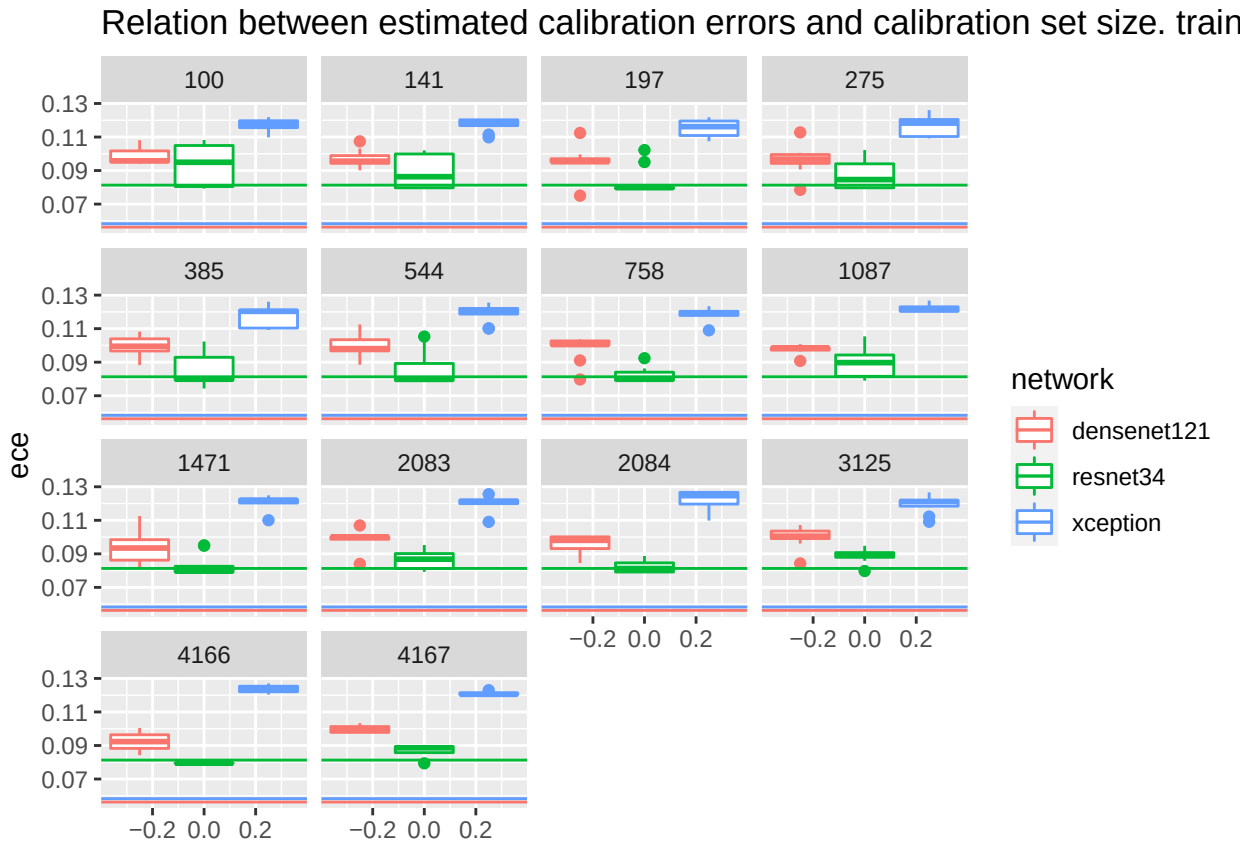
Experiment code is in the file calibration_subsets_sizes_experiment.py This experiment trains CalibrationEnsemble on random subsets of different sizes picked from the neural networks training set. Each subset size is tested several times. The goal of the experiment is to determine the best calibration subset size.

CIFAR-10

```
base_dir_tc <- "../data/data_train_val_half_c10/0/exp_subsets_sizes_calibration_outputs/" # nolint
metrics_ens_tc <- read.csv(file.path(base_dir_tc, "ens_metrics_train.csv"), stringsAsFactors = TRUE)
metrics_net_tc <- read.csv(file.path(base_dir_tc, "net_metrics_train.csv"), stringsAsFactors = TRUE)
metrics_net_cal_tc <- read.csv(file.path(base_dir_tc, "net_cal_metrics_train.csv"))

metrics_ens_vc <- read.csv(file.path(base_dir_tc, "ens_metrics_val.csv"), stringsAsFactors = TRUE)
metrics_net_vc <- read.csv(file.path(base_dir_tc, "net_metrics_val.csv"), stringsAsFactors = TRUE)
metrics_net_cal_vc <- read.csv(file.path(base_dir_tc, "net_cal_metrics_val.csv"))

metrics_ens_tc$cal_type <- "tc"
metrics_net_tc$cal_type <- "tc"
metrics_net_cal_tc$cal_type <- "tc"

metrics_ens_vc$cal_type <- "vc"
metrics_net_vc$cal_type <- "vc"
```

```
metrics_net_cal_vc$cal_type <- "vc"

metrics_ens <- rbind(metrics_ens_tc, metrics_ens_vc)
metrics_net <- rbind(metrics_net_tc, metrics_net_vc)
metrics_net_cal <- rbind(metrics_net_cal_tc, metrics_net_cal_vc)
```

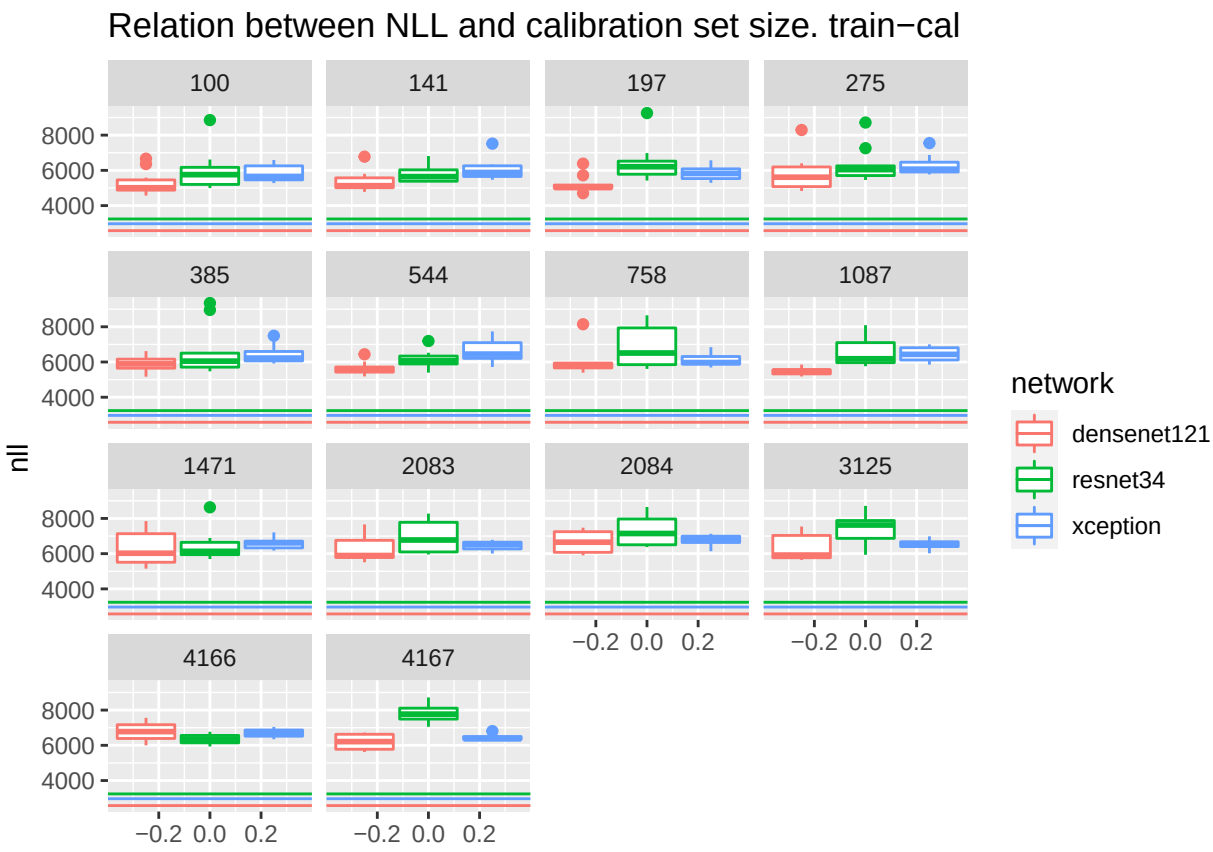
```
net_calibrations_ece <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "tc"),
    mapping = aes(y = ece, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "tc"),
    mapping = aes(yintercept = ece, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between estimated calibration errors and calibration set size. train-cal")
net_calibrations_ece
```



As we can see, calibration on training data worsens the estimated calibration error of the networks.

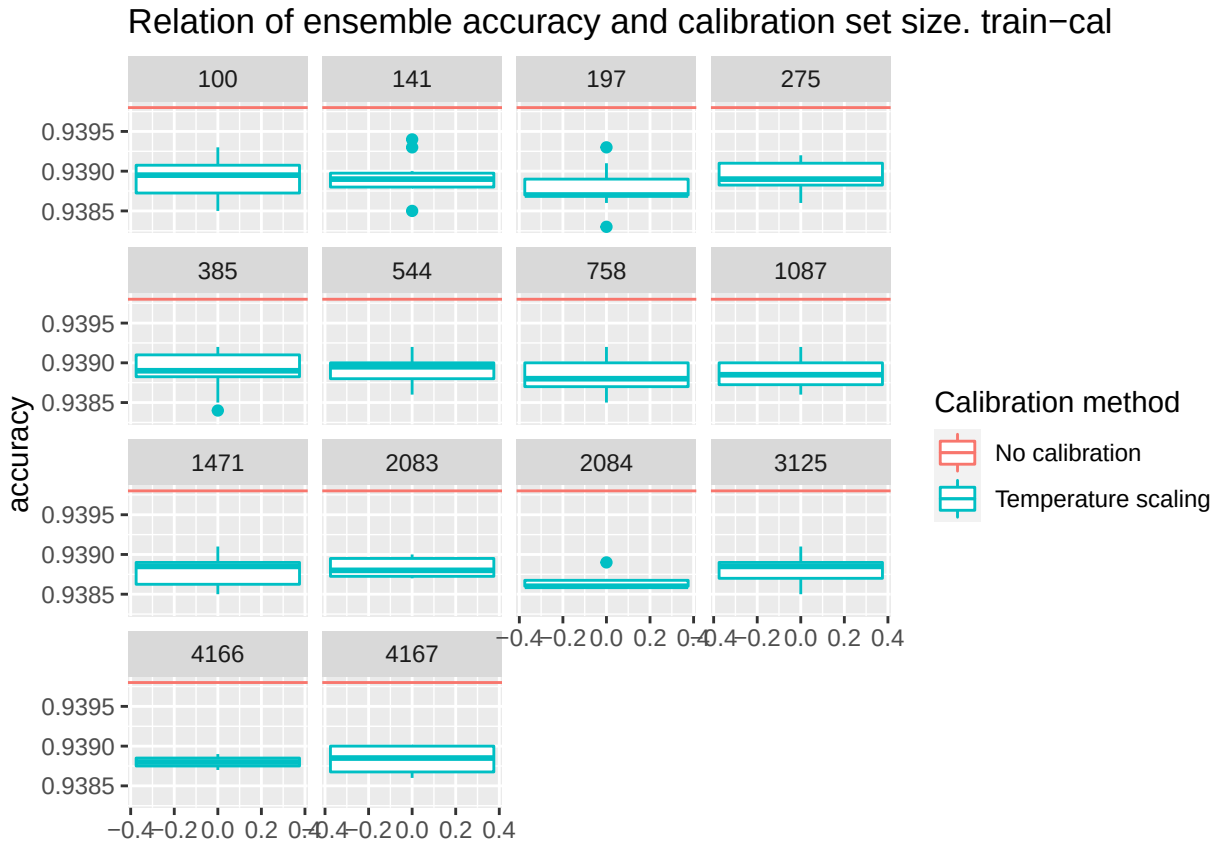
```
net_calibrations_nll <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "tc"),
    mapping = aes(y = nll, color = network)
  ) +
  geom_hline(
```

```
    data = metrics_net %>% filter(cal_type == "tc"),
    mapping = aes(yintercept = nll, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between NLL and calibration set size. train-cal")
net_calibrations_nll
```



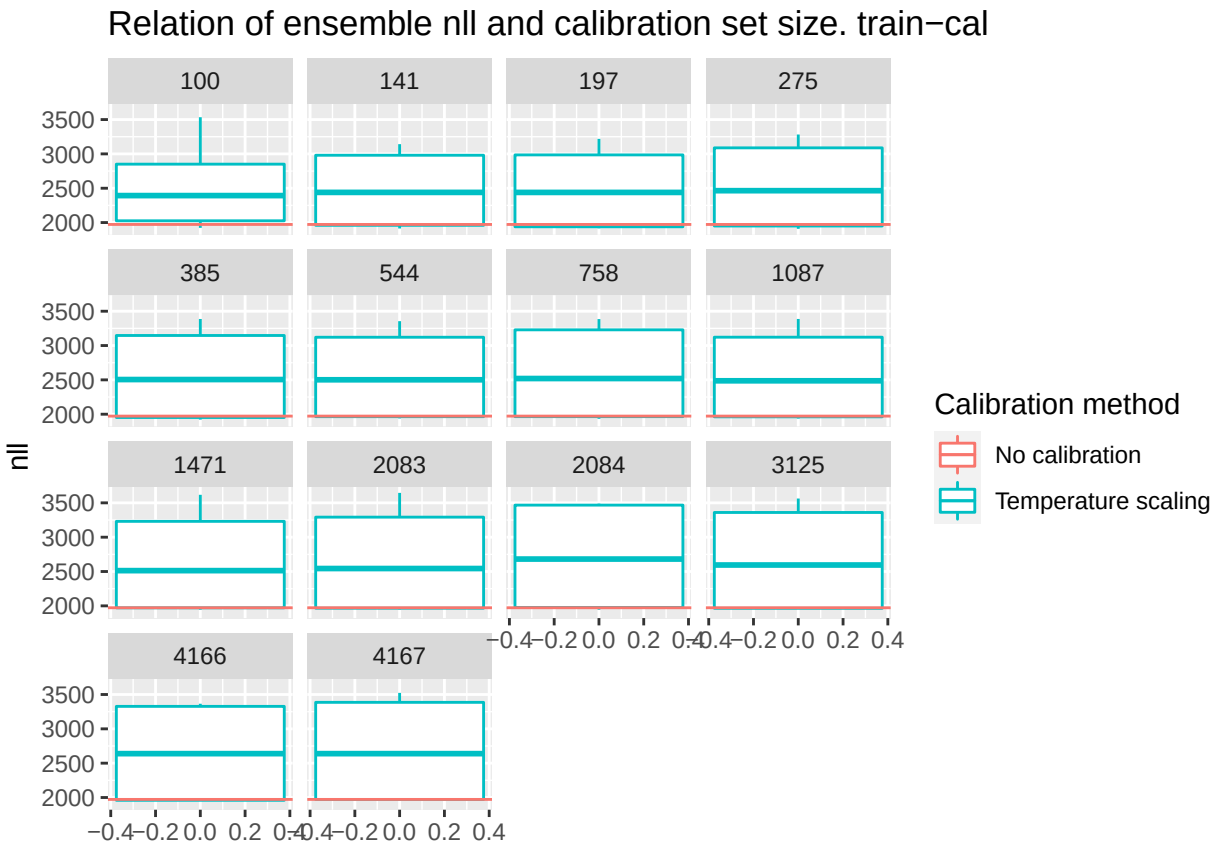
Calibration on training data also worsens the negative log likelihood.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration" &
      cal_type == "tc")),
    mapping = aes(y = accuracy, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration",
      cal_type == "tc") %>%
      select(!train_size)),
    mapping = aes(yintercept = accuracy, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble accuracy and calibration set size. train-cal")
ens_accuracy
```



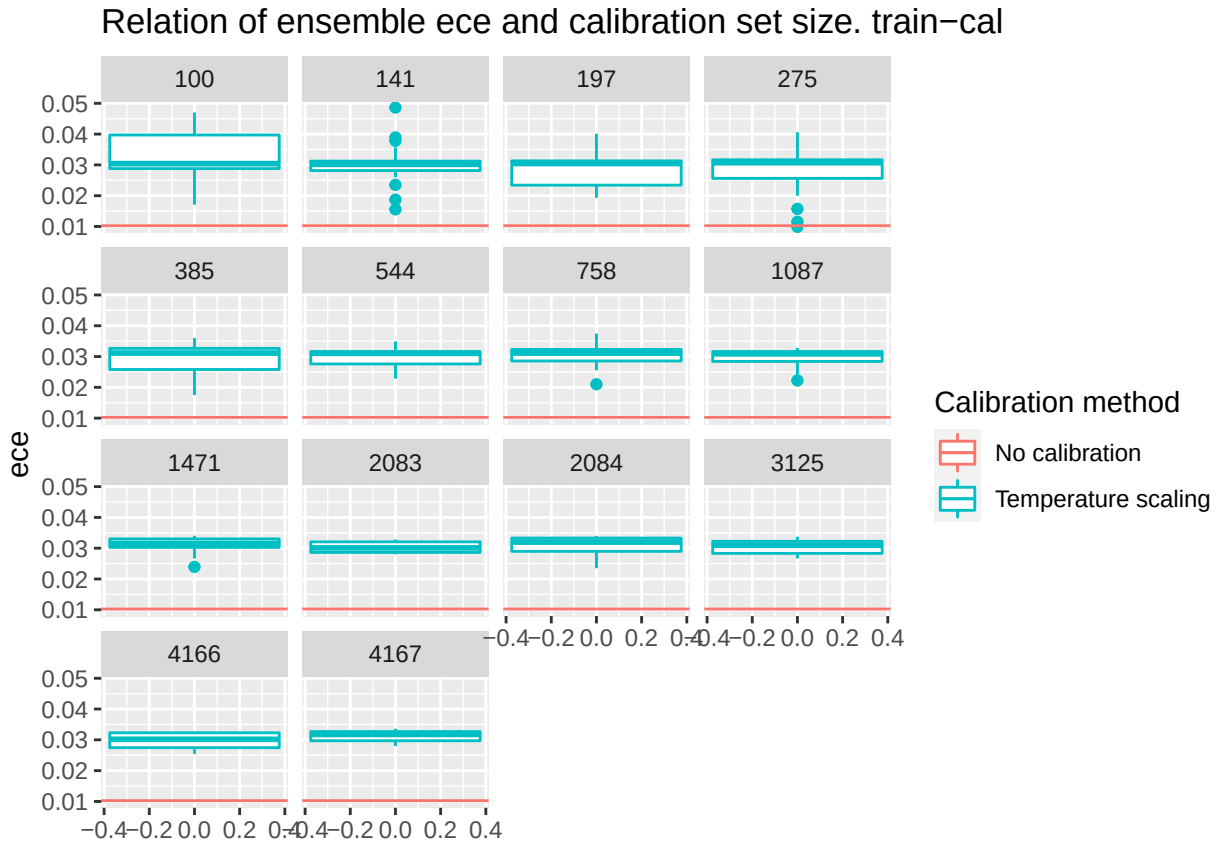
Calibration on train data worsens the ensemble accuracy.

```
ens_nll <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = nll, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = nll, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble nll and calibration set size. train-cal")
ens_nll
```



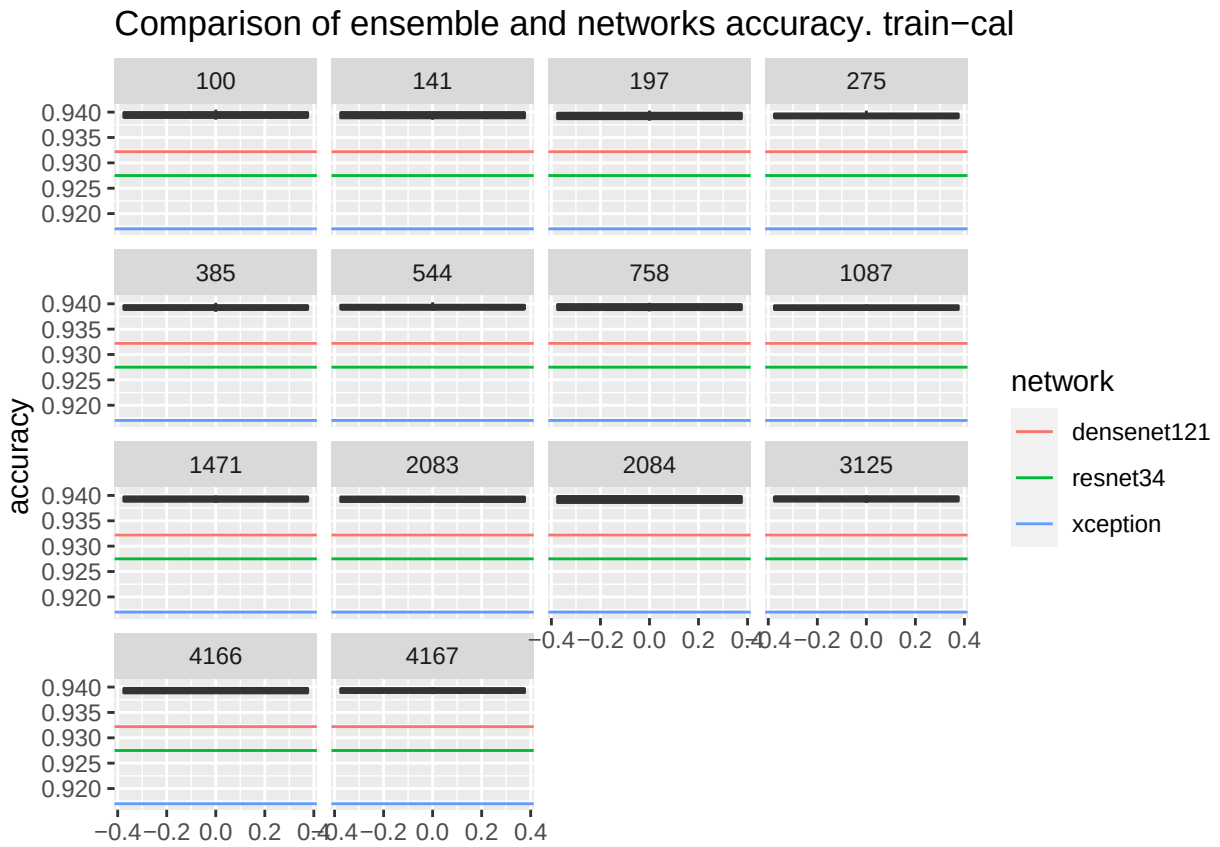
Calibration on train data worsens the ensemble nll.

```
ens_ece <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = ece, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = ece, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble ece and calibration set size. train-cal")
ens_ece
```



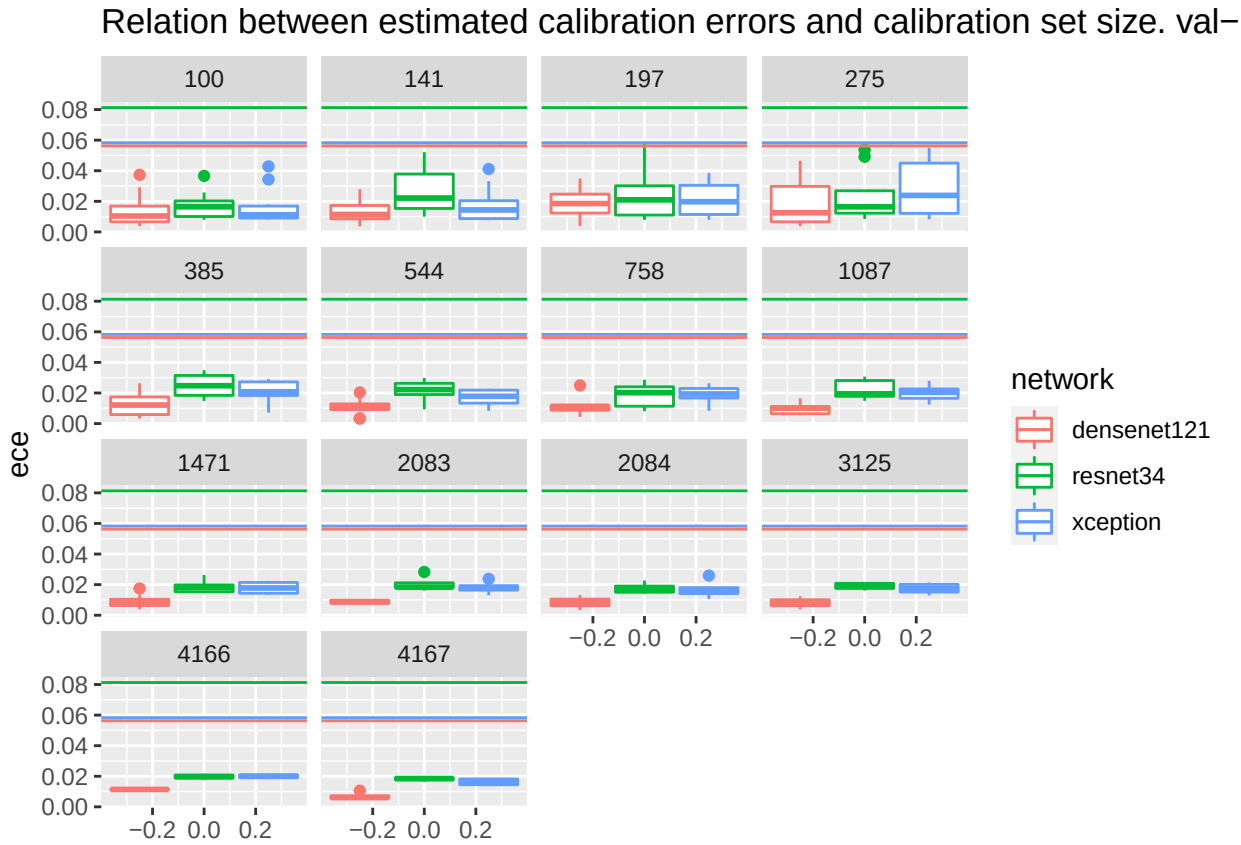
Calibration on train data worsens the ensemble estimated calibration error.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = accuracy)
  ) +
  geom_hline(
    data = metrics_net,
    mapping = aes(yintercept = accuracy, color = network)
  ) +
  facet_wrap(~train_size) +
  ggtitle("Comparison of ensemble and networks accuracy. train-cal")
ens_accuracy
```



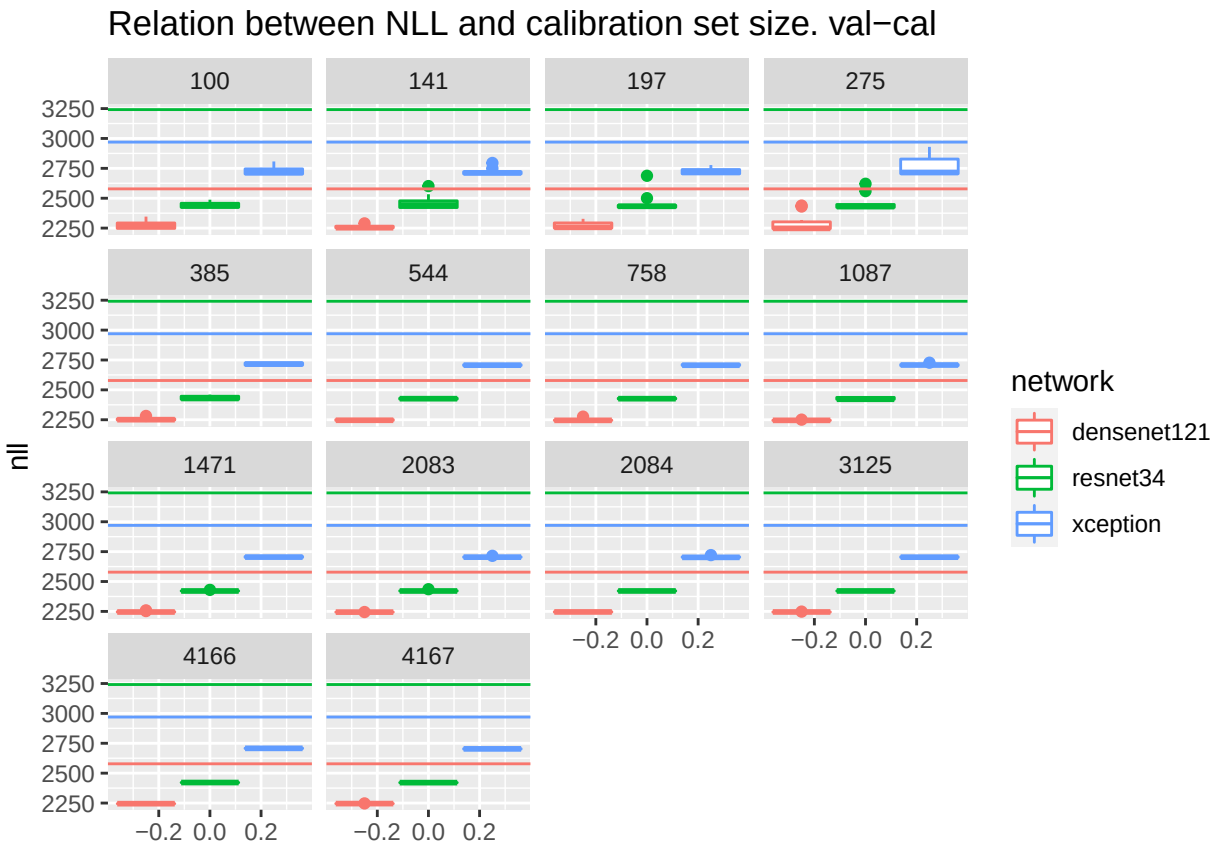
Since calibration on training data proved to be contraproductive, we also tested calibration on validation data.

```
net_calibrations_ece <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "vc"),
    mapping = aes(y = ece, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "vc"),
    mapping = aes(yintercept = ece, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between estimated calibration errors and calibration set size. val-cal")
net_calibrations_ece
```



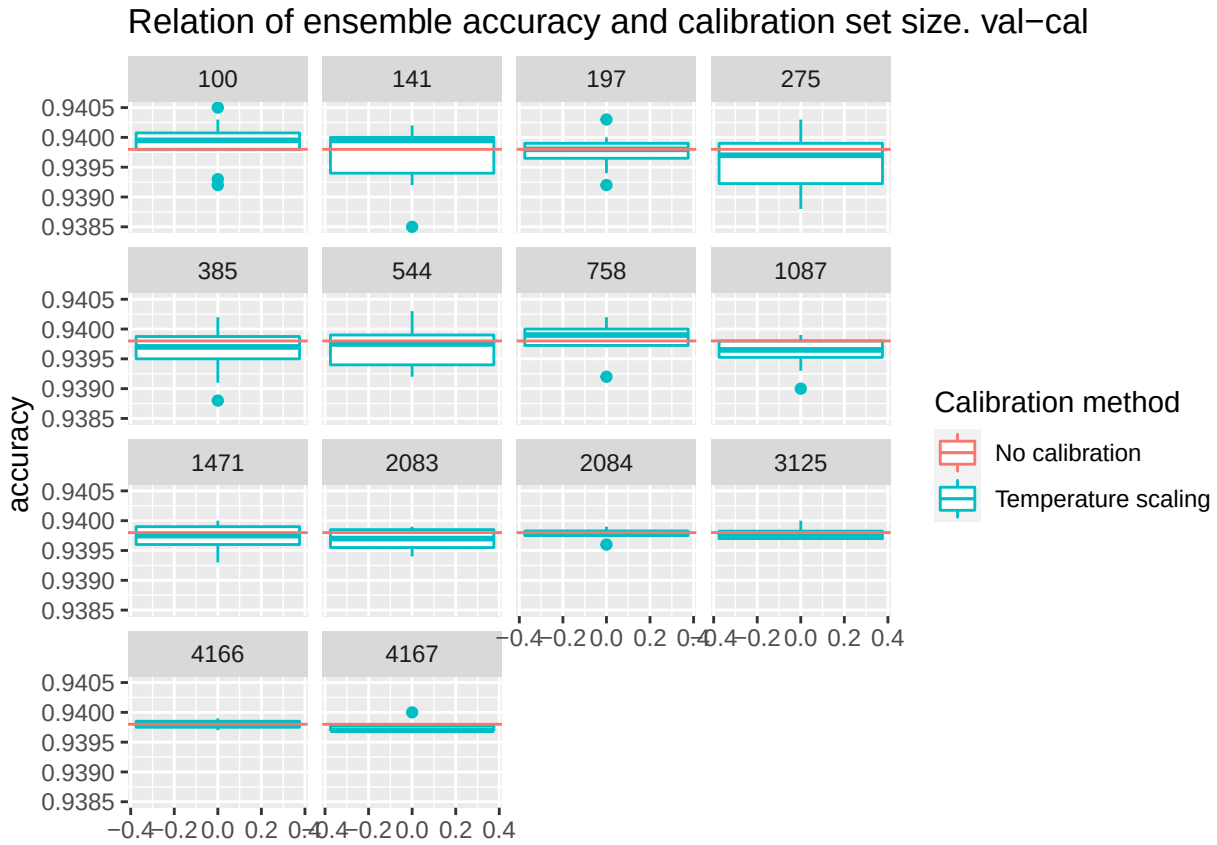
Calibration on validation data improves estimated calibration error for networks on testing data.

```
net_calibrations_nll <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "vc"),
    mapping = aes(y = nll, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "vc"),
    mapping = aes(yintercept = nll, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between NLL and calibration set size. val-cal")
net_calibrations_nll
```



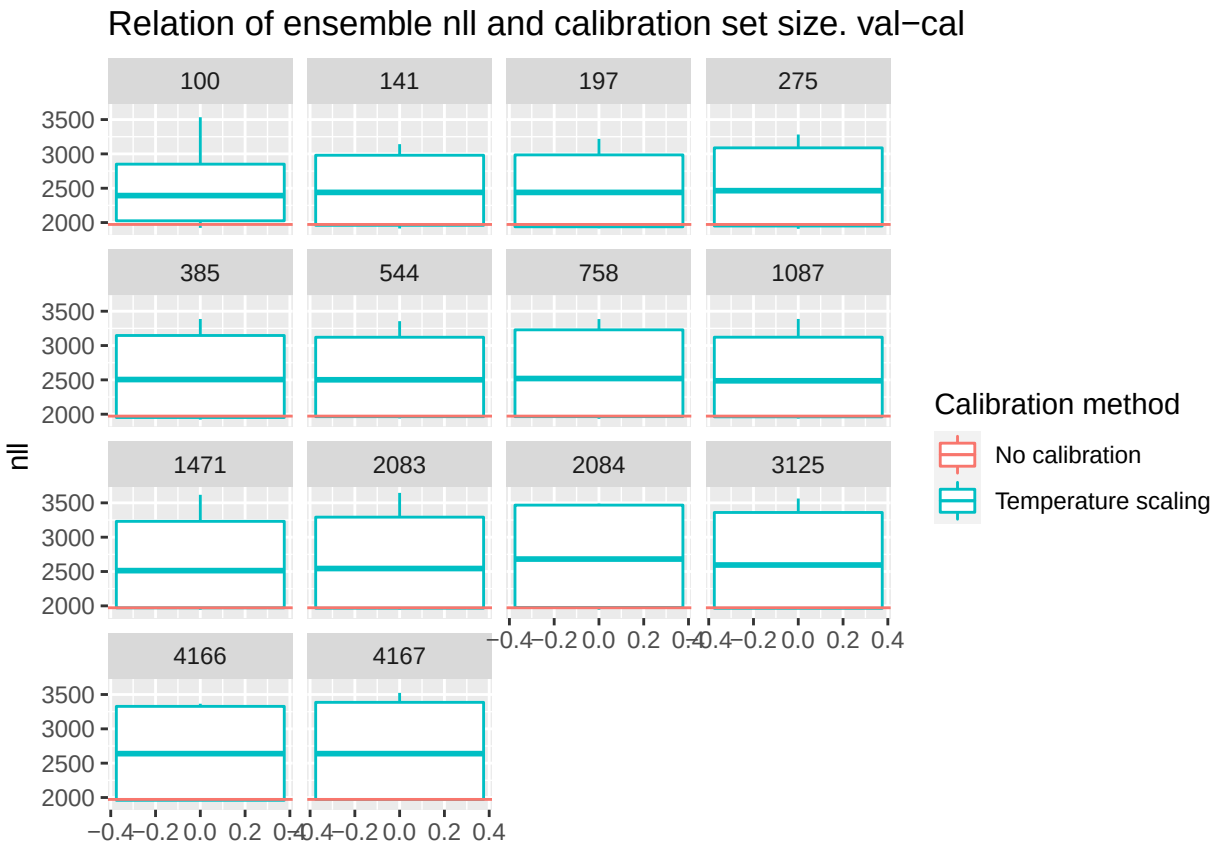
Calibration on validation data improves nll for networks on test set.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration" &
      cal_type == "vc")),
    mapping = aes(y = accuracy, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration",
      cal_type == "vc") %>%
      select(!train_size)),
    mapping = aes(yintercept = accuracy, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble accuracy and calibration set size. val-cal")
ens_accuracy
```



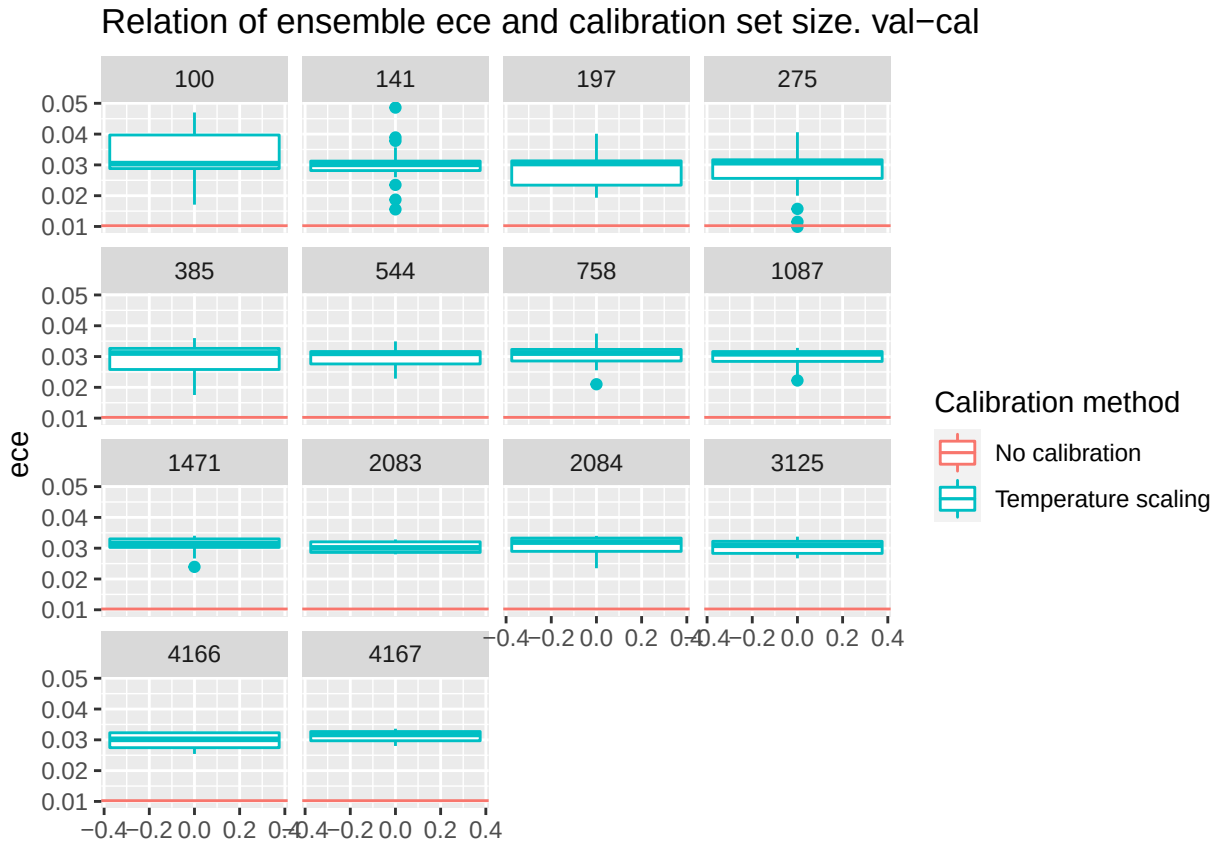
Calibration on validation data doesn't have large impact on ensemble accuracy.

```
ens_nll <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = nll, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = nll, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble nll and calibration set size. val-cal")
ens_nll
```



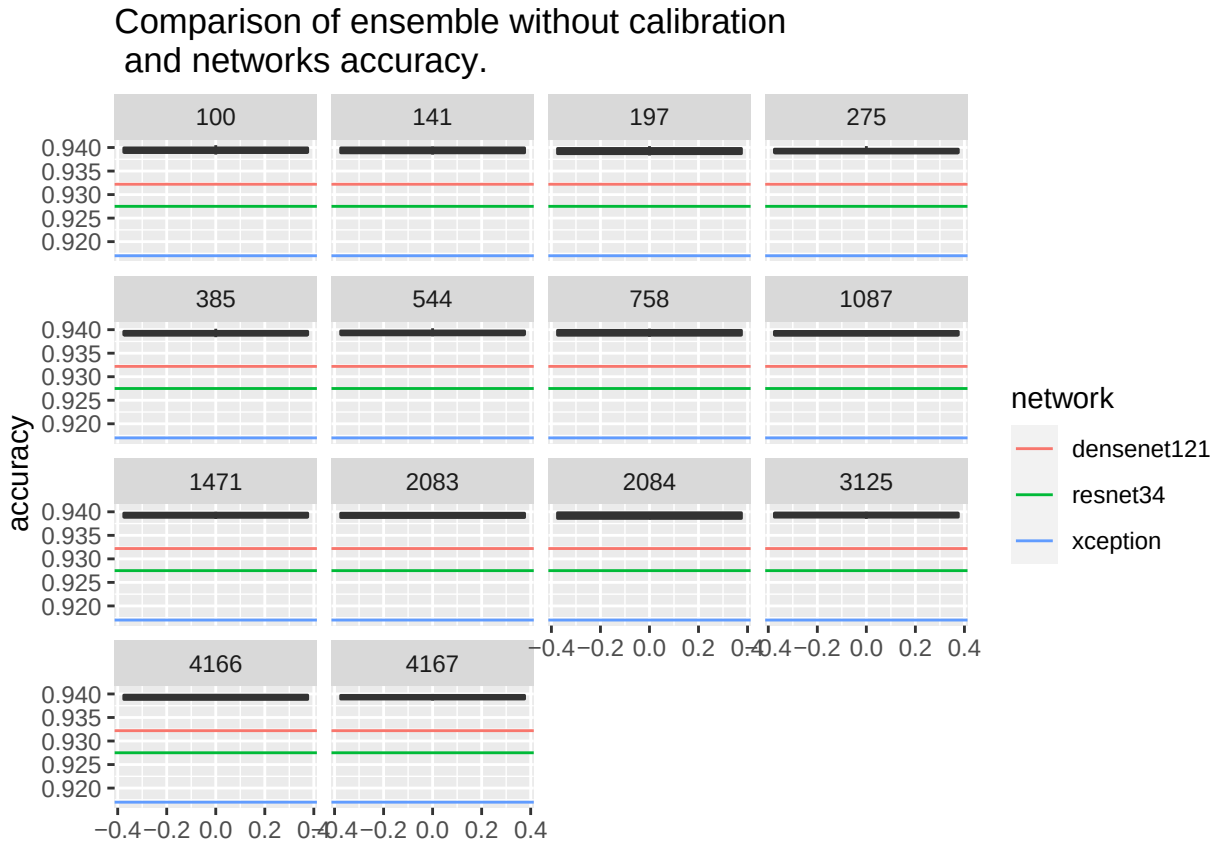
Calibration on validation data worsens the ensemble nll.

```
ens_ece <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = ece, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = ece, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble ece and calibration set size. val-cal")
ens_ece
```



Calibration on validation data worsens the ensemble estimated calibration error.

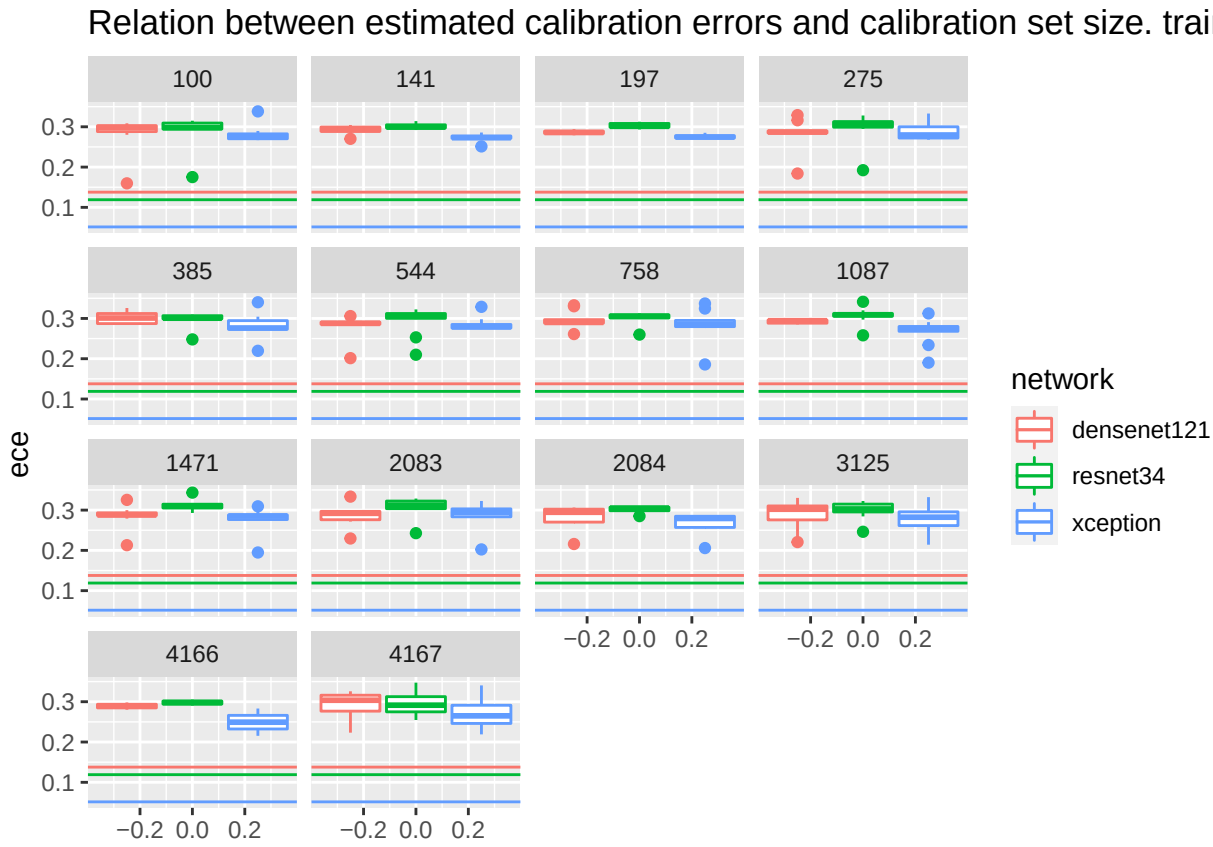
```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = accuracy)
  ) +
  geom_hline(
    data = metrics_net,
    mapping = aes(yintercept = accuracy, color = network)
  ) +
  facet_wrap(~train_size) +
  ggtitle("Comparison of ensemble without calibration \n and networks accuracy.")
ens_accuracy
```



CIFAR100

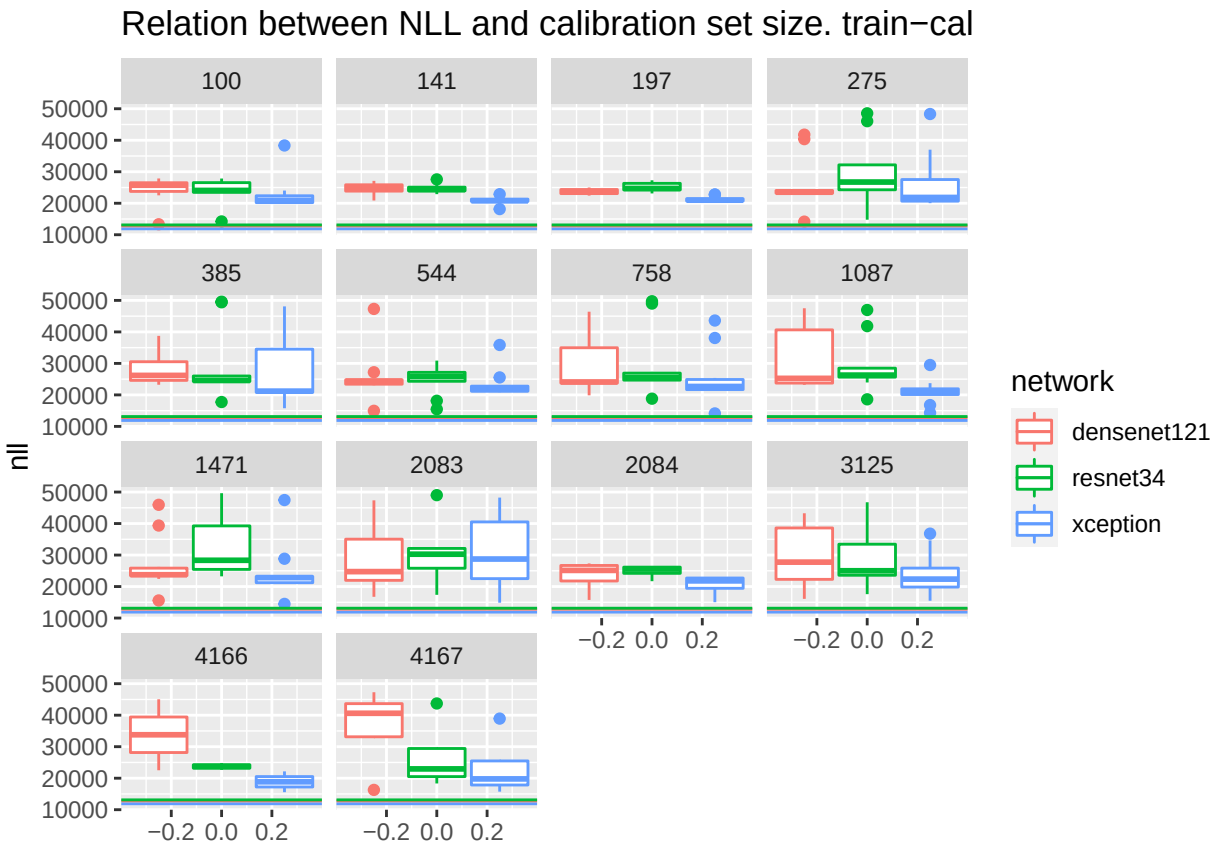
```
base_dir_tc <- "../data/data_train_val_half_c100/0/exp_subsets_sizes_calibration_outputs/" # nolint
metrics_ens_tc <- read.csv(file.path(base_dir_tc, "ens_metrics_train.csv"), stringsAsFactors = TRUE)
metrics_net_tc <- read.csv(file.path(base_dir_tc, "net_metrics_train.csv"), stringsAsFactors = TRUE)
metrics_net_cal_tc <- read.csv(file.path(base_dir_tc, "net_cal_metrics_train.csv"))

metrics_ens_vc <- read.csv(file.path(base_dir_tc, "ens_metrics_val.csv"), stringsAsFactors = TRUE)
metrics_net_vc <- read.csv(file.path(base_dir_tc, "net_metrics_val.csv"), stringsAsFactors = TRUE)
metrics_net_cal_vc <- read.csv(file.path(base_dir_tc, "net_cal_metrics_val.csv"))

metrics_ens_tc$cal_type <- "tc"
metrics_net_tc$cal_type <- "tc"
metrics_net_cal_tc$cal_type <- "tc"

metrics_ens_vc$cal_type <- "vc"
metrics_net_vc$cal_type <- "vc"
metrics_net_cal_vc$cal_type <- "vc"

metrics_ens <- rbind(metrics_ens_tc, metrics_ens_vc)
metrics_net <- rbind(metrics_net_tc, metrics_net_vc)
metrics_net_cal <- rbind(metrics_net_cal_tc, metrics_net_cal_vc)
```

```
net_calibrations_ece <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "tc"),
    mapping = aes(y = ece, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "tc"),
    mapping = aes(yintercept = ece, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between estimated calibration errors and calibration set size. train-cal")
net_calibrations_ece
```



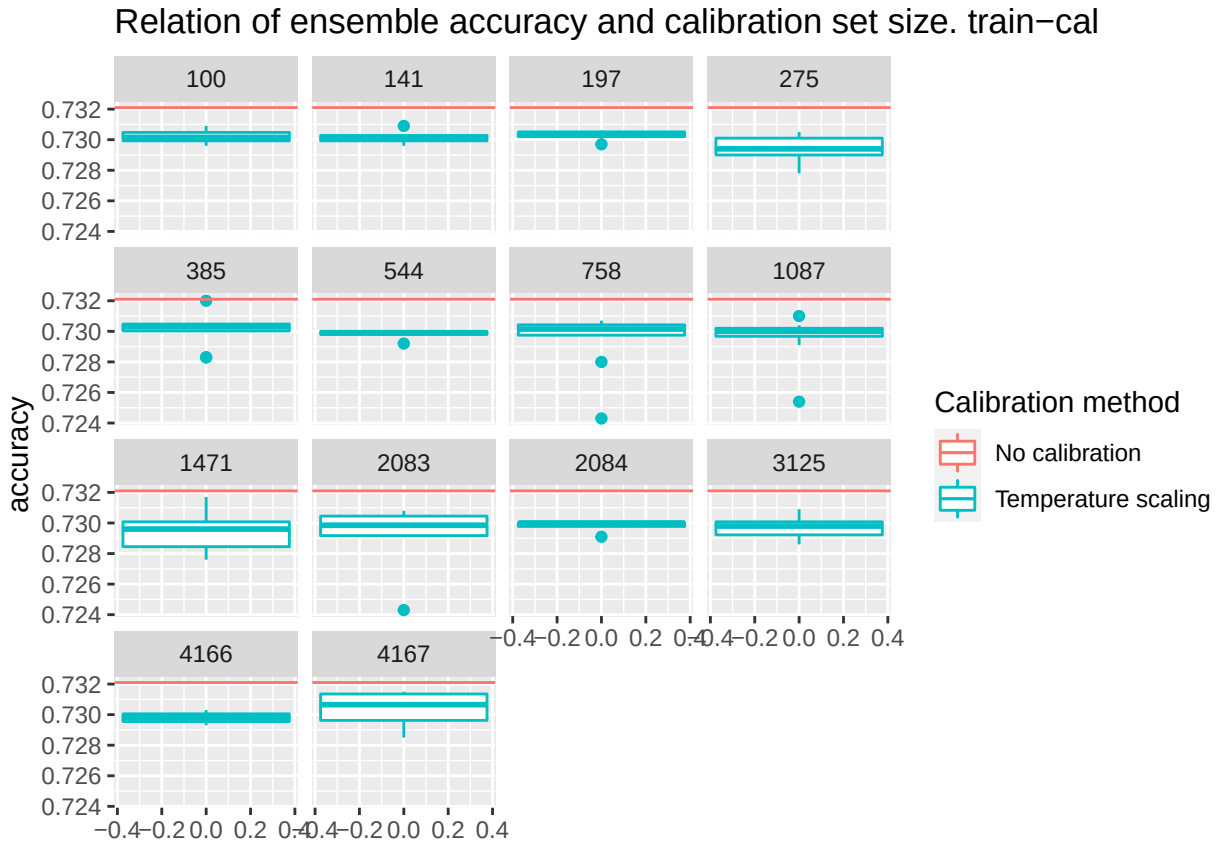
As we can see, calibration on training data worsens the estimated calibration error of the networks.

```
net_calibrations_nll <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "tc"),
    mapping = aes(y = nll, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "tc"),
    mapping = aes(yintercept = nll, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between NLL and calibration set size. train-cal")
net_calibrations_nll
```



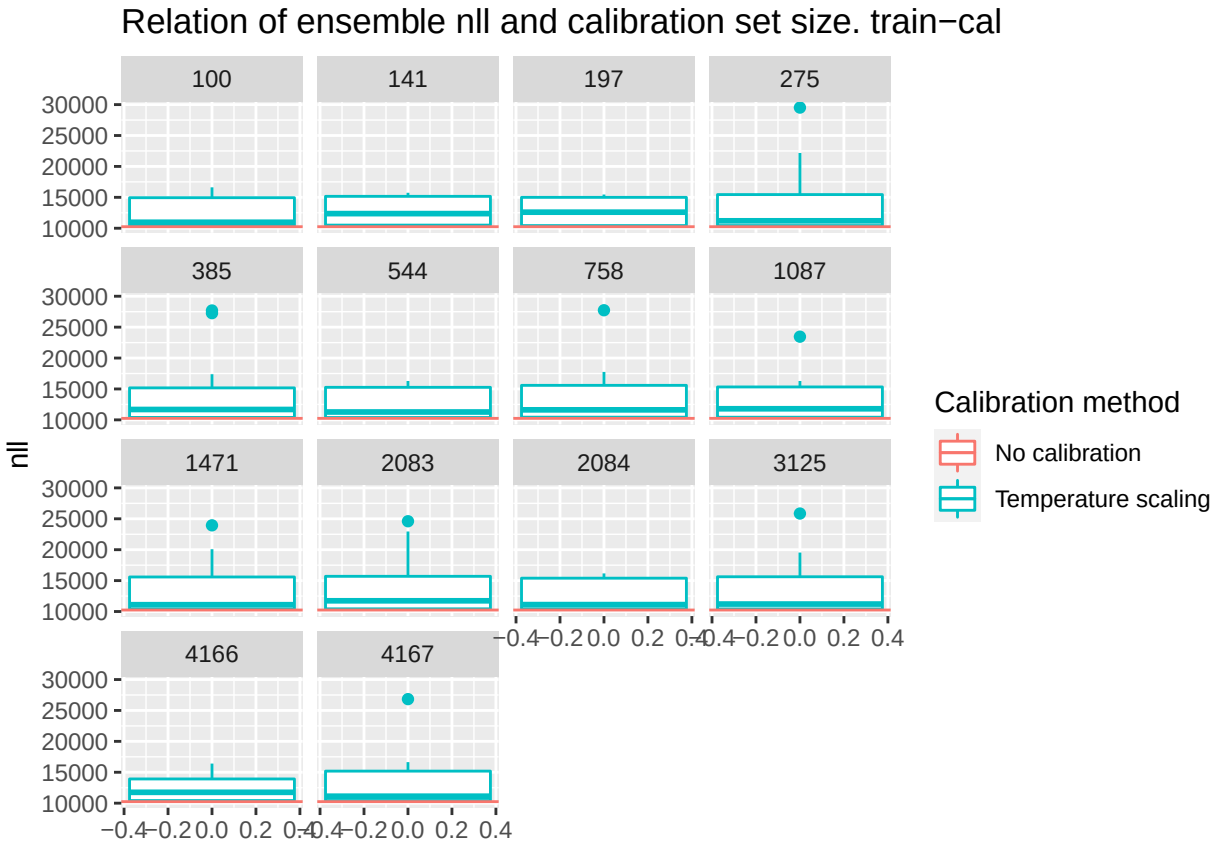
Calibration on training data also worsens the negative log likelihood.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration" &
      cal_type == "tc")),
    mapping = aes(y = accuracy, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration",
      cal_type == "tc") %>%
      select(!train_size)),
    mapping = aes(yintercept = accuracy, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble accuracy and calibration set size. train-cal")
ens_accuracy
```



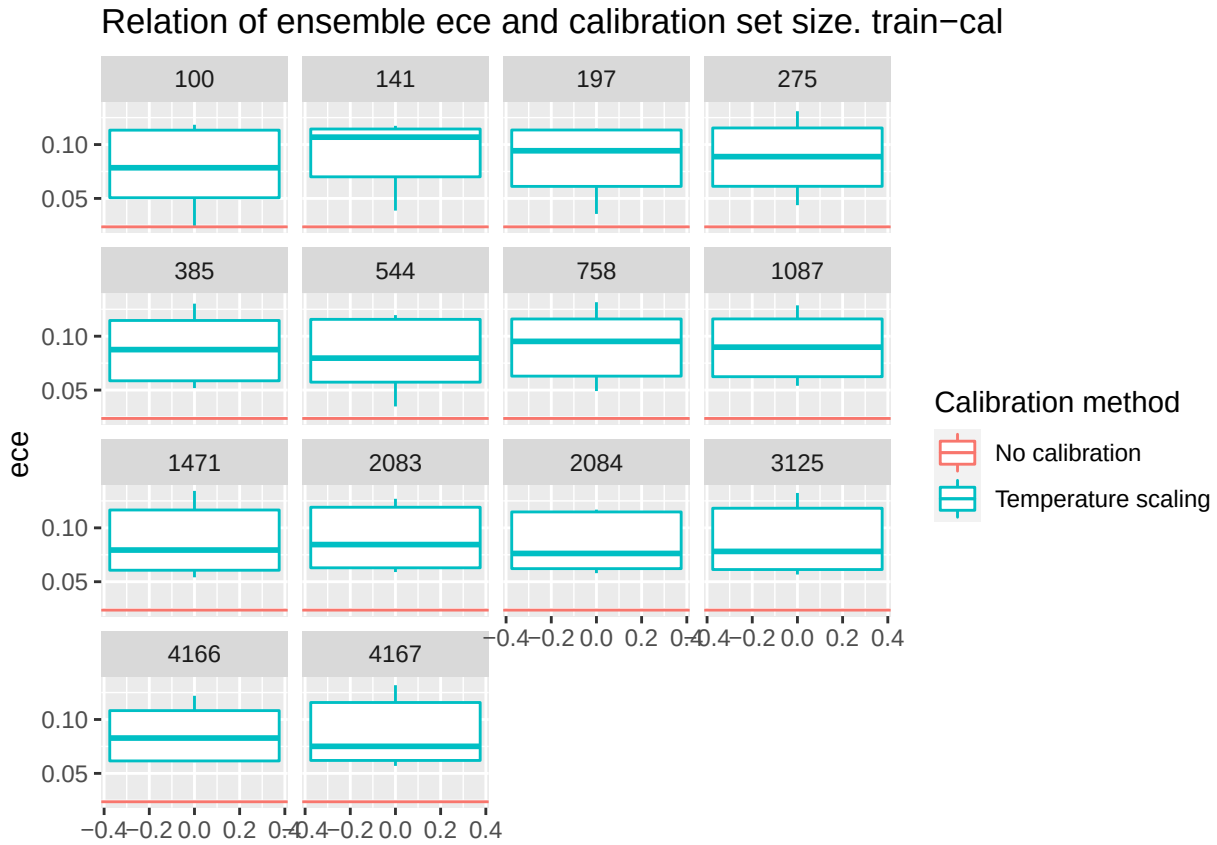
Calibration on train data worsens the ensemble accuracy.

```
ens_nll <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = nll, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = nll, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble nll and calibration set size. train-cal")
ens_nll
```



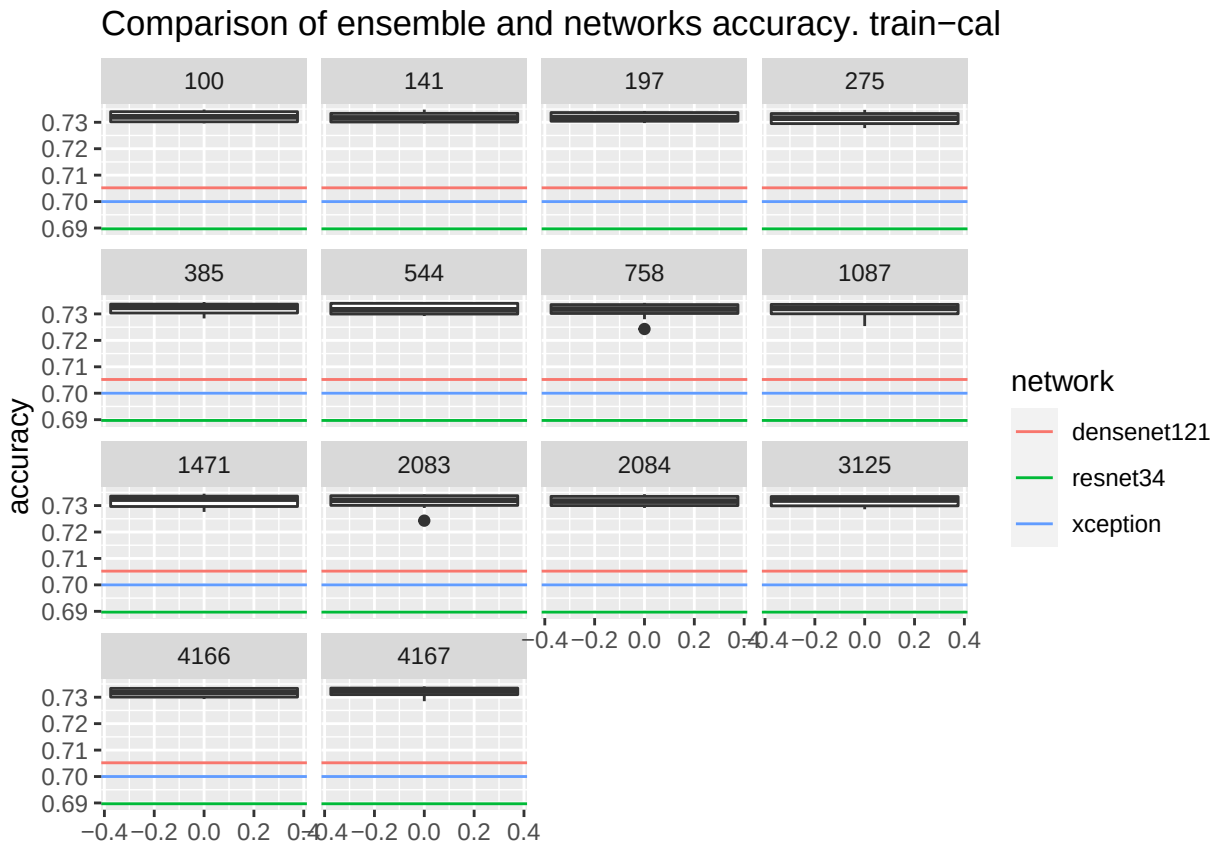
Calibration on train data worsens the ensemble nll.

```
ens_ece <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = ece, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = ece, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble ece and calibration set size. train-cal")
ens_ece
```



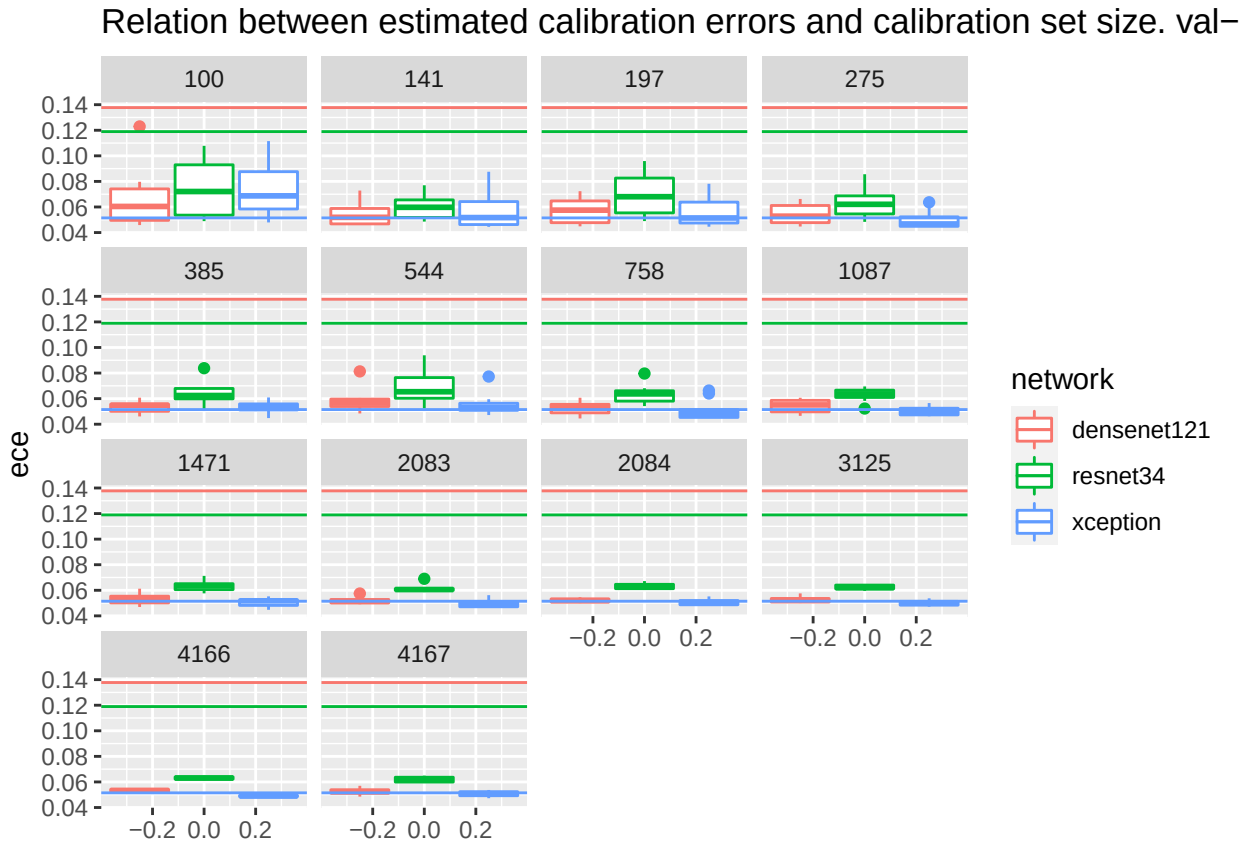
Calibration on train data worsens the ensemble estimated calibration error.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = accuracy)
  ) +
  geom_hline(
    data = metrics_net,
    mapping = aes(yintercept = accuracy, color = network)
  ) +
  facet_wrap(~train_size) +
  ggtitle("Comparison of ensemble and networks accuracy. train-cal")
ens_accuracy
```



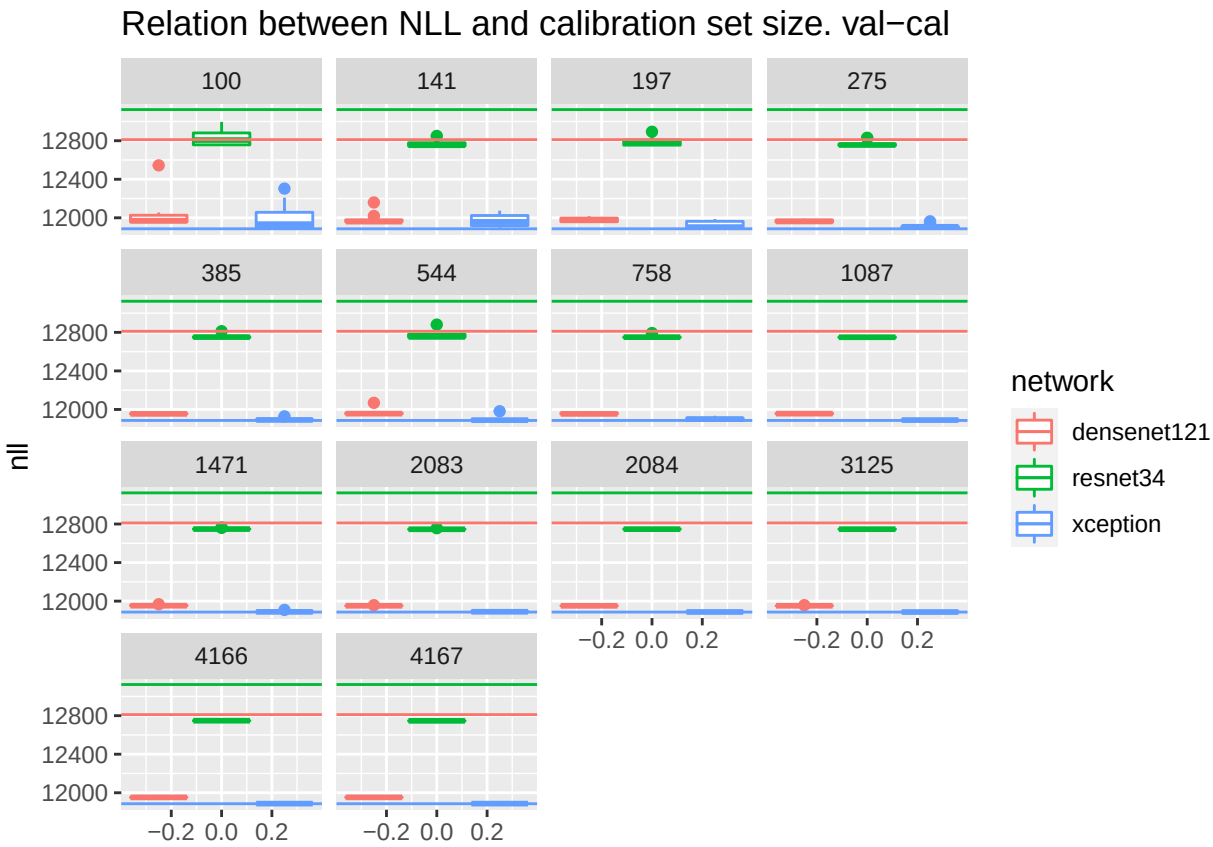
Since calibration on training data proved to be contraproductive, we also tested calibration on validation data.

```
net_calibrations_ece <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "vc"),
    mapping = aes(y = ece, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "vc"),
    mapping = aes(yintercept = ece, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between estimated calibration errors and calibration set size. val-cal")
net_calibrations_ece
```



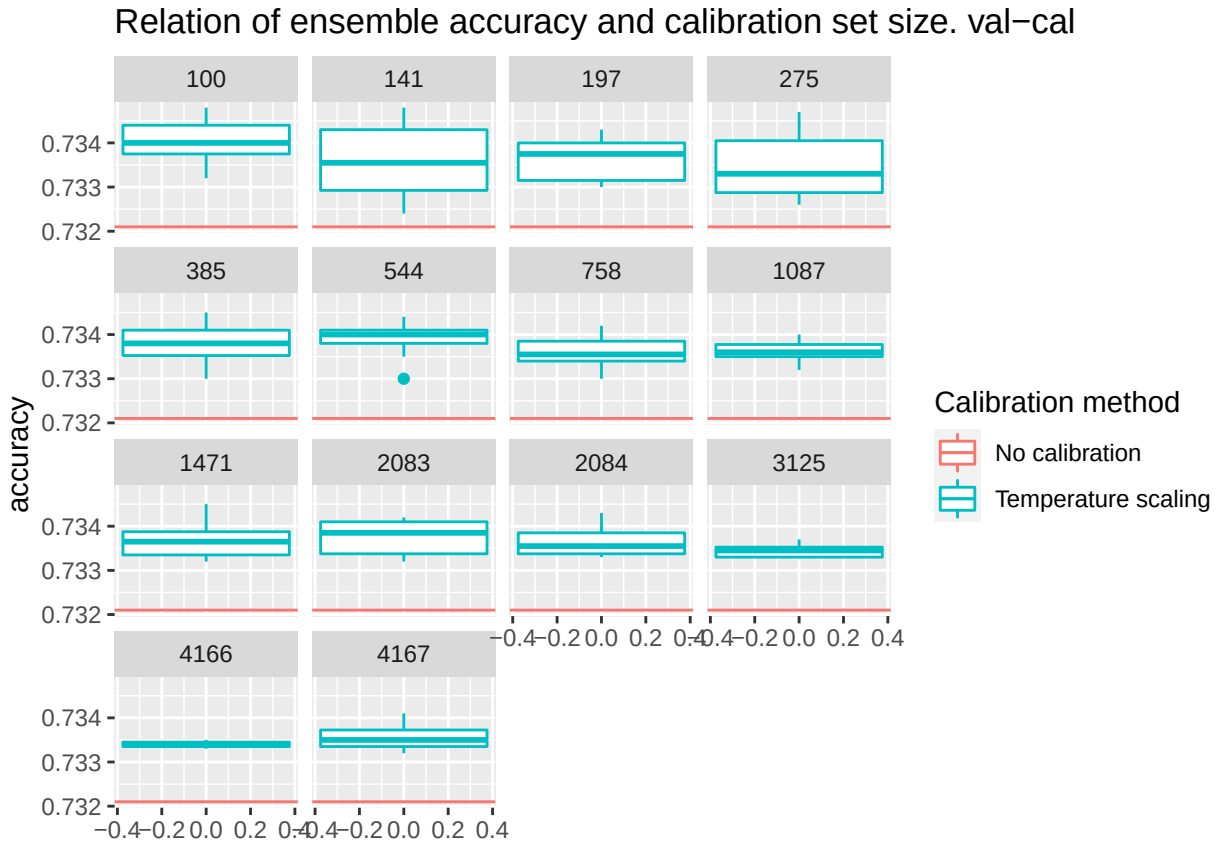
Calibration on validation data improves estimated calibration error for networks on testing data.

```
net_calibrations_nll <- ggplot() +
  geom_boxplot(
    data = metrics_net_cal %>% filter(cal_type == "vc"),
    mapping = aes(y = nll, color = network)
  ) +
  geom_hline(
    data = metrics_net %>% filter(cal_type == "vc"),
    mapping = aes(yintercept = nll, color = network)) +
  facet_wrap(~train_size) +
  ggtitle("Relation between NLL and calibration set size. val-cal")
net_calibrations_nll
```



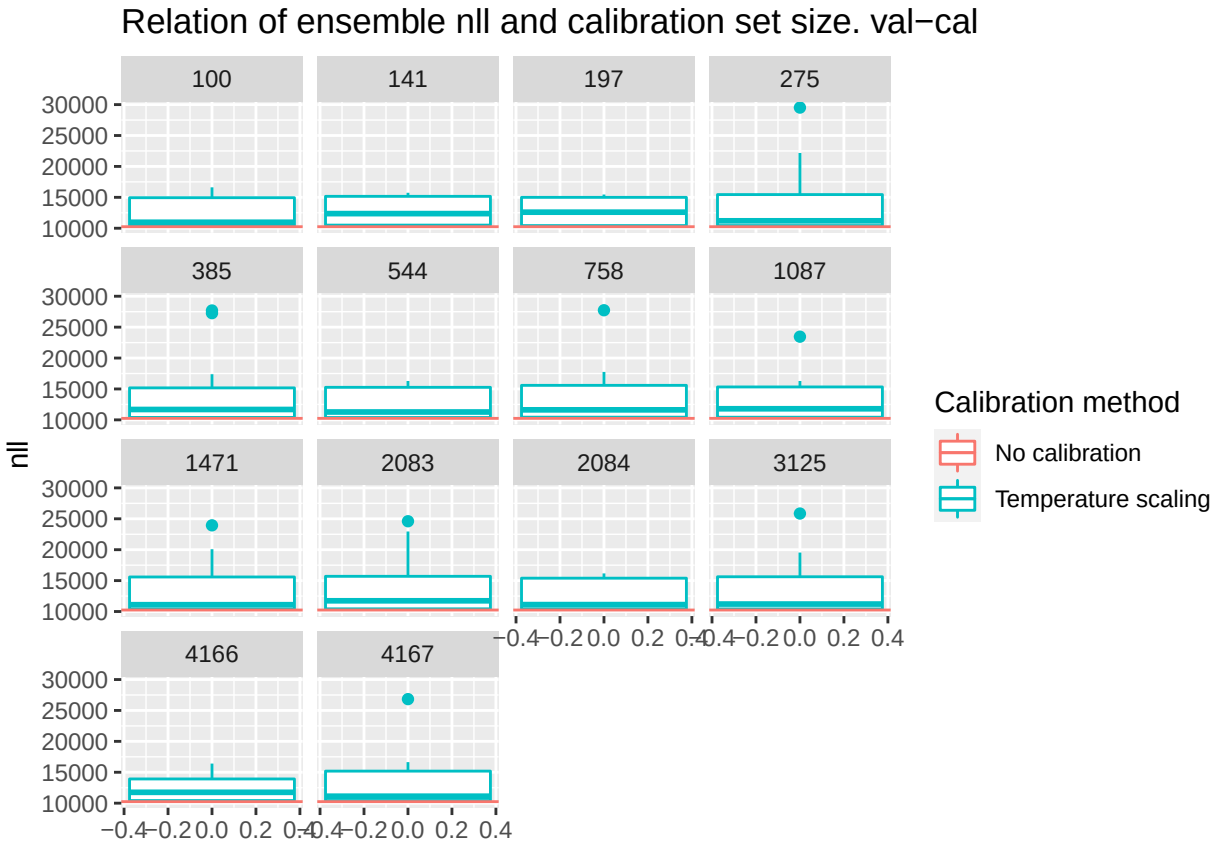
Calibration on validation data improves nll for networks on test set.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration" &
      cal_type == "vc")),
    mapping = aes(y = accuracy, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration",
      cal_type == "vc") %>%
      select(!train_size)),
    mapping = aes(yintercept = accuracy, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble accuracy and calibration set size. val-cal")
ens_accuracy
```



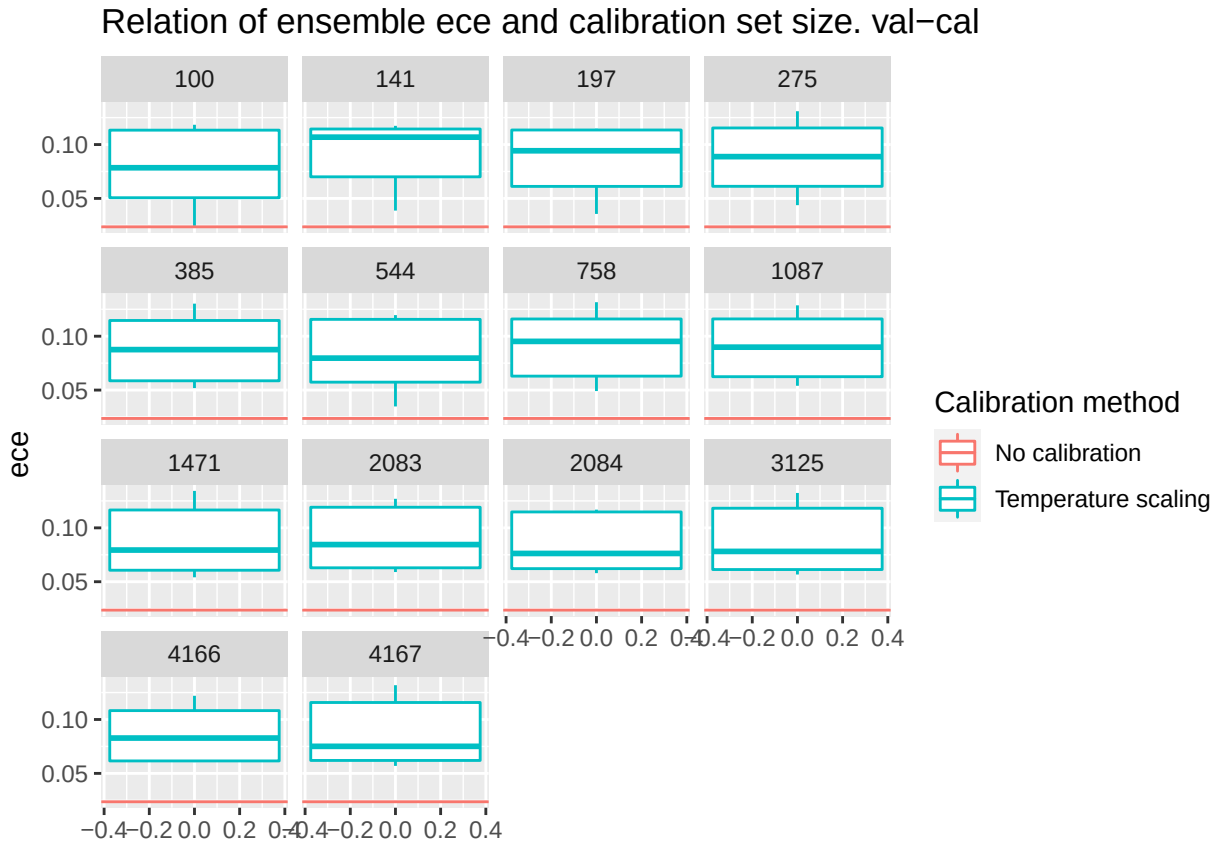
Calibration on validation data improves ensemble accuracy.

```
ens_nll <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = nll, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = nll, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble nll and calibration set size. val-cal")
ens_nll
```



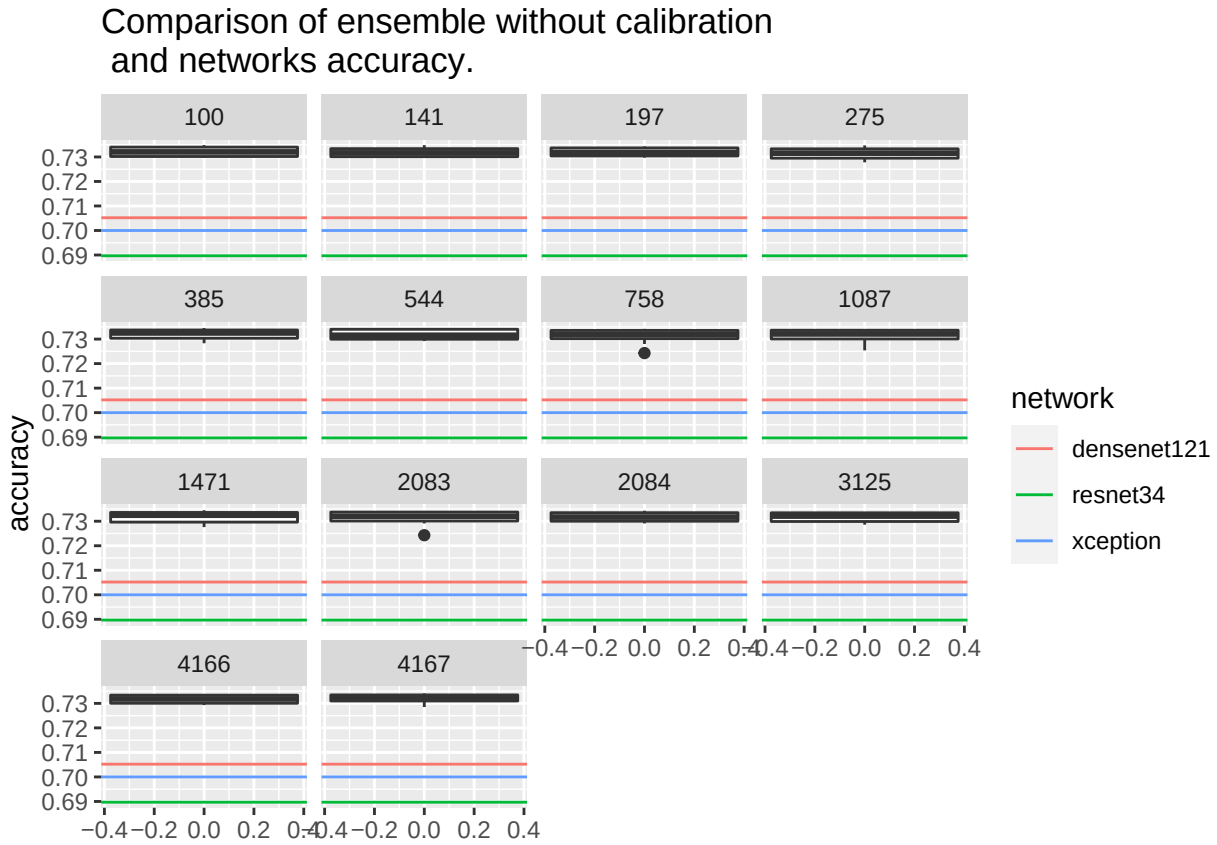
Calibration on validation data doesn't have large effect on the ensemble nll.

```
ens_ece <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = ece, color = "Temperature scaling")
  ) +
  geom_hline(
    data = (metrics_ens %>% filter(
      calibrating_method == "NoCalibration") %>%
      select(!train_size)),
    mapping = aes(yintercept = ece, color = "No calibration")
  ) +
  facet_wrap(~train_size) +
  guides(color = guide_legend(title = "Calibration method")) +
  ggtitle("Relation of ensemble ece and calibration set size. val-cal")
ens_ece
```



Calibration on validation data worsens the ensemble estimated calibration error.

```
ens_accuracy <- ggplot() +
  geom_boxplot(
    data = (metrics_ens %>% filter(calibrating_method != "NoCalibration")),
    mapping = aes(y = accuracy)
  ) +
  geom_hline(
    data = metrics_net,
    mapping = aes(yintercept = accuracy, color = network)
  ) +
  facet_wrap(~train_size) +
  ggtitle("Comparison of ensemble without calibration \n and networks accuracy.")
ens_accuracy
```


Comparison of ensemble without calibration
and networks accuracy.
100
141
197
275
385
544
758
1087
1471
2083
2084
3125
4166
4167
accuracy
0.73
0.72
0.71
0.70
0.69
−0.4
−0.2
0.0
0.2
0.4
network
densenet121
resnet34
xception