Table 3. The five best scoring methods of the longitudinal MS lesion segmentation challenge with challenge score, volume correlation (VC), Dice coefficient, positive predictive value (PPV), lesion false positive rate (LFPR), lesion true positive rate (LTPR). Dice, PPV, LFPR and LTPR are denoted in percent and best values out of five are printed in bold. In brackets we denote the relative weight of each metric on the final score.

	score	VC (1/4)	Dice (1/8)	PPV (1/8)	LFPR (1/4)	LTPR (1/4)
asmsl (proposed)	**92.076**	0.862	62.98	**84.46**	20.13	48.71
nic_vicorob_test	91.440	0.840	64.29	79.25	15.46	38.72
VIC_TF_FULL	91.331	0.866	63.05	78.67	**15.29**	36.40
MIPLAB_v3	91.267	0.823	62.74	79.97	23.17	45.40
miac_results [1]	91.011	**0.867**	**66.78**	74.05	40.73	**58.29**

4 Discussion

We encountered expected as well as odd behavior in our exploratory study. Contrary to what has been advised for in the literature [12], dropping information from the state weights results in better regularization, as the Dice coefficients in Table 1 B indicate. This behavior could be due to the fact that we only ignore part of the previous state per iteration and channel. Interestingly though, a combination of DC on both input and state produces worse results, even with a reduced drop rate. As dropout tends to prolong training, further experiments with longer training times might shed light on this effect. Another surprising result is the inability of BN to surpass baseline Dice scores in our preliminary tests in Table 1, in all variations we tested. Due to the correlation in our mini-batch of one and the varying weights in the case of the running average, the assumption does not hold that the statistics of our mini-batch are similar to the global statistics. Residual learning between MD-GRU layers seems to contribute to the overall improvement. Surprisingly, neither concatenating the C-GRU nor placing the reset gate as in the original GRU did result in an improved Dice.

The high pass filtering as preprocessing step proved to be fruitful, especially in the setting where we only trained for 3000 iterations, where leaving it out resulted in no segmentation at all. Using only original data, a visible tendency towards lesions could be found, but with probabilities well below 0.5. The main reason why this step is so important can be seen in Fig. 3, where values of the filtered image lie mostly around zero and in the original scan around two. All the weights of our network are initialized to handle data from a standard normal distribution. Inside the brain, filtering the original image would result in sums far away from zero. Using a hyperbolic tangent or sigmoid function on such a result will return a value close to 1 and hence a very flat gradient, which will not be able to help adjust the weights to correct for this in a fast manner.

Selective sampling and random deformation succeeded to be the most important improvements, which is easily explainable with the huge class imbalance present in our data and the low amount of training data. As the crossvalida-